IntervalGP-VAE: Learning Unobserved Confounders with Uncertainty for Personalized Causal Effect Estimation

Anonymous authors
Paper under double-blind review

Abstract

Estimating individual treatment effects (ITEs) in the presence of unobserved confounding remains a central challenge in causal inference. Existing proxy-based methods aim to recover latent confounders from observational proxies, but typically produce only point estimates without uncertainty quantification. This lack of uncertainty modeling provides incomplete and potentially insufficient information for downstream decision-making, especially when uncertainty is inherent in the data. We propose IntervalGP-VAE, a novel framework that combines variational autoencoders with Gaussian Processes (GPs) to model both the latent confounders and their associated uncertainty. At the core of our method is an interval-valued GP prior, which enables the model to capture a distribution over plausible latent confounders and treatment responses, rather than relying on potentially unreliable point estimates. This approach accounts for uncertainty arising from noisy and imperfect proxy variables and yields calibrated ITE intervals to support more robust causal decisions. We provide theoretical guarantees for identifiability of the latent confounder up to a smooth monotonic transformation under weak assumptions. Experiments on synthetic and semi-synthetic datasets demonstrate that IntervalGP-VAE achieves superior performance in ITE estimation and uncertainty calibration, outperforming existing methods.

1 Introduction

Estimating Individual Treatment Effects (ITEs) from observational data is a core challenge in causal inference, particularly when important confounding variables are not directly observed (Pearl, 2009; Peters et al., 2017). In many real-world applications, such as healthcare, economics, and social sciences, treatment assignment is influenced by latent factors that affect both the treatment and the outcome. Ignoring such hidden confounders can lead to biased causal estimates and unreliable treatment recommendations.

A promising line of recent research addresses this challenge by recovering latent confounders from observed *proxy variables* that carry indirect information about the hidden confounding structure (Louizos et al., 2017; Zhang et al., 2021; Wu et al., 2024; Harada & Kashima, 2024). Proxy-based approaches have demonstrated that under suitable structural assumptions, the latent confounder and causal effects can be recovered even when the confounder itself is unobserved. However, in practice proxy measurements are often noisy, incomplete, and weakly informative. As a result, recovering the latent confounder becomes an inverse problem under measurement error, where a single proxy observation corresponds not to a unique latent state but to a distribution of plausible latent values.

Consequently, two fundamental challenges arise in proxy-based causal inference. First, one must determine *when* recovery of latent confounders and individual causal effects is theoretically possible from proxy observations. Second, even when identifiability conditions hold, one must quantify *how much uncertainty remains* in the recovered latent structure and the resulting causal estimates due to measurement noise.

Our theoretical results (Theorem 1) address the first question by establishing minimal structural conditions, most notably the number of proxies, conditional independence, and variability, under which the latent confounder and ITE become identifiable in principle. Under these conditions, the latent confounder is identifiable up to a smooth invertible transformation, which does not affect causal identifiability (Theorem 2).

However, identifiability alone does not imply accurate pointwise recovery: even when the model is identifiable in principle, noisy proxy measurements induce unavoidable uncertainty in the inferred latent confounder and the resulting causal estimates. This observation highlights a fundamental gap between *identifiability* and *estimation*. While identifiability theory clarifies the structural conditions under which causal recovery is possible, practical inference must still account for the epistemic uncertainty arising from noisy proxy measurements and stochastic outcome generation mechanisms. Reliable decision support therefore requires explicit modeling and propagation of this uncertainty when estimating counterfactual outcomes and ITEs.

Our interval formulation is designed to address this second challenge by capturing the uncertainty induced by measurement noise. Because proxies are noisy measurements of the latent confounder, the posterior distribution of the confounding variable induced by proxy observations typically spans a range of plausible latent values rather than collapsing to a single point. Moreover, the structural outcome equations governing the data-generating process introduce additional exogenous noise in the outcome mechanism, implying that even conditional on the latent confounder, counterfactual outcomes remain uncertain. Especially, in high-dimensional and noisy settings, point estimates of the latent confounder can become unstable and overly sensitive to measurement noise. Instead, a more appropriate representation is to infer a *set* or *distribution* of plausible latent values that reflects the uncertainty induced by proxy noise while preserving similarity relationships in the proxy space. Interval representations therefore provide a concise and interpretable summary of the combined uncertainty induced by proxy measurement noise and outcome variability, avoiding overconfident point estimates while remaining faithful to the underlying structural model.

Beyond uncertainty quantification, effective proxy-based confounder recovery also requires coherent latent geometry: individuals with similar proxy profiles should map to nearby latent confounder values, whereas distinct proxy configurations should remain well separated in the latent space. To support principled inference under this uncertainty, we model the relationship between proxies and latent confounders using Gaussian Processes (GPs) (Rasmussen & Williams, 2006). GPs provide a flexible nonparametric Bayesian framework that captures uncertainty in function estimation while enforcing smoothness and similarity structure in the input space. By imposing a GP prior over the latent confounder as a function of the proxies, individuals with similar proxy profiles induce similar latent posterior distributions, while uncertainty increases in regions of the proxy space with limited data support. The GP posterior therefore provides calibrated uncertainty estimates that can be propagated through downstream causal predictions.

Building on these insights, we propose *IntervalGP-VAE*, a novel framework that integrates interval-valued Gaussian Processes with variational autoencoders (VAEs) to recover latent confounders with uncertainty for personalized causal effect estimation. The VAE encoder learns a structured latent representation from noisy proxy measurements, enabling flexible modeling of complex proxy–confounder relationships. An interval-valued GP prior is imposed over this latent representation to model the confounder as a smooth function of the proxies while explicitly accounting for measurement-induced uncertainty. Outcome prediction is further modeled using interval-valued GP regression that maps inferred confounder intervals to calibrated outcome bounds. This formulation enables coherent uncertainty propagation from noisy proxies through latent confounder recovery to interval-valued ITE estimates.

The resulting framework links theory and method in a coherent manner. Identifiability results determine *when* causal recovery from proxies is possible by specifying minimal structural conditions on the proxy generation mechanism. Once these conditions hold, the proposed IntervalGP-VAE provides a principled approach for *how* uncertainty-aware inference can be performed in practice. The interval representation captures unavoidable uncertainty arising from noisy proxy and outcome measurements, while the Gaussian Process provides a Bayesian mechanism for smooth, data-adaptive inference and calibrated uncertainty propagation.

To the best of our knowledge, IntervalGP-VAE is the first method that combines proxy-based latent variable modeling with interval-valued Gaussian Processes for causal effect estimation. While Gaussian Processes have been widely used for regression and uncertainty quantification, existing methods rarely address interval-valued observations, and none incorporate GP priors for recovering latent confounders from proxy variables. Existing approaches such as CEVAE (Louizos et al., 2017) typically rely on simple latent priors (e.g., isotropic Gaussian distributions) and do not explicitly model the structured uncertainty induced by

proxy measurements. In contrast, our framework introduces structured Bayesian priors that enable both smooth latent recovery and principled uncertainty propagation.

Our main contributions are summarised as follows:

- We present a theoretical analysis of the identifiability conditions under which ITE can be recovered from proxies, offering formal guarantees for the proposed method.
- We propose IntervalGP-VAE, a novel framework that combines VAEs with an interval-valued GP prior over the latent space and a GP-based interval likelihood head to enable smooth and uncertainty-aware estimation of counterfactual outcomes and treatment effects.
- We evaluate on 24 synthetic settings constructed to satisfy the identification conditions, and on the semi-synthetic IHDP benchmark across 100 replications. IntervalGP-VAE achieves lower or comparable PEHE and ATE error relative to TEDVAE (Zhang et al., 2021), while additionally providing calibrated ITE intervals.

2 Problem Setting

Key notations used in this paper are summarised in Table 1. We consider the following structural causal model.

Outcome model. The observed outcome is generated according to the structural equation

$$Y = f(T, U, Z_Y) + \epsilon_Y, \tag{1}$$

where $T \in \{0, 1\}$ is the treatment, $U \in \mathbb{R}^d$ denotes latent confounders, $Z_Y \subseteq Z$ is a subset of proxies that directly affect the outcome, and the exogenous noise satisfies $\epsilon_Y \perp (T, U, Z_Y)$.

Table 1: Summary of key notations.

Symbol	Description
$Z \in \mathbb{R}^k$	Proxy variables
$Z_Y \in \mathbb{R}^{k'}$	Outcome-related proxies
$T \in \{0, 1\}$	Binary treatment variable
$Y \in \mathbb{R}$	Outcome variable
$U \in \mathbb{R}^d$	Latent confounder(s)
ϵ_Y, ϵ_Z	Noise for Y and Z
$f(T, U, Z_Y)$	Outcome function
$g(U)$	Mapping from U to proxies
$\hat{Y}(t)$	Estimated outcome under t
$q_\phi(u \mid z)$	Posterior over U (encoder)
$p_\theta(z \mid u, \epsilon)$	Proxy likelihood (decoder)
$[\hat{\tau}_j^{\text{lower}}, \hat{\tau}_j^{\text{upper}}]$	ITE interval for individual j

ITE and counterfactual. Following Pearl's abduction–action–prediction semantics for counterfactuals, let $\epsilon_{Y,j}$ denote the realised exogenous noise in Eq. 1 for individual j. The potential outcome under the intervention $T := t$ is obtained by substituting t into the same structural equation:

$$Y_j(t) := f(t, U_j, Z_{Y,j}) + \epsilon_{Y,j}. \tag{2}$$

The *individual treatment effect (ITE)* is then defined as the difference between the two counterfactual outcomes for the *same* realisation of the exogenous variables:

$$\tau_j := Y_j(1) - Y_j(0) = f(1, U_j, Z_{Y,j}) - f(0, U_j, Z_{Y,j}), \tag{3}$$

where the exogenous noise cancels because it is held fixed across the two counterfactual worlds.

