UnCoVAEr: Estimating Causal Concept Effects under Visual Latent Confounding

Anonymous authors
Paper under double-blind review

Abstract

Estimating the effect of human-interpretable concepts on model predictions is crucial for explaining and auditing machine learning systems, as well as for mitigating their reliance on spurious correlations. Most existing approaches assume complete concept annotations, but in practice some concepts may remain unobserved and act as confounders, biasing causal effect estimates. We introduce **UnCoVAEr** (Unobserved Confounding Variational AutoEncoder), a latent-variable model that partitions image latent representations into confounder-related and non-confounding residual components. This allows us to (i) identify which observed concepts are confounded, (ii) obtain corrected unbiased effect estimates via backdoor adjustment, and (iii) learn confounder-proxy variables that align with underlying latent factors. On a controlled semi-synthetic MorphoMNIST benchmark, we show that UnCoVAEr yields substantially less biased effect estimates than prior methods, providing practitioners with a practical tool for trustworthy concept-level causal inference in partially annotated image datasets.

1 Introduction

Human-interpretable visual concepts are being increasingly used to explain, audit and control the behavior of machine learning models. Concept-based explanations enable domain experts to ask and answer targeted causal questions such as “how much does hippocampal atrophy, as seen on an MRI, contribute to an Alzheimer’s diagnosis?” (Castro et al., 2020) or “which facial attributes drive perceived attractiveness in our annotated dataset?” (Lingenfelter et al., 2022). In practice, however, causal statements at the concept level are only as valid as the assumptions that underlie them. Most existing concept effect estimators implicitly assume that we have measured all relevant visual factors that confound concept-label relations. When important factors are missing from the annotation set, naive observational estimates can be severely biased and lead practitioners to mistaken conclusions and harmful interventions.

Consider a scenario in medical imaging where an interpretable concept-based model concludes that hippocampal atrophy is a dominant predictor of Alzheimer’s disease. However, this association may be inflated by confounders such as scanner hardware or acquisition protocols: different scanners alter image appearance in ways that affect how atrophy is manifested and measured, while also correlating with hospital site and diagnostic practices that influence the diagnosis. As a result, a naive estimate of the effect of hippocampal atrophy on diagnosis may capture site- or device-specific artifacts rather than a genuine biological causal effect, potentially misleading auditors and downstream clinical decisions. Next, consider a facial-attribute dataset where a set of annotated concepts such as smiling, makeup or age are used to predict attractiveness. Yet, unannotated factors like skin tone, lighting, or demographic imbalance in the dataset can act as confounders, inflating the estimated influence of certain attributes and masking annotator prejudice (Lingenfelter et al., 2022). Obtaining unbiased causal effect estimates enables practitioners and researchers to estimate bias in datasets (Madras et al., 2019; Di Stefano et al., 2020), decide if they need to collect additional metadata or reweight training examples (Zhao et al., 2023). Moreover, obtaining corrected causal effect estimates can be useful to improve the performance of domain generalization methods by penalizing reliance on spurious attributes (Kumar et al., 2023).

Concept-based models are not new: they can be understood as a principled evolution of feature engineering in which model decisions are expressed in terms of semantically meaningful factors