Proxy model. We do not observe U directly but instead observe proxies $Z = (Z_1, \ldots, Z_k)$ generated via

$$Z_i = g_i(U) + \epsilon_{Z_i}, \qquad \epsilon_{Z_i} \perp U, \quad i = 1, \ldots, k, \tag{4}$$

where the noise terms are mutually independent and the structural functions g_i are unknown but sufficiently non-redundant to allow information about U to be recovered from Z.

Observed data There are n samples of (Z, T, Y) in the observed data, and our aim is twofold:

- **Identifiability:** To determine the structural conditions under which the latent confounder U can be recovered (up to an equivalence class) from the proxy variables, so that τ_j becomes identifiable from the observed distribution $p(Z, T, Y)$.
- **Uncertainty quantification:** To develop principled interval-valued estimates of treatment outcomes and ITEs with uncertainty.

To support identifiability of the latent confounder U and ITE τ_j, we adopt the following assumptions:

- **Positivity.** Each individual has a non-zero probability of receiving either treatment level:
$$0 < P(T = 1 \mid U) < 1.$$
- **Latent Ignorability.** Conditional on the latent confounder U, the treatment assignment is independent of potential outcomes:
$$(Y(0), Y(1)) \perp\!\!\!\perp T \mid U.$$
Thus U fully accounts for confounding between T and Y.
- **Consistency and well-defined outcomes.** For each $t \in \{0, 1\}$, the potential outcome $Y(t)$ is defined by substituting the intervention $T := t$ into the structural outcome equation (Eq. 1).

In addition to the standard causal assumptions above, identifiability from proxies requires two structural conditions on the proxy-generating process:

- **Conditional independence of proxies.** For all $i \neq j$,
$$Z_i \perp\!\!\!\perp Z_j \mid U. \tag{5}$$
This is consistent with the structural model $Z_i = g_i(U) + \epsilon_{Z_i}$ with mutually independent noise terms.
- **Variability (injectivity-in-distribution).** For any two distinct latent states $u_1 \neq u_2$, the conditional proxy distributions differ across many coordinates:
$$p(Z_i \mid U = u_1) \neq p(Z_i \mid U = u_2) \quad \text{for many } i. \tag{6}$$
This ensures that different values of U induce distinguishable proxy distributions, enabling identification up to an appropriate reparameterisation.

The assumptions of positivity, latent ignorability, and consistency are standard in the causal inference literature and foundational for potential outcomes frameworks. In our setting, identifiability further relies on structural assumptions for the proxy and outcome models. We assume that the proxies are generated via an additive noise model (Eq. 4) with mutually independent noise terms independent of U. This formulation is widely used in nonlinear latent variable models and additive noise models (ANMs). The conditional independence of proxies given U, as well as the *variability* condition, requiring that different values of U induce distinguishable distributions over the proxies, are consistent with established identifiability results in proxy or auxiliary variable-based latent models (e.g., nonlinear ICA (Hyvärinen et al., 2019; Khemakhem et al., 2020)).

3 Related Work

We categorize related work into three areas: latent confounder modeling, Gaussian Processes and VAE, and interval-valued Gaussian Processes.

3.1 From Proxy-Based Identification to Personalized Latent Recovery

Leveraging proxy variables to identify causal effects in the presence of unobserved confounding is a classical strategy in causal inference. Early work (Kuroki & Pearl, 2014; Miao et al., 2018) derived identifiability conditions under linear or parametric assumptions, and subsequent neural approaches such as DeepPCL (Xu et al., 2021) extended these ideas to nonlinear settings. However, these methods primarily address population-level identification and sometimes rely on unverifiable assumptions, such as splitting proxies into treatment- or outcome-specific groups. Recent research has shifted toward *personalized* latent recovery. Several VAE-based models, CEVAE (Louizos et al., 2017), TEDVAE (Zhang et al., 2021), CEMVAE (Wu et al., 2024), InfoVAE (Zhao et al., 2019), and InfoCEVAE (Harada & Kashima, 2024), aim to learn latent confounders from proxies, and GP-based alternatives such as the Sequential Deconfounder and Structured GP Confounder (Kuzmanovic et al., 2021; Hatt & Feuerriegel, 2024; Witty et al., 2020) provide additional modelling flexibility. Nevertheless, these approaches largely lack structural identifiability guarantees. While InfoVAE (Zhao et al., 2019) and InfoCEVAE (Harada & Kashima, 2024) encourage informative latent recovery via mutual information, they do not provide identifiability guarantees linking proxies to the latent space.

In contrast, our method establishes structural identifiability through the tensor decomposition framework of Kruskal's theorem (Kruskal, 1977; Allman et al., 2009). We show that under nonlinear and injective proxy mappings, the latent confounder $U \in \mathbb{R}^d$ is identifiable (up to a smooth invertible transformation) from $k \geq 2d + 4$ proxy variables. This ensures that the recovered latent representation supports valid causal inference, including ITE estimation.

Notably, we compare performance with TEDVAE (Zhang et al., 2021), a widely used deep latent-variable approach for treatment effect estimation. TEDVAE assumes unconfoundedness of the form $T \perp (Y(0), Y(1)) \mid X$, where X denotes a learned representation of the observed covariates. However, the structural graph underlying TEDVAE explicitly includes an unobserved confounder Z_c that affects both treatment and outcome. This introduces a backdoor path $T \leftarrow Z_c \rightarrow Y$, implying that the causal effect $T \rightarrow Y$ is not identifiable unless the learned representation X fully captures the latent confounder Z_c. In other words, unconfoundedness holds only if X constitutes a complete representation of the confounding structure. While TEDVAE assumes that such a representation can be recovered during learning, the conditions under which this assumption holds are not analysed or characterised in the original work.

In contrast, our framework starts from the more general setting in which treatment effects are not identifiable from the observed variables alone due to latent confounding. We explicitly derive structural conditions, including conditional independence of proxies, sufficient variability of proxy responses, and a sufficient number of proxy variables, under which the latent confounder U becomes identifiable from noisy proxy measurements. Once these conditions hold, the recovered latent confounder induces an unconfounded representation and thereby restores causal identifiability. Thus, while both methods ultimately rely on an unconfounded representation to estimate treatment effects, our approach provides the theoretical conditions under which such unconfoundedness can be guaranteed, whereas TEDVAE assumes it without analysing when it holds.

Another related line of work in identifiable representation learning, including nonlinear ICA (Hyvärinen et al., 2019), identifiable VAEs (Khemakhem et al., 2020), and causal-effect VAEs such as (Wu & Fukumizu, 2021), exploits the principle that sufficiently rich variability in conditional distributions can render latent representations identifiable up to an invertible transformation. These approaches build on the idea that when an auxiliary or modulation variable induces sufficiently diverse changes in conditional distributions (e.g. $p(Z \mid U)$), the latent representation can be uniquely determined up to an invertible mapping. This forms a necessary ingredient within our broader identifiability result for proxy-based latent confounder recovery. However, despite this conceptual similarity, the identification problems addressed in these works differ fundamentally from ours. In nonlinear ICA and related identifiable generative models, the auxiliary or modulation variable driving the variability (for example, the variable u in nonlinear ICA or identifiable VAE) is *observed*. Identifiability arises because the entire conditional distribution of the latent variables changes in sufficiently rich ways as the observed auxiliary variable varies. For each fixed value of the auxiliary variable, the model generates a full distribution over the latent components, and the observed modulation enables recovery of the latent sources.

In contrast, our setting concerns proxy-based latent confounder recovery for causal inference. Here the latent confounder U is *unobserved* and must be inferred from multiple noisy proxy variables $\{Z_i\}_{i=1}^k$ generated from U. The goal is not source separation but the recovery of latent confounders for personalised treatment effect estimation. In this setting the variability condition is necessary but not sufficient: because U itself is hidden, one must additionally determine how many conditionally independent proxy variables are required to recover U from its measurements. Our Theorem 1 provides precisely such a result, establishing a proxy-count requirement together with structural conditions under which the latent confounder becomes identifiable from noisy proxies. Thus, while our work is conceptually related to identifiable generative modelling through the shared variability principle, it addresses a distinct identification problem: recovering latent confounders from proxy variables to enable personalised causal effect estimation under hidden confounding. Existing identifiable representation learning approaches do not provide such proxy-based identifiability guarantees.

3.2 GP-VAE and Uncertainty-Aware Estimation

Gaussian Processes (GPs) offer a flexible, nonparametric framework for uncertainty modeling (Rasmussen & Williams, 2006). The foundational work by (Casale et al., 2018) introduced GP-VAEs, replacing the isotropic Gaussian prior in VAEs with a GP prior to induce structured latent representations that vary smoothly with the input. Subsequent work extended this idea to sequential data (Fortuin et al., 2020) and latent confounder trajectories (Hatt & Feuerriegel, 2024). In the Structured GP Confounder model (Witty et al., 2020), separate GPs are used to model treatment and outcome given latent confounders. Our proposed IntervalGP-VAE introduces an interval-valued GP prior over the latent space to capture both structured dependencies and epistemic uncertainty in confounder recovery from noisy proxies. For outcome prediction, we incorporate a GP head trained on interval-valued targets to produces calibrated predictive uncertainty for counterfactual queries. To the best of our knowledge, this is the first model to integrate GP-based latent inference with interval-valued uncertainty propagation for ITE estimation.

3.3 Interval-Valued Gaussian Processes

Modeling interval-valued outputs with GPs has received limited attention in the literature. To the best of our knowledge, the only existing work that explicitly supports interval observations in a GP setting is the Generalized Multi-Output Censored GP model by (Gammelli et al., 2020; 2022). Their framework introduces a likelihood formulation capable of handling output intervals across multiple outputs. However, their method is designed for multi-output regression tasks and does not address causal inference, latent confounders, or proxy variables. Our approach supports interval supervision of the latent space and, through a GP, interval prediction of individual treatment effects. This allows us to model both uncertainty in confounder inference and outcome prediction in a principled and calibrated manner.

4 Identifiability Analysis

Definition 1 (Identifiability of the latent confounder for ITE estimation)**.** Under the structural equations (Eqs. 1–4), the latent confounder U is said to be *identifiable for ITE estimation* from the observational distribution $p(Z, T, Y)$ if there exists a measurable mapping $\phi : \mathbb{R}^k \to \mathbb{R}^d$ such that the recovered representation $\hat{U} = \phi(Z) = h(U)$ for some smooth and invertible transformation $h : \mathbb{R}^d \to \mathbb{R}^d$, and the individual treatment effect (Eq. 3) is unchanged under this transformation:

$$\tau_j(h(U), Z_{Y,j}) = \tau_j(U, Z_{Y,j}) \qquad \text{for all individuals } j.$$

Thus identifiability requires recovering U up to an invertible reparameterisation that preserves the ITE.

The purpose of Definition 1 is to formalise the notion of identifiability that is necessary for ITE estimation. The primary objective of this work is the recovery of ITEs, for which the latent variable U serves as an intermediate representation enabling personalised causal effect estimation. Definition 1 is formulated to capture precisely the level of identifiability needed for personalised causal effect estimation: the latent state is identifiable up to an invertible transformation, and this degree of identifiability is fully sufficient for ITE estimation.

Remark 1. Thus, U need not be identifiable in an absolute sense; it is sufficient that the aspects of U required for computing the ITE in Eq. 3 are uniquely recoverable from the proxy variables up to an invertible change of coordinates.

Remark 2. In our framework, the latent confounder U is an intermediate representation. Because Definition 1 implies that $p(Z)$ identifies U up to a smooth invertible transformation under which the ITE is invariant, recovering any element of this equivalence class is sufficient. Our focus is thus on causal identifiability rather than exact reconstruction of U.

In the ideal case where U is directly observable, the structural outcome equation (Eq. 1) yields identifiable causal effects under standard assumptions. However, in realistic settings U is unobserved and must be inferred from proxy variables. Thus, the central question becomes: under what conditions can U be recovered from proxies (up to a smooth and invertible transformation), thereby ensuring identifiability of causal effects? Theorem 1 provides the formal conditions required to achieve this.

Theorem 1 (Identifiability of the latent confounder from noisy proxies)

Consider the proxy–latent relationship defined in the structural equation (Eq. 4), and suppose the proxy variables satisfy the assumptions of conditional independence (Eq. 5) and variability (Eq. 6). If the total number of proxies satisfies

$$k \;\geq\; 2d + 4,$$

then the latent confounder $U \in \mathbb{R}^d$ is identifiable from the observed proxy distribution $p(Z)$ up to a smooth, invertible reparameterisation.

Proof. See Appendix A for the proof of Theorem 1. □

The key reason the latent confounder becomes identifiable from the proxy distribution relies on the combination of two structural properties: conditional independence of the proxies given the latent variable (Eq. 5), and sufficient variability of the proxy responses across latent states (Eq. 6),.

First, conditional independence means that once the latent confounder is fixed, each proxy behaves as an independent observation channel of the same hidden state. In other words, the proxies provide multiple independent views of the latent confounder. This structure allows information about the latent state to accumulate across proxies: each proxy contributes partial information, and the joint behaviour of several proxies becomes progressively more informative about the underlying latent variable.

Second, the variability assumption ensures that different latent states induce genuinely different proxy distributions. Intuitively, this condition guarantees that changes in the latent confounder produce observable changes in the proxy behaviour. If the proxy distributions were identical across latent states, the proxies would carry no information about the latent variable and identifiability would be impossible. The variability condition therefore ensures that the mapping from the latent confounder to the proxy distribution is injective at the level of conditional distributions.

When these two properties are combined, the joint proxy distribution becomes highly structured. Each proxy contributes a distinct and independent signal about the latent state, and the variability condition ensures that these signals vary across latent values rather than collapsing into redundant information. As more independent proxy channels are observed, the joint proxy distribution becomes increasingly constrained, and the latent state responsible for generating it becomes uniquely determined up to a reparameterisation.

The requirement on the number of proxies reflects the amount of independent information needed to resolve the latent confounder. Because the confounder lives in a d-dimensional space, sufficiently many independent proxy channels are required to distinguish between different latent states. Once the number of proxies exceeds this threshold, the joint proxy distribution contains enough independent variation to uniquely determine the latent confounder structure (up to a smooth invertible transformation). This is precisely the regime captured by the condition $k \geq 2d + 4$ in Theorem 1.

More details can be found in Appendix A

Since identifiability only holds up to a smooth, strictly monotonic, and invertible transformation of U, a natural question arises: does this ambiguity affect ITE estimation? Theorem 2 shows that ITEs are invariant to such reparameterisations of the latent space, for any fixed Z_Y[1].

Theorem 2 (ITE invariance under latent reparameterisation)

Let $h : \mathbb{R}^d \to \mathbb{R}^d$ be a smooth and invertible transformation, and define the reparameterised latent variable $\hat{U} = h(U)$. Then the structural individual treatment effect defined in equation 3 is invariant under such reparameterisations:

$$\tau_j(\hat{U}_j, Z_{Y,j}) \;=\; \tau_j(U_j, Z_{Y,j}) \qquad \text{for all individuals } j.$$

Proof. See Appendix B for the proof of Theorem 2. □

The central goal of the paper is personalised ITE estimation, for which recovering the latent confounder U is a necessary but not final step. Theorem 2 showing that the ITE is invariant under any bijective reparameterisation of U. In other words, even if U is recovered only up to an invertible map, the personalised causal quantity of interest remains uniquely defined and unaffected. Thus, Theorem 2 is an essential conceptual bridge between latent-variable identifiability and the ultimate causal target.

5 IntervalGP-VAE

5.1 Motivation for GP Prior

Section 4 establishes that, under the conditional-independence and variability assumptions, the latent confounder U is identifiable from the proxy variables Z up to a smooth invertible reparameterisation. In practice, however, learning the inverse mapping $Z \mapsto \hat{U}$ from noisy proxies is statistically ill posed: even when the proxy mechanism is identifiable in theory, finite-sample noise and encoder stochasticity can yield unstable or overly dispersed latent estimates. To regularise this inversion, we place a GP prior over $\hat{U}(Z)$ using a kernel K that enforces *latent-space coherence* while retaining uncertainty in regions where proxy information is weak. Proposition 1 examines two questions arising from this construction: (i) whether introducing a GP prior alters identifiability of causal quantities such as the ITE, and (ii) whether the ITE's invariance to smooth invertible transformations of U remains compatible with GP-based latent modelling.

Proposition 1 (Compatibility of GP priors with identifiability and latent reparameterisations)

Let $\hat{U} = (\hat{U}_1, \ldots, \hat{U}_n)$ denote the latent confounder values inferred from the proxy variables $Z^1, \ldots, Z^n$ under the structural proxy model (Eq. 4). Suppose a Gaussian Process prior

$$\hat{U} \sim \mathcal{GP}\big(0, K(Z^i, Z^j)\big)$$

is placed on the inferred latent representation, where K is any positive-definite kernel on the proxy space. Then:

1. Introducing a GP prior on $\hat{U}$ does not alter the identifiability of the individual treatment effect (ITE) established in Theorems 1.
2. For any smooth and invertible reparameterisation $h : \mathbb{R}^d \to \mathbb{R}^d$, replacing U by $\tilde{U} = h(U)$ does not affect the role of K as a covariance function over the latent representation.

Proof. See Appendix C for the detailed proof of Proposition 1. □

[1] If Z_Y appears in the outcome model, it is assumed to be observed and held fixed when evaluating counterfactual outcomes.

Proposition 1 shows the answer to both questions: the GP prior leaves identifiability intact, and the latent reparameterisation freedom described in Theorem 2 is fully compatible with GP modelling through an induced kernel. Therefore, the GP component plays no role in our identifiability analysis. Theorem 1 establishes identifiability of U (up to an invertible reparameterisation) using only conditional independence and generic variability of the proxies; this result is entirely independent of GP modelling. The role of the GP prior is instead to induce latent-space coherence and uncertainty quantification during amortised inference. The smoothness induced by the kernel acts only on the inference-time mapping $Z \mapsto \hat{U}$ and should be understood as encouraging a geometric regularity in the recovered latent manifold: proxy observations that are similar should lead to similar latent representations. This GP-induced coherence (by our IntervalGP) directly enables interval-valued uncertainty quantification for the ITE by capturing epistemic uncertainty arising from noisy proxies and by propagating this uncertainty through the outcome model in a mathematically principled way. Hence, the IntervalGP prior provides a natural way to model such uncertainty while preserving local structure. This coherent latent geometry is crucial for reliable personalised treatment effect estimation, especially in settings where unobserved confounding is strong and proxy noise varies substantially across samples.

Also, our framework is compatible with a wide range of kernels, including Matérn kernels with controllable smoothness, rational quadratic kernels, and non-stationary kernels that adapt to heterogeneous regions of the latent space. Each defines a different notion of latent-space coherence.

5.2 IntervalGPs for Latent Representation and ITE Estimation

We extend the VAE framework by introducing a structured prior over latent variables using *Interval Gaussian Processes (IntervalGP)* (Gammelli et al., 2020; 2022). Unlike standard VAEs that impose an isotropic Gaussian prior over latent representations, we define a GP prior over the proxy space Z and model the latent variable U as an interval-valued random function of Z.