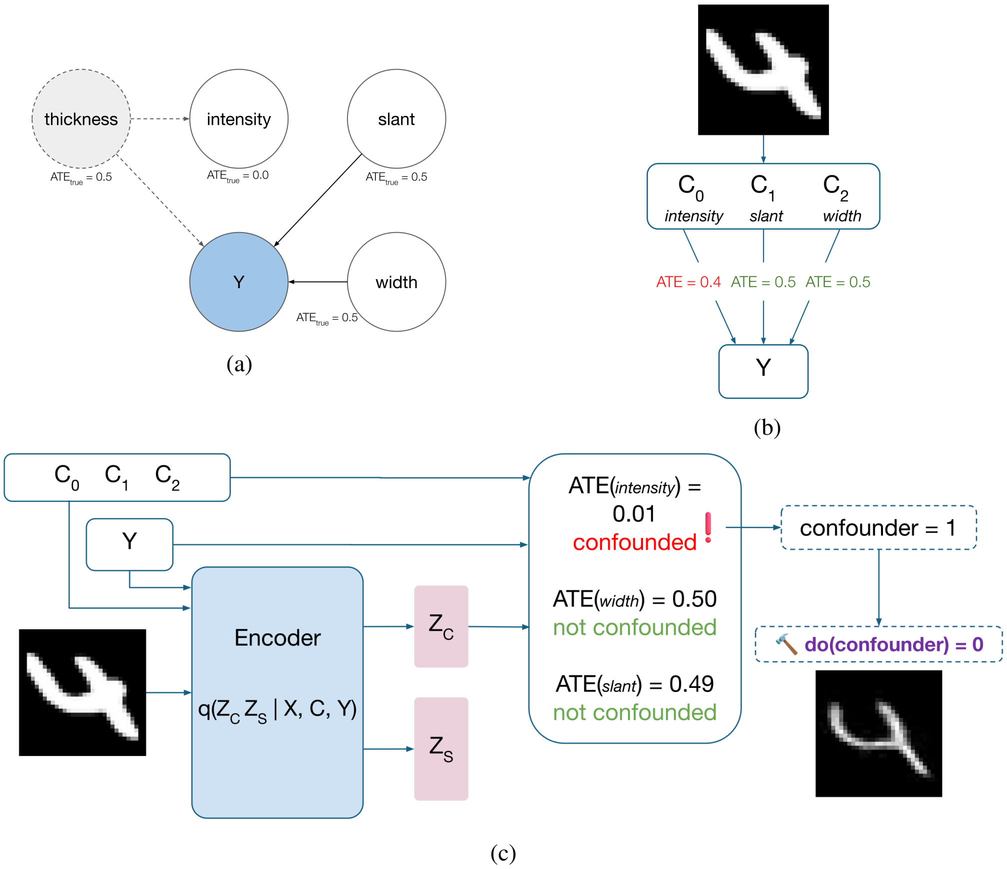

Figure 1: **(a)** The causal graph shows a case where one of our observed concepts C (*intensity*) and our label of interest Y are caused by an unobserved confounder (*thickness*). **(b)** As a result, when computing the effect of *intensity* on Y without taking this confounding into account , we get a wrong estimate. **(c)** Our method *Unobserved Confounding Variational AutoEncoder (UnCoVAEr)* estimates the correct causal effects and finds which observed concepts are confounded. After training our model, we provide C, Y, and the image X to the encoder, which outputs a confounder-related latent Z_C and a non-confounding residual Z_S used only for image reconstruction. We then perform backdoor-adjustment using the learned proxy Z_C to debias the Average Treatment Effect (ATE) estimation. Additionally, we can intervene on the confounder proxy and interpret its effect.

rather than opaque input dimensions. Early work assumed a fixed, complete set of predefined concepts (Koh et al., 2020); subsequent methods allow concepts to be learned post-hoc or discovered from images (Yuksekgonul et al., 2023; Oikarinen et al., 2023; Sawada & Nakamura, 2022; Shang et al., 2024; Rao et al., 2024), and a growing body of research takes a causal formulation on concept models Dominici et al. (2025b;a); Moreira et al. (2024). Separately, a line of work that builds on proximal causal inference (Tchetgen et al., 2020) studies causal question in the presence of latent confounders that manifest through proxies, estimating treatment effects originally in tabular settings (Louizos et al., 2017; Wu & Fukumizu, 2022; Zhang et al., 2020; Miao et al., 2018; Wang & Blei, 2021) and extending to high-dimensional data such as images (Kaddour et al., 2021; Kompa et al., 2022; Israel et al., 2023; Jerzak et al., 2023; Schulte et al., 2025). Our work sits at the intersection of these two threads: we incorporate a deep latent-variable method from the proximal causal inference literature to robustly estimate causal quantities when unobserved visual concepts confound both observed concepts and the label.