Interval-Valued GP Regression. Given input features $X^i \in \mathbb{R}^d$ and scalar targets $y^i \in \mathbb{R}$, standard GP regression assumes a latent function $f(\mathbf{x}) \sim \mathcal{GP}(0, K(X, X'))$, where $K(\cdot, \cdot)$ is the kernel:

$$k(X^i, X^j) = \sigma_f^2 \exp\left(-\frac{\|X^i - X^j\|^2}{2\ell^2}\right), \tag{7}$$

with hyperparameters σ_f^2 (variance) and ℓ(lengthscale). In many practical settings, however, targets are not point-valued but known to lie within intervals: $y^i \in \left[y^{\text{lower},\,i},\, y^{\text{upper},\,i}\right]$. To accommodate this, IntervalGP (Gammelli et al., 2020; 2022) generalizes standard GP regression to interval targets by replacing the Gaussian likelihood with a truncated Gaussian likelihood:

$$p\big(y^{\text{lower},i} \leq f(X^i) \leq y^{\text{upper},i} \mid X^i\big) = \Phi\left(\frac{y^{\text{upper},i} - \mu^i}{\sigma^i}\right) - \Phi\left(\frac{y^{\text{lower},i} - \mu^i}{\sigma^i}\right). \tag{8}$$

where μ^i and σ^i denote the predictive mean and standard deviation of the GP at X^i, and $\Phi(\cdot)$ is the standard Gaussian cumulative distribution function (CDF).

IntervalGP Prior on the Latent Confounder. To infer the latent confounder U from observed proxies Z, we employ a variational encoder:

$$q(U \mid Z) = \mathcal{N}(\mu_u(Z),\, \sigma_u^2(Z)), \tag{9}$$

where each latent estimate U_i is treated as an interval-valued random variable, interpreted as $u_i \in [\mu_u^i - \sigma_u^i,\ \mu_u^i + \sigma_u^i]$. To regularize the learned latent space U, we impose a GP prior over $U(Z)$, i.e., $U(Z) \sim \mathcal{GP}(0,\ k(Z, Z'))$, is placed over the latent space to regularize the learned U-space. This GP prior treats each encoder-derived interval as an interval-valued observation under the GP posterior, enabling structured and uncertainty-aware regularization. For each data point i:

$$\log p_{\text{GP}}\big(u^i \in [\mu_u^i \pm \sigma_u^i] \mid Z^i\big) = \log\left[\Phi\left(\frac{\mu_u^i + \sigma_u^i - \mu^{\text{GP},i}}{\sigma^{\text{GP},i}}\right) - \Phi\left(\frac{\mu_u^i - \sigma_u^i - \mu^{\text{GP},i}}{\sigma^{\text{GP},i}}\right)\right]. \tag{10}$$

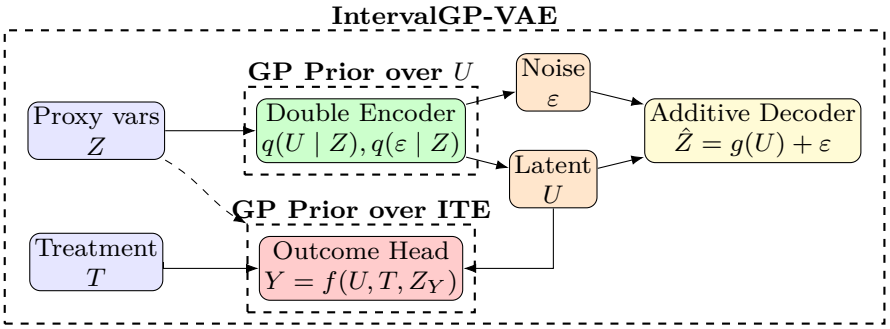

Figure 1: Architecture of the **IntervalGP-VAE** model.

where $\mu^{\text{GP},\,i}$ and $\sigma^{\text{GP},\,i}$ are the posterior predictive mean and standard deviation at input Z^i.

Latent Confounder Prediction via GP Posterior. The GP prior enables coherent prediction for unseen inputs using interval-valued observations. Specifically, we define:

- $K \in \mathbb{R}^{n\times n}$: the kernel matrix over training inputs, with $K_{ij} = K(Z^i, Z^j) + \sigma^2\delta^{ij}$,
- $K_s \in \mathbb{R}^{n\times 1}$: the cross-covariance vector between training inputs and a test input Z^*,
- $K_{ss} \in \mathbb{R}$: the prior variance at Z^*, i.e., $K_{ss} = k(Z^*, Z^*) + \sigma^2$.

The GP posterior over the latent confounder at Z^* is:

$$p(u^* \mid Z^*, \mathcal{D}) = \mathcal{N}(\mu^*, \sigma^{*2}), \tag{11}$$

where $\mu_u^* = K_s^\top K^{-1}\mu_u, \quad \sigma_u^{*2} = K_{ss} - K_s^\top K^{-1}K_s$. The resulting latent interval is interpreted as: $u^* \in [\mu_u^* - \sigma_u^*,\ \mu_u^* + \sigma_u^*]$.

ITE Prediction via GP Posterior. Similarly, the GP posterior for the ITE at test input Z^* is computed as:

$$\mu_{\text{ITE}}^{\text{lower}}(Z^*) = K_s^\top K^{-1}\mu_{\text{ITE}}^{\text{lower}}, \quad \mu_{\text{ITE}}^{\text{upper}}(Z^*) = K_s^\top K^{-1}\mu_{\text{ITE}}^{\text{upper}}, \quad \sigma_{\text{ITE}}^2(Z^*) = K_{ss} - K_s^\top K^{-1}K_s. \tag{12}$$

where $\mu_{\text{ITE}}^{\text{lower}}$ and $\mu_{\text{ITE}}^{\text{upper}}$ are obtained by drawing multiple samples from the GP posterior $q(u \mid Z^*)$ and computing empirical quantiles:

$$\text{ITE}_{\text{lower}} = \text{Quantile}_\alpha(Y_1 - Y_0), \quad \text{ITE}_{\text{upper}} = \text{Quantile}_{1-\alpha}(Y_1 - Y_0). \tag{13}$$

The final ITE prediction is expressed as an interval: $\text{ITE}_{\text{GP}}(Z^*) \in [\,\mu_{\text{ITE}}^{\text{lower}}(Z^*),\ \mu_{\text{ITE}}^{\text{upper}}(Z^*)\,]$, with uncertainty quantified by the predictive variance $\sigma_{\text{ITE}}^2(Z^*)$.

5.3 IntervalGP-VAE Architecture

The proposed *IntervalGP-VAE* extends the standard VAE framework by incorporating structured GP priors over both the latent confounder space U and ITE. This architecture enables the model to produce both point estimates and calibrated uncertainty intervals for latent variables and causal effects. The key components are outlined below and illustrated in Fig. 1: Encoder: Maps each input Z^i to a variational posterior over U, defining a latent interval $u^i \in [\mu_u^i - \sigma_u^i,\ \mu_u^i + \sigma_u^i]$ via Eq. 9 to capture epistemic uncertainty in the latent representation. IntervalGP over U: The GP prior regularizes the latent space by maximizing the interval-based log-likelihood $\log p_{\text{GP}}\left(\mu_u \pm \sigma_u \mid Z\right)$ as given in Eq. 10. Decoder: Reconstructs the proxies via $\hat{\mathbf{z}} = g_{\text{dec}}(u, \epsilon)$, using sampled latent variables and additive noise. Outcome Head: Predicts the outcome y from u, t, enabling estimation of potential outcomes and corresponding ITEs. IntervalGP over ITE: A second GP regressor predicts calibrated ITE intervals, propagating uncertainty from latent inference to treatment effect estimation. To train the model, we formulate a composite training objective that jointly optimizes

reconstruction accuracy and GP-based prior regularization. The overall training procedure is summarized in Algorithm 1. The weights and implementation-specific parameters used in the IntervalGP-VAE method are detailed in the Appendix D and Experiments section.

Algorithm 1: IntervalGP–VAE Algorithm

Input: Train set $\mathcal{D}_{\text{train}}$, test set $\mathcal{D}_{\text{test}}$, joint training epochs E
Output: Estimated ITEs $\{\hat{\tau}_j\}_{j=1}^m$; ITE intervals $[\hat{\tau}_j^{\text{lower}}, \hat{\tau}_j^{\text{upper}}]$

Initialize model $\mathcal{M}$ with
  encoder $q_\phi(u, \epsilon \mid z)$, decoder $p_\theta(z \mid u, \epsilon)$,
  and causal head $f_\psi(u, t)$;
for epoch $= 1$ **to** E **do**
  for *each mini-batch* (z, t, y) **do**
    Sample $(u, \epsilon) \sim q_\phi(u, \epsilon \mid z)$;
    Reconstruct $\hat{z} \sim p_\theta(z \mid u, \epsilon)$;
    Predict $\hat{y} = f_\psi(u, t, z_y)$;
    Compute the GP prior regularisation term;
    Compute total loss $\mathcal{L}$ and update (ϕ, θ, ψ);
  end
end
Post–training inference (for each test point z_j)
Compute GP posterior for U using Eq. 11 to obtain $[u_j^{\text{lower}}, u_j^{\text{upper}}]$ with posterior uncertainty;
Compute GP posterior for ITE using Eq. 12 to obtain $[\tau_j^{\text{lower}}, \tau_j^{\text{upper}}]$ with posterior uncertainty;
Output GP–based interval with posterior uncertainty.

5.4 Computational complexity

Although exact GP inference scales as $\mathcal{O}(n^3)$, the practical cost in our model is substantially lower. The interval-GP prior operates in the low-dimensional latent space, and all kernel evaluations, Cholesky factorizations, and predictive computations are performed on minibatches (128 in all experiments), yielding small kernel matrices and inexpensive decomposition steps. As a result, the computational overhead is comparable to standard VAE training. For higher-dimensional latent spaces or larger latent sets, the framework supports sparse GP approximations, such as inducing-point or structured, kernel methods—which reduce complexity to $\mathcal{O}(m^3)$ with $m \ll n$. These approximations do not affect our identifiability results, which rely on proxy structure rather than the specific GP implementation.

Notably, in standard Gaussian Process regression, the full kernel matrix across all training inputs is essential, and minibatching is nontrivial. However, our model employs an *interval-valued GP prior*, which is conceptually and mathematically distinct from the posterior inference typically associated with GP regression.

Specifically, as defined in Equation (10), our prior imposes a structured regularization over the encoder output $U \sim q(U \mid Z) = \mathcal{N}(\mu_u(Z), \sigma_u^2(Z))$, treating each latent variable u_i as lying within the interval $[\mu_u^i - \sigma_u^i,\ \mu_u^i + \sigma_u^i]$. The GP prior then computes the marginal likelihood of each interval under the GP prior $U(Z) \sim \mathcal{GP}(0, K(Z, Z))$, and regularizes the encoder by maximizing the total log probability that each u_i lies within this interval. Importantly, the resulting prior term, $\log p_{\text{GP}}\left(u^i \in [\mu_u^i \pm \sigma_u^i] \mid Z^i\right)$, depends only on the marginal GP distribution at each point Z^i, i.e., $\mathcal{N}(0, k(Z^i, Z^i))$, and not on cross-point correlations. This makes the prior computation fully local, univariate, and independent across data points, thereby fully compatible with minibatch training. As a result, training complexity is significantly reduced compared to standard GP methods. As such, the resulting regularizer is fully decoupled across samples.

This form of prior emerges when the latent variable is represented not as a point estimate, but as an interval-valued random variable. In such cases, the probabilistic constraint that a latent sample $u_i \in [\mu_u^i - \sigma_u^i,\ \mu_u^i + \sigma_u^i]$ under a GP prior corresponds to evaluating the Gaussian probability mass within that interval. i.e., the difference between two CDF values of a univariate normal distribution centered at zero with variance $k(Z^i, Z^i)$. Since this prior only depends on the marginal variance at each Z^i, it avoids the need to evaluate full joint Gaussian densities across data points.

Table 2: Treatment mechanisms and proxy/outcome functions used in synthetic experiments.

	Functions
Proxy Functions	$\{u,\ u^2,\ u^3,\ \sin(u),\ \tanh(u),\ \exp\left(-0.5\,u^2\right)\}$, $\{u,\ \sin(u),\ \tanh(0.5u),\ \exp(0.3u),\ \log(1+u^2),\ \sigma(u)\}$, $\{u,\ u^2,\ \sin(u),\ \tanh(u),\ \exp\left(-0.2\,u^2\right),\ \sin(2u)\}$ $\{u,\ u^3,\ \tanh(u),\ \sigma(u),\ \log(1+u^2),\ \sin(u)\}$ $\{u,\ \tanh(0.5u),\ \sin(u),\ \exp\left(-0.3\,u^2\right),\ \sigma(u),\ u^2\}$ $\{u,\ \log(1+e^u),\ \exp(0.5u),\ \tanh(u),\ \sigma(u),\ \log(1+u^2)\}$
Treatment Functions	$\{\text{Bernoulli}\left(\sigma(0.8u)\right)\}$, $\{\text{Bernoulli}\left(\sigma(0.4u+0.3)\right)\}$
Outcome Functions	$\{0.5u + 0.2\sin(u) + t\,(0.8 + 0.3u) + \varepsilon\}$, $\{0.5u + 0.3\cos(u) + t\,(0.5 + 0.5u) + \varepsilon\}$

This pointwise structure makes the intervalGP prior *naturally compatible with mini-batch training.* The computational complexity is therefore drastically reduced, and the model can scale to large datasets using stochastic optimization. This is particularly important for practical implementation on high-dimensional proxy data Z, and aligns with the goal of enabling uncertainty-aware latent regularization without sacrificing scalability.

We emphasize that this GP prior is used during training for regularization. At test time, we revert to a full posterior GP computation (Equations 11–12), which recovers standard GP properties, such as smoothness, confidence calibration, and extrapolation behavior, by leveraging the full kernel structure. This clarifies that the simplification only affects the training prior, not the inference procedure, ensuring both efficiency and fidelity.

6 Experiments

We conduct both synthetic and semi-synthetic experiments to evaluate the effectiveness of our proposed method. All experiments were conducted on a laptop running *Windows 11 Home* (version 22H2, build 22631), equipped with a 13th Gen Intel® Core™ i9-13900H processor (14 cores, 20 threads, 2.6 GHz), 32 GB of RAM, and a 1 TB SSD. We compare *IntervalGP-VAE* primarily against **TEDVAE** (Zhang et al., 2021), which outperforms a range of state-of-the-art methods, such as **CEVAE**(Louizos et al., 2017).

Although TEDVAE assumes unconfoundedness, the structural graph in Figure 1 of **TEDVAE** (Zhang et al., 2021) includes the backdoor path $T \leftarrow Z_c \rightarrow Y$, implying that causal identification requires the observed covariates to contain a complete and noise-free representation of the latent confounder, a condition assumed but not analysed in TEDVAE. In contrast, our framework provides explicit conditions under which the latent confounder becomes identifiable from proxy variables, thereby inducing unconfoundedness. For this reason, comparing our method with TEDVAE in settings where our identifiability conditions hold is appropriate and reflects the intended operational regimes of both methods.

We assess performance using two standard metrics: the ATE error between the estimated and true ATE, where the ATE is computed as the expectation of ITE

$$\widehat{\text{ATE}} = \frac{1}{m}\sum_{j=1}^{m}\hat{\tau}_j, \qquad \text{ATE-error} = \left|\widehat{\text{ATE}} - \text{ATE}_{\text{true}}\right|. \tag{14}$$

and *Precision in Estimation of Heterogeneous Effect* (PEHE), which evaluates at individual levels,

$$\text{PEHE} := \sqrt{\frac{1}{m}\sum_{j=1}^{m}(\hat{\tau}_j - \tau_j)^2}, \tag{15}$$

both of which are commonly used in the causal inference literature (Hill, 2011; Shalit et al., 2017; Louizos et al., 2017; Yao et al., 2018). In addition, we also report the coverage rate of the estimated ITE intervals, a distinguishing feature that quantifies its ability to capture uncertainty in counterfactual predictions. Model

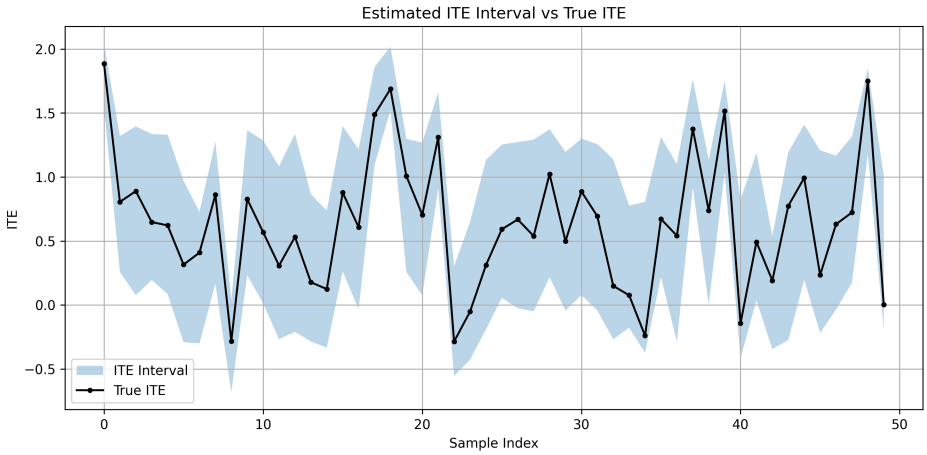

Figure 2: ITE Intervals (Combination 12)

training follows a staged strategy: joint training of the encoder, decoder, and causal head for $E = 200$ epochs using the Adam optimizer with a batch size of 128 and learning rate of 10^{-3}. The implementation of *IntervalGP-VAE* uses the following parameters: latent dimension is 1, hidden layer width is 64, GP lengthscale $\ell = 0.4$, GP variance $\sigma_f^2 = 5.0$, and GP noise variance $= 10^{-4}$.

6.1 Synthetic Experiments

We evaluate the methods on 24 ($6 \times 2 \times 2$) synthetic settings combining different proxy functions, binary treatment mechanisms, and outcome functions, as summarised in Table 2. These settings are deliberately constructed to satisfy the identifiability conditions of Theorem 1. In this identifiable regime, *IntervalGP-VAE* behaves as expected: it recovers the latent confounder up to an invertible transformation and propagates uncertainty through the causal decoder, producing calibrated ITE intervals rather than collapsing to overconfident point estimates. Figure 2 illustrates the estimated 90% confidence intervals for the ITE in one synthetic replicate (Combination 12). The near-nominal coverage of these 90% intervals demonstrates the model's ability to enforce latent-space coherence through the GP prior and to propagate proxy-level uncertainty into reliable ITE estimates. The average training time of the *IntervalGP-VAE* method per synthetic dataset is approximately 173.9s. As shown in Figure 3, *IntervalGP-VAE* consistently achieves lower average PEHE (orange dashed line) than *TEDVAE* (blue dashed line), while maintaining comparable ATE error. The PEHE reflects individual-level accuracy, whereas the ATE error reflects only the mean bias of the ITE estimates. Therefore, a model can achieve low ATE error while having a large PEHE. Notably, *IntervalGP-VAE* attains a coverage rate over 95%, indicating that its predicted ITE intervals are well-calibrated with respect to the 90% confidence