We propose **UnCoVAEr**, a latent-variable model inspired by Causal Effect Variational AutoEncoder (CEVAE) that explicitly accounts for concept incompleteness by partitioning the latent space into two parts: a confounder-related component that explains variation shared between concepts and label, and a non-confounding block that captures residual image variation. This structured decomposition lets us (i) recover proxies for the confounders, (ii) use them to estimate concept effects via backdoor-adjustment, and (iii) identify which observed concepts are substantially confounded by latent visual factors. We validate our approach on a controlled semi-synthetic Morpho-MNIST benchmark. UnCoVAEr reduces bias in concept-effect estimates compared to prior concept-based and latent-variable baselines.

Our contributions are as follows: (1) we formalize concept incompleteness as a latent confounding problem in image datasets and introduce partitioned-latent representations as an effective inductive bias; (2) we propose a principled criterion to distinguish confounded from unconfounded concepts and correct their effect estimates via backdoor adjustment; and (3) we provide empirical evidence that UnCoVAEr reduces bias in causal effect estimates compared to strong baselines and learns proxy variables that correlate with underlying latent confounders on a controlled semi-synthetic MorphoMNIST dataset.

2 Related work

Latent-variable proximal causal inference UnCoVAEr builds on a line of work that utilizes deep latent-variable models to estimate causal quantities like ATE in the presence of unobserved confounders. CEVAE (Louizos et al., 2017) assumes a causal graph where latent confounders are also causes of observed proxies and uses a Variational AutoEncoder (VAE) formulation to model the data-generating process and estimate ATE with backdoor adjustment. While CEVAE has demonstrated promising empirical performance, its reliance on variational inference raises concerns about identifiability. Rissanen & Marttinen (2021) provide an extensive critique, showing analytically and experimentally that CEVAE can fail when the latent space is misspecified or when the data distribution is complex, despite working in simple synthetic setups, while they also provide simple experiments on digit images. Follow-up works provide identification under limited overlap assumptions (Wu & Fukumizu, 2022) and disentangle instrumental, risk, and confounding factors to better isolate causal effects (Zhang et al., 2020), while Madras et al. (2019) utilize CEVAE to improve causal effect estimates between sensitive attributes and outcome in a fairness setting. At the same time, proximal causal inference literature (Tchetgen et al., 2020) and related proxy-variable identification results provide formal conditions (completeness / rank) under which proxies identify causal effects (Miao et al., 2018; Wang & Blei, 2021). A number of recent works tackles causal effect estimation assuming that images or image-derived features act as proxies for latent confounders. Some approaches apply standard adjustment ideas by learning a model that extracts confounding information from the image via propensity score matching (Jerzak et al., 2023) or by extracting image features (Xu et al., 2021; Schulte et al., 2025). Others develop neural methods that directly learn the necessary adjustment functions from high-dimensional proxies (Kompa et al., 2022). Kumar et al. (2023) also use the image directly to perform backdoor adjustment and use the estimated causal effects of the attributes to regularize classifiers for domain generalization.

Concept-based explanations Concept-based explanation methods such as TCAV (Kim et al., 2018) or Concept Bottleneck Model (CBM) (Koh et al., 2020) treat concepts as interpretable primitives for explaining image classification, enabling interventions and attributing predictions directly to concepts. Follow-up work has extended CBMs to incorporate concepts not predefined in the concept set (Yuksekgonul et al., 2023; Oikarinen et al., 2023; Sawada & Nakamura, 2022; Shang et al., 2024; Rao et al., 2024), while also revealing important limitations, such as whether the learned concepts truly correspond to human-understandable semantics or instead capture spurious shortcuts (Mahinpei et al., 2021; Margeloiu et al., 2021; Havasi et al., 2022). In this line of work, Bahadori & Heckerman (2021) address biases in concept-based explanations arising from confounding information. They propose a two-stage regression technique, inspired by instrumental variable methods, to remove the impact of confounders and noise. Their approach also considers the completeness of the concept set (Yeh et al., 2020), demonstrating effectiveness even when the set is incomplete. Goyal et al. (2020) introduce the notion of *CaCE (Causal Concept Effect)*, defining it as the effect of the presence or absence of a human-interpretable concept on a deep neural network's prediction. They train a conditional VAE to generate counterfactuals by disentangling and intervening on the concept of interest. While they highlight the importance of causality for concept explanations, they rely on the assumption that unobserved confounders do not significantly impact the observed concepts. Gao & Chen (2024) explicitly tackle concept incompleteness by constructing pseudo-concepts orthogonal to the observed ones and using a linear predictor to capture residual bias. However, their orthogonality assumption does not allow for confounding.