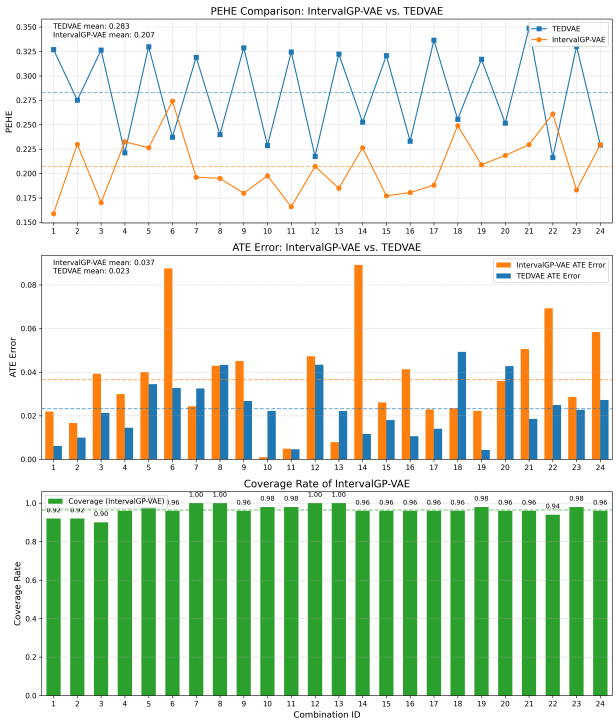

Figure 3: Experimental results on the synthetic dataset.

level. These results align with our theoretical claims: when the proxy functions satisfy the identifiability conditions of Theorem 1, the latent confounder becomes recoverable up to an equivalence class, enabling accurate estimation of individual treatment effects.

6.2 Semi-synthetic Experiments

Beyond the synthetic settings, we conduct a 100-replication study on the semi-synthetic IHDP benchmark dataset (Hill, 2011). In this setting, the observed covariates Z are treated as proxies for latent socio-demographic confounders U, providing a realistic yet evaluable testbed with ground-truth counterfactuals. To reduce variance and ensure reproducibility, we use only the continuous covariates specified by the IHDP index. We follow the standard evaluation protocol: train on the prescribed training split and assess on the held-out test split across multiple realizations, reporting PEHE, ATE error, and the empirical coverage of the 90% confidence intervals. The results are presented in Figure 4. The average training time of the *IntervalGP-VAE* method on the IHDP dataset was approximately 366.9 s. On the IHDP benchmark, *IntervalGP-VAE* achieves comparable performance in terms of PEHE and ATE error, relative to *TEDVAE*. However, the empirical coverage rates of the 90% confidence intervals are lower than those observed on the synthetic datasets (53% vs 95%). This discrepancy can be theoretically explained by the identifiability conditions outlined in Theorem 1. Specifically, the synthetic datasets are intentionally constructed to satisfy the minimum sufficient conditions for identifying the latent confounder via multiple complementary proxy variables, enabling reliable latent recovery. In contrast, the proxy mechanism in the IHDP dataset may violate these requirements, such as lacking nonlinear and nonredundant proxy functions, thus weakening the model's ability to infer the true latent structure. Moreover, *IntervalGP-VAE* assumes conditional independence and variability of the proxy functions. If these assumptions are violated, both posterior inference and interval calibration may degrade. TEDVAE attains lower PEHE and ATE error on IHDP, which is consistent with its unconfoundedness assumption: when the observed covariates themselves encode most of the confounding information, TEDVAE can perform well even without formal identifiability guarantees. By contrast, IntervalGP-VAE is intended for settings where the confounder is genuinely unobserved and is recoverable only under multiple complementary proxy conditions—conditions that IHDP does not satisfy. Our contribution is therefore not universal performance improvement but providing explicit identifiability conditions and principled interval-based uncertainty quantification, which existing VAE-based causal models typically lack.

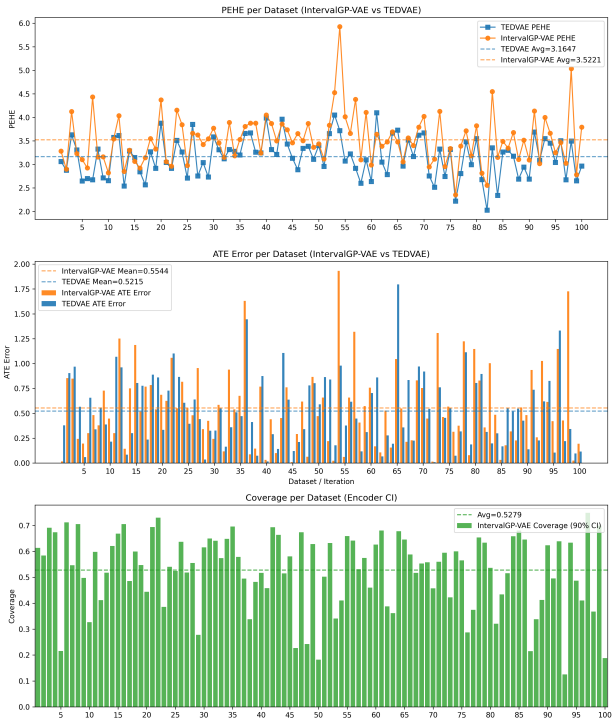

Figure 4: Experimental results on the IHDP dataset.

7 Conclusion

This work demonstrates that *IntervalGP-VAE* provides a principled framework for causal inference in the presence of unobserved confounding. By combining latent variable modelling with interval-valued Gaussian process inference, the framework enables both recovery of latent confounders from noisy proxy measurements and coherent quantification of the uncertainty that such recovery inevitably entails. In contrast to approaches that produce only point estimates of individual treatment effects (ITEs), our model provides calibrated ITE

intervals, thereby explicitly representing uncertainty arising from latent confounding and imperfect proxy measurements.

A central contribution of this work lies in the identifiability analysis. We show that, under structural assumptions on the proxy generation mechanism, the latent confounder can be recovered up to a smooth invertible transformation. Importantly, we further establish that counterfactual treatment effects remain invariant under such transformations, ensuring that causal conclusions are preserved despite this residual ambiguity. These results provide theoretical justification for proxy-based causal inference and clarify the conditions under which unmeasured confounding can be addressed using noisy proxy variables. Nevertheless, a limitation of our framework is that these identifiability guarantees rely on structural assumptions about the proxy generation mechanism which, in practice, may be challenging to verify. Developing data-driven methods to test or relax these assumptions remains an important direction for future work.

Empirically, the results support our central hypothesis: integrating VAE-based latent representation learning with interval-valued GP inference yields reliable and uncertainty-aware estimation of ITEs when the proxies satisfy the identifiability conditions. Experiments on synthetic and semi-synthetic benchmarks demonstrate that the proposed approach produces well-calibrated uncertainty intervals while maintaining competitive estimation accuracy.

Several promising directions follow from this work. Extending the framework to multivariate or structured latent confounders would broaden its applicability in more complex causal systems. Incorporating temporal or spatial structures could further enable counterfactual reasoning in longitudinal health, education, or environmental studies. In addition, improving robustness to incomplete or partially informative proxies—through adaptive kernel learning or causal feature selection—may enhance the applicability of proxy-based causal inference in real-world observational datasets. Finally, connecting interval-valued ITE estimates with decision-theoretic objectives, such as risk-sensitive treatment recommendation or fairness-aware policy optimisation, offers a promising direction for developing uncertainty-aware causal models that directly support practical decision making. A limitation of our framework is that the identifiability guarantees rely on structural assumptions about the proxy generation mechanism. In practice, verifying these assumptions may be challenging and remains an important direction for future work.

References

Elizabeth S Allman, Catherine Matias, and John A Rhodes. Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, 37(6A):3099–3132, 2009.

Francesco Paolo Casale, Adrian V. Dalca, Luca Saglietti, Jennifer Listgarten, and Nicolo Fusi. Gaussian process prior variational autoencoders. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 10390–10401. Curran Associates, Inc., 2018.

Vincent Fortuin, Dmitry Baranchuk, Gunnar Rätsch, and Stephan Mandt. GP-VAE: Deep probabilistic time series imputation. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 1651–1661, 2020.

Daniele Gammelli, Inon Peled, Filipe Rodrigues, Dario Pacino, Haci A. Kurtaran, and Francisco C. Pereira. Estimating latent demand of shared mobility through censored gaussian processes. *Transportation Research Part C: Emerging Technologies*, 120:102775, 2020.

Daniele Gammelli, Kasper Pryds Rolsted, Dario Pacino, and Filipe Rodrigues. Generalized multi-output gaussian process censored regression. *Pattern Recognition*, 129:108751, 2022.

Shonosuke Harada and Hisashi Kashima. Infocevae: Treatment effect estimation with hidden confounding variables matching. *Machine Learning*, 113:1799–1817, 2024.

Tobias Hatt and Stefan Feuerriegel. Sequential deconfounding for causal inference with unobserved confounders. In *Proceedings of the Third Conference on Causal Learning and Reasoning (CLeaR)*, volume 236 of *Proceedings of Machine Learning Research (PMLR)*, pp. 934–956, 2024.

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

Aapo Hyvärinen, Hiroaki Sasaki, and Richard E. Turner. Nonlinear ica using auxiliary variables and generalized contrastive learning. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, pp. 859–868. PMLR, 2019.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvärinen. Variational autoencoders and nonlinear ica: A unifying framework. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, pp. 2207–2216. PMLR, 2020.

Joseph B. Kruskal. Three-way arrays: Rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear Algebra and Its Applications*, 18:95–138, 1977.

Manabu Kuroki and Judea Pearl. Measurement bias and effect restoration in causal inference. *Annals of Mathematics and Artificial Intelligence*, 72(1):81–102, 2014.

Milan Kuzmanovic, Tobias Hatt, and Stefan Feuerriegel. Deconfounding temporal autoencoder: Estimating treatment effects over time using noisy proxies. In *Proceedings of the First Machine Learning for Health (ML4H 2021)*, pp. 143–155. PMLR, 2021.

Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Rich Zemel, and Max Welling. Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 6446–6456, 2017.

Wang Miao, Zhi Geng, and Eric J. Tchetgen Tchetgen. Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, 105(4):987–993, 2018.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2 edition, 2009.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, 2017.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA, 2006.

Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3076–3085. PMLR, 2017.

Sam Witty, Kenta Takatsu, David Jensen, and Vikash Mansinghka. Causal inference using gaussian processes with structured latent confounders. In *Proceedings of the 37th International Conference on Machine Learning (ICML 2020)*, pp. 10313–10323. PMLR, 2020.

Feilong Wu, Kejing Yin, and William K. Cheung. An end-to-end learning approach for counterfactual generation and individual treatment effect estimation. In *Proceedings of the 2024 IEEE Conference on Artificial Intelligence (CAI)*. IEEE, 2024.

Pengzhou Wu and Kenji Fukumizu. Towards principled causal effect estimation by deep identifiable models. *arXiv preprint arXiv:2109.15062*, 2021.

Liyuan Xu, Heishiro Kanagawa, and Arthur Gretton. Deep Proxy Causal Learning and its Application to Confounded Bandit Policy Evaluation. In *Proceedings of the 35th International Conference on Neural Information Processing Systems (NeurIPS 2021)*, pp. 26264–26275, Virtual Conference, 2021. Curran Associates, Inc.

Liuyi Yao, Sheng Li, Yaliang Li, Mengdi Huai, Jing Gao, and Aidong Zhang. Representation learning for treatment effect estimation from observational data. In *Advances in Neural Information Processing Systems*, pp. 2638–2648, 2018.

Weijia Zhang, Lin Liu, and Jiuyong Li. Treatment Effect Estimation with Disentangled Latent Factors. In *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21)*, pp. 10923–10930, Virtual Conference, 2021. AAAI Press.

Shengjia Zhao, Jiaming Song, and Stefano Ermon. Infovae: Balancing learning and inference in variational autoencoders. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 5885–5892. AAAI Press, 2019.

Appendix

A Proof of Theorem 1

Theorem 1 (Identifiability of the latent confounder from noisy proxies)

Consider the proxy–latent relationship defined in the structural equation (Eq. 4), and suppose the proxy variables satisfy the assumptions of conditional independence (Eq. 5) and variability (Eq. 6). If the total number of proxies satisfies

$$k \;\geq\; 2d + 4,$$

then the latent confounder $U \in \mathbb{R}^d$ is identifiable from the observed proxy distribution $p(Z)$ up to a smooth, invertible reparameterisation.

Proof. We establish identifiability through a discretisation argument combined with Kruskal's theorem for finite mixtures (Kruskal, 1977; Allman et al., 2009), followed by a refinement limiting argument to recover the continuous model.

Step 1: Minimal discretisation size $r \geq d+1$. Since $U \in \mathbb{R}^d$, the conditional proxy form a d-dimensional differentiable manifold

$$\mathcal{M} := \{\, p(Z \mid U = u) : u \in \mathbb{R}^d \,\}$$

A discretization with r latent states selects representatives $\{u_\ell\}_{\ell=1}^r$. To preserve the d degrees of freedom of $\mathcal{M}$, this requires at least $d+1$ points, hence $r \;\geq\; d+1$.

From conditional independence $Z_i \perp Z_j \mid U$, we have in the discretization:

$$p(Z_i, Z_j \mid \tilde{U} = \ell) = p(Z_i \mid U = u_\ell)\, p(Z_j \mid U = u_\ell),$$

By the variability condition (Eq. 6), whenever $u_\ell \neq u_{\ell'}$, we have for many i.

$$p(Z_i \mid U = u_\ell) \neq p(Z_i \mid U = u_{\ell'})$$

Step 2: Partitioning proxies. We split the k proxies into three nonempty groups L_1, L_2, L_3 such that $|L_j| \leq r$. Since $|L_1| + |L_2| + |L_3| = k$, the bounds $|L_j| \leq r$ imply

$$2r + 2 \leq k \leq 3r,$$

so a valid partition exists whenever $k \geq 2r + 2$.

Step 3: Conditional matrices and Kruskal ranks. For each block L_j, define

$$M_j(\ell, m) := \mathbb{P}(\tilde{Z}_{L_j} = m \mid \tilde{U} = \ell) \in \mathbb{R}^{r \times n_j},$$

where m indexes joint outcomes of $\tilde{Z}_{L_j}$.

By conditional independence,

$$M_j(\ell, m) = \prod_{i \in L_j} \mathbb{P}(\tilde{Z}_i = m_i \mid \tilde{U} = \ell).$$

For each $i \in L_j$, define the per-proxy conditional matrix

$$A_i(q, \ell) := \mathbb{P}(\tilde{Z}_i = q \mid \tilde{U} = \ell).$$

Then, viewing $M_j^\top$ columnwise, we have

$$M_j^\top = \bigodot_{i \in L_j} A_i,$$

i.e. $M_j^\top$ is the Khatri–Rao (columnwise Kronecker) product of the $\{A_i\}_{i\in L_j}$.

By the variability assumption (Eq. 6), the columns of each A_i vary generically across latent states, implying $k(A_i) \geq 2$ for many i. Standard rank bounds for Khatri–Rao products then give

$$k(M_j) = k(M_j^\top) \;\geq\; \min\left\{ r,\; \sum_{i\in L_j} k(A_i) - (|L_j| - 1) \right\}.$$

Since $|L_j| \leq r$, this yields $k(M_j) \geq |L_j|$, and therefore

$$k(M_j) = |L_j| = \min(|L_j|, r).$$

Step 4: Identifiability at the discrete level. By Kruskal's theorem (Kruskal, 1977; Allman et al., 2009), the finite-mixture representation

$$p(\tilde{Z}_{L_1} = m_1, \tilde{Z}_{L_2} = m_2, \tilde{Z}_{L_3} = m_3) = \sum_{\ell=1}^{r} \pi_\ell \, M_1(\ell, m_1)\, M_2(\ell, m_2)\, M_3(\ell, m_3),$$

with mixing weights $\pi_\ell = p(\tilde{U} = \ell)$, has a unique decomposition (up to permutation of the latent states) whenever

$$k(M_1) + k(M_2) + k(M_3) = |L_1| + |L_2| + |L_3| = k \;\geq\; 2r + 2.$$

Thus the discrete factors $\{\pi_\ell, M_1, M_2, M_3\}$ are identifiable at each finite resolution m.

Step 5: Refining the discretisation (discrete → continuous). Consider a sequence of discretisations indexed by m, where the discretizations become arbitrarily fine as $m \to \infty$. Let $(M_j^{(m)}, \pi^{(m)})$ denote the identifiable discrete factors at resolution m.

(i) Convergence of conditionals. By definition of $M_j^{(m)}$ as averages of $p(Z_{L_j} \mid U = u)$ over the discretizations, the Lebesgue differentiation theorem yields

$$M_j^{(m)}(\ell, \cdot) \;\to\; p(Z_{L_j} \mid U = u_\ell^{(m)}) \quad \text{in } L^1.$$

(ii) Convergence of mixing weights. Since $\pi_\ell^{(m)}$ is the mass of $p(U)$ on the ℓ-th latent discretization, the discrete measures $\sum_\ell \pi_\ell^{(m)} \delta_{u_\ell^{(m)}}$ converge weakly to $p(U)\, du$.

(iii) Preservation of uniqueness. All discretisations at level m produce the same coarse-grained distribution of Z, hence the same discrete factorisation (up to permutation). Taking $m \to \infty$ and using (i)–(ii) yields a unique continuous limit.

(iv) Stability. Because the factorisation is stable under L^1 perturbations, the limiting conditional densities and mixing distribution are uniquely determined by $p(Z)$.

Step 6: Recovering the latent manifold. The full collection of conditionals $p(Z \mid U = u)$ is identified from $p(Z)$ up to a reparameterisation of u. Under the structural equation (Eq. 4), $p(Z \mid U = u) = p_\epsilon(Z - g(u))$, so the noise distribution and the image manifold $\{g(u)\}$ are determined up to a smooth change of coordinates. Any alternative latent variable U' producing the same conditionals must therefore satisfy $U' = h(U)$ for some smooth invertible h. By the variability condition (Eq. 6), the map $u \mapsto p(Z \mid U = u)$ is injective, so U (and any function of it) is identifiable from $p(Z)$ up to smooth reparameterisation.

Step 7: Conclusion. Since $r \geq d + 1$ is necessary for identifiability (Step 1), and $k \geq 2r + 2$ guarantees a valid proxy partition and satisfies Kruskal's condition, the smallest sufficient choice is $r = d + 1$, giving the requirement

$$k \;\geq\; 2(d + 1) + 2 \;=\; 2d + 4.$$

Hence the latent confounder U is identifiable from $p(Z)$ up to smooth invertible reparameterisation. □

B Proof of Theorem 2

Theorem 2 (ITE invariance under latent reparameterisation)

Let $h : \mathbb{R}^d \to \mathbb{R}^d$ be a smooth and invertible transformation, and define the reparameterised latent variable $\hat{U} = h(U)$. Then the structural individual treatment effect defined in Eq. 3 is invariant under such reparameterisations:

$$\tau_j(\hat{U}_j, Z_{Y,j}) \;=\; \tau_j(U_j, Z_{Y,j}) \qquad \text{for all individuals } j.$$

Proof. Let $\hat{U}_j = h(U_j)$ for some diffeomorphism h. Since h is invertible, $U_j = h^{-1}(\hat{U}_j)$. Evaluating the structural equation (Eq. 1) at the same exogenous noise realisation $\epsilon_{Y,j}$ yields

$$f(t, U_j, Z_{Y,j}) = f(t, h^{-1}(\hat{U}_j), Z_{Y,j}), \qquad t \in \{0, 1\}.$$

Thus by definition of the potential outcomes in Eq. 3,

$$\begin{aligned}\tau_j(\hat{U}_j, Z_{Y,j}) &= f(1, h^{-1}(\hat{U}_j), Z_{Y,j}) - f(0, h^{-1}(\hat{U}_j), Z_{Y,j}) \\ &= f(1, U_j, Z_{Y,j}) - f(0, U_j, Z_{Y,j}) = \tau_j(U_j, Z_{Y,j}).\end{aligned}$$

Hence the ITE is invariant under smooth invertible reparameterisations of the latent space. □

Remark 3 (Role of Theorem 2). The latent confounder U is identifiable from proxy variables only up to a smooth and invertible reparameterisation. Theorem 2 shows that the structural ITE defined in Eq. 3 is invariant under all smooth invertible transformations of the latent space, and therefore remains uniquely determined even when U is identifiable only up to an equivalence class.