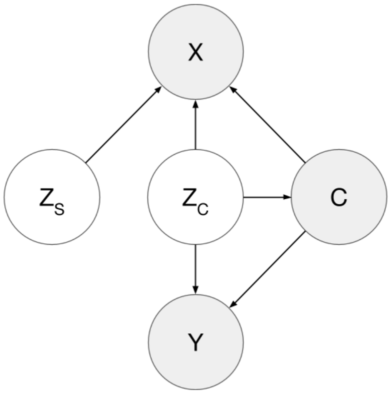

Figure 2: Our causal graph assumption. We assume that the image X is jointly caused by a set of observed concepts C, the unobserved confounder Z_C and some Z_S (e.g. writing style, point of view) that is irrelevant for C and the outcome Y. C cause Y, while the unobserved confounder Z_C causes both C and Y.

3 Preliminaries and Problem Setup

We observe i.i.d. samples $(X, C, Y) \sim \mathcal{D}$, where $X \in \mathcal{X}$ is an image, $C = (C_1, \ldots, C_M) \in \{0,1\}^M$ are observed binary concept annotations, and $Y \in \{0,1\}$ is a binary outcome of interest. We are interested in quantifying how changes to a single concept C_i causally affect Y. Two common estimands for this are the individual (or conditional) treatment effect (ITE) and the population Average Treatment Effect (ATE):

$$\begin{aligned} \mathrm{ITE}_i(x) &:= \mathbb{E}\big[\, Y \mid do(C_i = 1), X = x \,\big] \;-\; \mathbb{E}\big[\, Y \mid do(C_i = 0), X = x \,\big] \\ \mathrm{ATE}_i &= \mathbb{E}\big[\, Y \mid do(C_i = 1) \,\big] \;-\; \mathbb{E}\big[\, Y \mid do(C_i = 0) \,\big]. \end{aligned} \tag{1}$$

Concept-based explanation methods such as CBMs (Koh et al., 2020) perform *concept-interventions* by editing intermediate concept values and re-evaluating the outcome classifier. These types of interventions and the treatment effects they yield are similar to those obtained by meta-learners in the causal inference literature (Künzel et al., 2019). For example, the S-learner, which is one of the simplest methods for treatment effect estimation, performs the same operation: it fits a predictor of Y given covariates and treatment (concepts in this case) and then estimates treatment effects by changing the treatment value in the input. These estimators coincide with the do-intervention only under the ignorability assumption (no unobserved confounding) and can work well enough only in this setting. However, in the presence of unobserved confounders, the estimates will be biased.

Confounding and backdoor adjustment Let V denote an observed variable that jointly causes some observed concepts C and the outcome Y. If V blocks all backdoor paths from C_i to Y, the interventional mean is given by the backdoor formula:

$$\mathbb{E}[Y \mid do(C_i = c)] \;=\; \int_V \mathbb{E}[Y \mid C_i = c, V = v]\, p(v)\, dv. \tag{2}$$

Thus, if V is observed, then Eq. 2 gives an unbiased estimand for the ATE as an immediate consequence of the backdoor-criterion Pearl (1993).

Proxy learning from images In our setup, instead of an observed V, we have a set of unobserved variables U. We assume that the unobserved confounders of interest manifest themselves in the image X (e.g. scanner type, lighting or a separate attribute that affects both annotations and labels). Thus X serves as a high-dimensional proxy for U. Because X also contains many features irrelevant to the causal relation between C and Y, we aim to learn a lower-dimensional proxy latent Z_C from (X, C, Y) with the following operational properties:

(P1) *Adjustment sufficiency:* conditioning on Z_C blocks the backdoor paths between C_i and Y (so Z_C plays the role of V in Eq. 2);

(P2) *Parsimony:* Z_C is low-dimensional and amenable to downstream estimation and marginalization.

(P3) *Interpretability:* We model Z_C as a binary variable to fit well in our setup of binary concepts and outcomes

We assume the causal structure shown in Figure 2. As we want Z_C to contain only the confounding information for adjustment, we decompose the image into two parts: a discrete confounder-specific latent $Z_C \in \{0,1\}^K$ and a continuous residual latent $Z_S \in \mathbb{R}^d$. Intuitively, Z_C captures the visual variation that (partially) explains both some observed concepts C and the label Y, while Z_S contains the remaining image variation that is irrelevant for the causal relationship between C and Y, but necessary to model X accurately.

Under (P1), we have:

$$p(Y \mid X, do(C_i = c)) \;=\; \sum_{Z_C} p(Y \mid X, C_i = c, Z_C)\, p(Z_C \mid X)\;. \tag{3}$$

Thus, to estimate $\mathrm{ITE_i(x)} = \mathbb{E}\big[Y \mid X = x, do(C_i = c)\big]$ and $\mathrm{ATE_i} = \mathbb{E}\big[ITE_i(x)\big]$ we need to learn the conditional $p(Y \mid X, C, Z_C)$[1] and the posterior $p(Z_C \mid X)$.

Identification requirements and limitations Identification of causal effects from observational (X, C, Y) rests on standard proxy-type and overlap assumptions. First, the observed image X must carry *sufficient proxy signal* of the unobserved confounder: if the confounder leaves no detectable trace in pixels, then no method can recover its effect. Second, the support of relevant predictive distributions must overlap (positivity), so that the requisite conditional expectations are well defined. Additionally, we assume that there are no *unobserved colliders*: no latent variables caused by both the outcome Y and a concept C_i. If such a collider exists, conditioning on it opens a spurious path. Under these conditions, and assuming a sufficiently expressive latent-variable model, our adjusted estimator using Z_C is consistent in principle. We offer three clarifications: (i) we do not require recovery of the true confounder, only that the learned proxy Z_C suffices for valid adjustment; (ii) if the confounder leaves no observable imprint, identification is impossible for *any* method; and (iii) while practical estimation is subject to approximation error, our results show that UnCoVAEr recovers unbiased ATEs whenever these identification assumptions hold.

4 UnCoVAEr: Unobserved Confounding Variational AutoEncoder

We now introduce *UnCoVAEr*, a variational autoencoder designed to recover causal concept effects under unobserved confounding.

Generative model The assumed causal graph of Figure 2 leads to the following factorization of the joint distribution:

$$p(X, C, Y, Z_C, Z_S) = p(Z_C)\, p(Z_S)\, p(C \mid Z_C)\, p(X \mid C, Z_C, Z_S)\, p(Y \mid C, Z_C). \tag{4}$$

Our model parameterizes three decoders: $p_{\theta_x}(X \mid C, Z_C, Z_S)$, $p_{\theta_c}(C \mid Z_C)$, and $p_{\theta_y}(Y \mid C, Z_C)$. Because the exact posterior $p(Z_C, Z_S \mid X, C, Y)$ is intractable, we introduce a variational encoder:

$$q_{\phi_c,\phi_s}(Z_C, Z_S \mid X, C, Y) = q_{\phi_c}(Z_C \mid X, C, Y)\, q_{\phi_s}(Z_S \mid X, C, Y).$$

We implement a shared backbone with separate output heads: logits for the discrete confounder-related latent Z_C (parameters ϕ_c) and Gaussian parameters $(\mu_{\phi_s}, \sigma_{\phi_s})$ for the continuous residual latent Z_S. Z_C is sampled with the Gumbel–Softmax relaxation during training (Jang et al., 2017). To reduce information leakage between the two blocks, we additionally include a mutual-information regularizer that minimizes $\mathrm{MI}(Z_C, Z_S)$ using the CLUB estimator with parameters ψ (Cheng et al., 2020). This encourages Z_C to capture confounder-related variation distinct from the residual information in Z_S.

[1] Note that under the assumed causal graph of Figure 2 Y is independent of X given C and Z_C, so $p(Y \mid C, Z_C)$ suffices.