Consequently, Theorem 2 ensures that the equivalence class of valid latent representations all yield the same causal conclusions.

C Proof of Proposition 1

Proposition 1 (Compatibility of GP priors with identifiability and latent reparameterisations)

Let $\hat{U} = (\hat{U}_1, \ldots, \hat{U}_n)$ denote the latent confounder values inferred from the proxy variables $Z^1, \ldots, Z^n$ under the structural proxy model (Eq. 4). Suppose a Gaussian Process prior

$$\hat{U} \sim \mathcal{GP}\big(0, K(Z^i, Z^j)\big)$$

is placed on the inferred latent representation, where K is any positive-definite kernel on the proxy space.
Then:

1. Introducing a GP prior on $\hat{U}$ does not alter the identifiability of the individual treatment effect (ITE) established in Theorems 1.
2. For any smooth and invertible reparameterisation $h : \mathbb{R}^d \to \mathbb{R}^d$, replacing U by $\tilde{U} = h(U)$ does not affect the role of K as a covariance function over the latent representation.

Proof. (1) Identifiability in Theorem 4 is derived solely from the conditional proxy mechanism $p(Z \mid U)$ implied by the structural equation (Eq. 4) together with the assumptions of conditional independence (Eq. 5) and variability (Eq. 6). The observational law admits the mixture representation

$$p(Z, T, Y) \;=\; \int p(Z, T, Y \mid U)\, p(U)\, dU,$$

where $p(U)$ acts only as the mixing distribution. Replacing $p(U)$ by any alternative prior $\tilde{p}(U)$, such as a GP prior, changes the mixing weights but leaves the conditional model $p(Z, T, Y \mid U)$, and hence all structural relationships between Z, T, Y and U, unchanged. Therefore the identifiability of U and of the ITE from $p(Z, T, Y)$ is unaffected by the choice of latent prior.

(2) Suppose, as in Theorem 1, that the latent confounder is identifiable only up to a smooth and invertible mapping $h : \mathbb{R}^d \to \mathbb{R}^d$, so that any observationally equivalent model may replace U by $\tilde{U} = h(U)$. A Gaussian Process prior specifies the covariance structure

$$\mathrm{Cov}(U_i, U_j) \;=\; K(Z^i, Z^j),$$

which depends only on the proxy inputs (Z^i, Z^j) and not on the coordinate representation of U. Under the reparameterisation $U \mapsto \tilde{U} = h(U)$, smooth invertibility of h implies

$$\|U_i - U_j\| \to 0 \quad \Longleftrightarrow \quad \|\tilde{U}_i - \tilde{U}_j\| \to 0,$$

so U and $\tilde{U}$ induce exactly the same neighbourhood structure in the latent space. The kernel $K(Z^i, Z^j)$ therefore remains a valid covariance function for both U and $\tilde{U}$, and the GP prior imposes the same geometric regularisation on either parameterisation.

Hence the GP prior is fully compatible with the equivalence class $\{h(U)\}$ of latent representations and does not restrict the reparameterisation freedom assumed in the identifiability analysis. □

Remark 4 (Latent-space coherence and uncertainty propagation). The GP prior induces geometric coherence in the learned latent space, ensuring that similar proxy observations correspond to similar inferred latent values. This regularisation stabilises the inverse problem $Z \mapsto U$ once identifiability has been established. In addition, the interVAlGP posterior enables interval-valued uncertainty quantification, allowing uncertainty in the proxies to propagate in a principled manner to the ITE estimates—an essential capability when the proxies are noisy.

Remark 5 (Flexibility of kernel design). Proposition 1 does not require a kernel tailored to the structural functions g_i. The kernel governs coherence in the latent inference space rather than modelling properties of g_i. While an RBF kernel is used as a numerically stable default, alternative kernels (e.g. Matérn, rational quadratic, or non-stationary kernels) may be employed without affecting identifiability guarantees or the conclusions of Proposition 1.

D IntervalGP-VAE Architecture Details

This appendix provides additional details of the IntervalGP-VAE architecture and its training procedure, complementing the discussion in Section 5. The model is composed of three neural networks (an encoder, a decoder, and an outcome head) together with two Gaussian Process (GP) components: (1) an interval-valued GP prior over the latent confounder, and (2) an IntervalGP head for ITE and its uncertainty. All neural components are implemented as multilayer perceptrons. The VAE encoder and decoder use a hidden width of 64, while the causal prediction head uses a hidden width of 32.

D.1 Encoder

Given a proxy vector $z \in \mathbb{R}^k$, the encoder outputs the parameters of a factorised Gaussian posterior:

$$q_\phi(u, \varepsilon \mid z) = \mathcal{N}(u; \mu_u(z), \sigma_u^2(z))\, \mathcal{N}(\varepsilon; \mu_\varepsilon(z), \sigma_\varepsilon^2(z)).$$

The pair (μ_u, σ_u) induces an interval estimate

$$u \in [\, \mu_u(z) - \sigma_u(z),\ \mu_u(z) + \sigma_u(z)\,].$$

D.2 Decoder

The decoder reconstructs proxies using an additive structure,

$$\hat{z} = g_{\text{dec},\theta}(u) + \varepsilon,$$

where $g_{\text{dec},\theta}$ is an MLP applied to u.

D.3 Causal/outcome head

Given the latent confounder u, treatment $t \in \{0, 1\}$, optional covariates z_Y, the causal head models

$$y = f_\psi(u, t, z_Y),$$

enabling estimation of $Y(0)$, $Y(1)$, and ITE $= Y(1) - Y(0)$.

D.4 IntervalGP Prior over the Latent Confounder

We place an IntervalGP prior over the latent confounder as a function of the proxies,

$$U(Z) \sim \mathcal{GP}(0, k(Z, Z')),$$

using an RBF kernel k. The GP takes each proxy vector Z_i as input and uses the encoder-derived interval

$$u_i \in [\mu_{u,i} - \sigma_{u,i},\ \mu_{u,i} + \sigma_{u,i}]$$

as its supervision signal. This provides a smooth latent representation while retaining uncertainty from the encoder.

For a new proxy vector Z^*, the GP posterior

$$p(u^* \mid Z^*, \mathcal{D})$$

is evaluated using Eq. 11, yielding a predictive mean and variance. These are turned into an interval

$$[u^{\text{lower}}(Z^*),\ u^{\text{upper}}(Z^*)],$$

which serves as the uncertainty-aware estimate of the latent confounder for downstream counterfactual prediction.

D.5 IntervalGP Regression for Counterfactual ITE Intervals

To obtain uncertainty-aware estimates of individual treatment effects, we apply a second IntervalGP model in the ITE space. This GP takes the proxy vector Z_i as input, and uses ITE interval

$$\tau_i \in \left[\tau_i^{\text{lower}},\ \tau_i^{\text{upper}}\right],$$

as its target. These interval targets arise from propagating the latent-confounder uncertainty $[u_i^{\text{lower}}, u_i^{\text{upper}}]$ through the outcome model.

For a new proxy vector Z^*, the IntervalGP computes the posterior distribution

$$p(\tau^* \mid Z^*, \mathcal{D}),$$

as given in Eq. 12. The predictive mean and variance are then converted into an ITE interval,

$$\left[\tau^{\text{lower}}(Z^*),\ \tau^{\text{upper}}(Z^*)\right],$$

which forms the final uncertainty-aware counterfactual treatment effect estimate.

D.6 Training Objective

We optimise the following loss:

$$\mathcal{L}_{\text{IntervalGP-VAE}} = \underbrace{\mathbb{E}_{q_\phi(u,\varepsilon|z)}\left[\|z - g_\theta(u,\varepsilon)\|_2^2\right]}_{\text{Proxy reconstruction loss}} + \lambda_y \underbrace{\mathbb{E}_{q_\phi(u|z)}\left[(y - f_\psi(u,t,z_y))^2\right]}_{\text{Outcome/causal prediction loss}} \tag{16}$$

$$+ \beta_\varepsilon \underbrace{D_{\text{KL}}\big(q_\phi(\varepsilon \mid z) \,\big\|\, \mathcal{N}(0,I)\big)}_{\text{KL on noise } \varepsilon} - \gamma \underbrace{\sum_{i=1}^{n} \log p_{\text{GP}}(u_i \in [\mu_{u,i} \pm \sigma_{u,i}] \mid z_i)}_{\text{IntervalGP prior over } U}. \tag{17}$$

This objective captures the core components of our framework: proxy reconstruction, outcome prediction, a VAE prior on the noise ε, and the IntervalGP prior enforcing structural constraints on the latent confounder U. In all reported experiments, we set

$$\lambda_y = 1, \quad \beta_\varepsilon = 1, \quad \gamma = 1.$$