Training objective We maximize the evidence lower bound (ELBO):

$$\begin{aligned}\mathcal{L}_{\text{ELBO}} = \mathbb{E}_{q_{\phi_c,\phi_s}(Z_C,Z_S|X,C,Y)}\big[\log p_{\theta_x}(X \mid C, Z_C, Z_S) + \log p_{\theta_c}(C \mid Z_C) + \log p_{\theta_y}(Y \mid C, Z_C)\big] \\ - \mathrm{KL}\big(q_{\phi_c}(Z_C \mid X, C, Y) \,\|\, p(Z_C)\big) - \mathrm{KL}\big(q_{\phi_s}(Z_S \mid X, C, Y) \,\|\, p(Z_S)\big), \qquad (5)\end{aligned}$$

using the following priors: $p(Z_C) = \prod_{j=1}^{K} \mathrm{Bern}(\pi = 0.5)$ and $p(Z_S) = \prod_{j=1}^{d} \mathcal{N}\big(Z_{S_j} \mid 0, 1\big)$.

We augment $\mathcal{L}_{\text{ELBO}}$ with two auxiliary discriminative losses, implemented as small classification heads (following Louizos et al. (2017)):

$$\mathcal{L}_{\text{aux},C} = -\mathbb{E}_{p_{\text{data}}(x,c)}\big[\log q_{\xi_C}(C \mid X)\big], \quad \mathcal{L}_{\text{aux},Y} = -\mathbb{E}_{p_{\text{data}}(x,c,y)}\big[\log q_{\xi_Y}(Y \mid X, C)\big].$$

The auxiliary losses serve two roles: (i) they are used during inference, providing predictors for C and Y, and (ii) they encourage representations that capture task-relevant information, sharpening the posterior and improving the quality of learned Z_C as a confounder proxy. Adding the CLUB-based mutual-information estimate $\mathcal{L}_{\text{MI}} = \widehat{\text{MI}}_{\psi}(Z_C, Z_S)$, The overall training objective is therefore

$$\mathcal{L}_{\text{train}} \;=\; -\mathcal{L}_{\text{ELBO}} + \lambda_C \mathcal{L}_{\text{aux},C} + \lambda_Y \mathcal{L}_{\text{aux},Y} + \lambda_{\text{MI}} \mathcal{L}_{\text{MI}}, \qquad (6)$$

where $\lambda_C, \lambda_Y, \lambda_{\text{MI}} \geq 0$ balance the auxiliary and independence terms. In our experiments, setting $(\lambda_C, \lambda_Y, \lambda_{\text{MI}}) = (1.0, 1.0, 0.1)$ yielded the best performance.

The auxiliary $q_{\xi_C}(C \mid X)$ is analogous to the concept layer in concept-bottleneck models (it provides an image-to-concept mapping), while $q_{\xi_Y}(Y \mid X, C)$ functions similar to an outcome layer with a residual connection (Yuksekgonul et al., 2023).

While the original CEVAE utilizes a TARNET-style architecture (Shalit et al., 2017) that fits separate outcome heads per treatment, our model shares decoders and conditions on C, since the networks would scale exponentially with the number of concepts. In our experiments we also explore a variant that allocates an independent discrete latent Z_{C_i} for each concept C_i (i.e., separate confounder proxies per concept). This allows us to estimate confounder proxies separately per concept and better interpret their relation with the observed proxies.

We use KL-annealing for the latent KL terms (gradually increasing their weight from 0 to 1 during early epochs) to avoid posterior collapse (Bowman et al., 2016) and temperature annealing for the Gumbel-Softmax relaxation of Z_C (start at τ_0 and reduce to $\tau_{\min}$) to transition from smooth relaxation to near-discrete samples (Jang et al., 2017).

ATE estimation After training, we estimate interventional means by marginalizing over the aggregated posterior of the confounder-latent Z_C. Concretely, for each test image x we draw samples (c, y, z) with $c \sim q_{\xi_C}(C \mid x)$, $y \sim q_{\xi_Y}(Y \mid x, c)$, and $z \sim q_{\phi_c}(Z_C \mid x, c, y)$. This yields approximate draws from $q_{\phi_c}(Z_C \mid x)$, analogous to the marginalization strategy in Louizos et al. (2017). For each target concept C_i, we intervene by setting it to $c \in \{0, 1\}$ while leaving the remaining concepts C_{-i} at their sampled values, and evaluate

$$\widehat{\mathbb{E}}[Y \mid do(C_i = c)] \approx \frac{1}{N}\sum_{n=1}^{N}\frac{1}{M}\sum_{m=1}^{M}\mathbb{E}_{\theta_y}\Big[Y \mid C_i = c,\ C_{-i} = c_{-i}^{(m)},\ Z_C = z^{(m)}\Big], \qquad (7)$$

where M (100 in our experiments) is the number of posterior samples per image. For each C_i, the ATE is the difference in predicted outcomes under interventions $C_i = 1$ and $C_i = 0$. To detect confounding, we compare the above ATE with the estimated difference in conditional means: $\text{ATE}_{\text{naive}} = \mathbb{E}[Y \mid C_i = 1] - \mathbb{E}[Y \mid C_i = 0]$, which would . We flag a concept as confounded when the computed ATE significantly and systematically differs from $\text{ATE}_{\text{naive}}$. For this, we employ a bootstrap test, in which we resample and recompute ATEs per batch and flag a concept as confounded if the 95% confidence intervals of the ATE do not overlap.

5 EXPERIMENTAL SETUP

5.1 DATASET

We evaluate UnCoVAEr on a controlled semi-synthetic benchmark derived from Morpho-MNIST (Castro et al., 2019), where digit images are systematically modified along interpretable morphological axes. All experiments use 5 random seeds; for each seed we select a different digit class (0–4).

This design minimizes variation due to digit identity and isolates causal effects arising purely from morphology.

We focus on four pixel-level morphological attributes as binary concepts: *thickness*, *intensity*, *slant*, and *width*. Continuous values for each concept are sampled conditionally from $\mathcal{N}(0.25, 0.01)$ when $C_i = 0$ and $\mathcal{N}(0.75, 0.01)$ when $C_i = 1$. The values are then scaled according to the attribute.

The outcome Y is a synthetic label constructed as a logical rule over the concepts:

$$Y = \mathbf{1}\{\text{thickness} + \text{slant} + \text{width} \geq 2\},$$

i.e., $Y = 1$ if at least two of these three concepts are active, with *intensity* not causing Y.

We design three dataset variants to probe distinct confounding patterns:

(i) **Single confounder:** observed concepts: $\{$*intensity, slant, width*$\}$; unobserved: $\{$*thickness*$\}$. Thickness causally influences intensity, making it the only observed concept affected by an unobserved confounder.

(ii) **Common confounder:** observed: $\{$ emphintensity, slant, width$\}$; unobserved: $\{$*thickness*$\}$. Here thickness jointly drives both intensity and slant, acting as a shared confounder across multiple observed concepts.

(iii) **Multiple confounders:** observed: $\{$*intensity, width*$\}$; unobserved: $\{$*thickness, slant*$\}$. Both thickness and slant affect intensity through a non-linear XOR causal mechanism, so a single observed concept is influenced by two distinct unobserved confounders.

Across all variants we control the *confounding strength* α. For a causal link $C_i \to C_j$, the label of C_j is set equal to C_i with probability α. We evaluate under two regimes: an *in-distribution* (ID) test set with strong confounding ($\alpha = 0.9$) and an *out-of-distribution* (OOD) test set with much weaker confounding ($\alpha = 0.6$), enabling us to assess robustness of ATE estimation under shifts in the confounding mechanism.

Lastly, to ensure that our method correctly adjusts for observed confounders, we construct an additional experimental setup by modifying the *Multiple confounders* variant. We assume that thickness is now observed (slant remains unobserved).

5.2 Baselines and ablations

We benchmark UnCoVAEr against latent-variable, concept-based, and feature-adjustment methods.

CEVAE (Louizos et al., 2017) is adapted for the image domain with convolutional encoders/decoders. Its difference from our method is that it does not partition the latent space but uses a single continuous latent.

CaCE (Goyal et al., 2020) estimates causal concept effects via counterfactual generation. Its original formulation refers to estimating effect on a classifier, rather than the true causal effect. For fair comparison we use the same architecture for encoder/decoder and we train an auxiliary predictor $q_\xi(Y \mid X)$, which we use to assess change in outcome.

Image-adjustment (Jerzak et al., 2023) conditions directly on image embeddings by fitting a propensity score model $\hat{e}_i(x) \approx p(C_i = 1 \mid X = x)$ and applying inverse-probability weighting (IPW) to estimate $E[Y \mid do(C_i)]$.

Concept Bottleneck Model (CBM) (Koh et al., 2020) predicts Y through an intermediate concept layer $\hat{C} = f(X)$ and enables interventions by editing $\hat{C}_i$.

Residual CBM (Res-CBM) augments standard CBM by explicitly modeling variation unexplained by observed concepts. Predictions are of the form $\hat{Y} = g(\hat{C}, r(X))$, where $r(X)$ is a residual representation. During training the concept layer remains fixed and we discretize $r(X)$ with Gumbel-Softmax. To estimate causal effects, we use IPW with $\hat{e}_i(C_i \mid r(X))$.

Finally, we include two meta-learners in the style of S-learners: (i) a **Naive Estimator**, which conditions only on C (biased under unobserved confounding); and (ii) an **Oracle Estimator**, which additionally conditions on the true latent confounder(s), providing an empirical upper bound.

To assess the contribution of each component of UnCoVAEr, we perform the following ablations: (i) removing the image reconstruction term $p_{\theta_x}(X \mid C, Z_C, Z_S)$, (ii) using only a shared discrete latent Z_C, (iii) the default model with shared Z_C and residual Z_S; and (iv) a variant with separate per-concept confounder proxies Z_{C_i}.

6 Results

Table 1 reports ATE estimation error across methods, datasets, and test regimes. Several consistent patterns emerge. First, in the *single confounder* setting, UnCoVAEr substantially outperforms all baselines, apart from the oracle which has access to the true confounder. The closest competitor is CEVAE, which itself can be seen as a restricted instance of our model using only a continuous latent.

Second, in the *common confounder* scenario, UnCoVAEr again improves upon feature-adjustment and CBM-based approaches. Interestingly, CaCE performs competitively here. Counterfactual concept editing remains effective when a single latent factor drives multiple observed pathways. Nevertheless, UnCoVAEr maintains strong performance, especially in-distribution.

Third, the *multiple confounders* variant exposes an interesting case. Since intensity is caused by the logical XOR of two latent factors, naive conditioning and CBMs manage to directly learn and exploit the *intenisty-Y* relation without accounting for the unobserved confounder at all, performing unexpectedly close to the oracle. Image-based methods, by contrast, are misled by this non-linear dependence. Among them, UnCoVAEr provides the lowest error, though the per-concept Z_{C_i} variant proves unstable in this regime. This suggests that while our structured latent partition is generally robust, learning disentangled proxies remains challenging under interacting confounders.

Finally, across all scenarios, UnCoVAEr shows improved out-of-distribution robustness: errors remain consistently lower than baselines when the strength of confounding shifts from $\alpha = 0.9$ to $\alpha = 0.6$. This supports our central claim that learning an explicit confounder proxy yields more stable causal effect estimates under distributional change.

Ablations and Diagnostics Table 1 further shows the effectiveness of the partitioned latent space design of UnCoVAEr. When the same bottleneck latent is used for reconstruction and for recovering the confounder proxy, the method underperforms. Moreover, our hypothesis that the proxy should guide image reconstruction is validated, as is evident in the performance drop. Finally, per-concept latents Z_{C_i} provide marginal gains, but become unstable under complex confounding. Figure 3 indicates that our confounding-detection criterion is generally effective, especially in the common confounder scenario, where it correctly characterizes both observed concepts. We report occasional false positives in the single case and a significant deterioration in the multiple-confounder case, where the naive estimators approximate the true ATE more closely than our estimands.

Table 1: Mean ATE estimation error (MAE, lower is better) across methods, datasets, and test regimes (averaged across concepts). Results are reported as mean ± std over 5 seeds. ID: in-distribution test set ($\alpha = 0.9$); OOD: out-of-distribution test set ($\alpha = 0.6$). Best non-oracle baseline per column is in **bold**.

	Single confounder		Common confounder		Multiple confounders	
Method	ID	OOD	ID	OOD	ID	OOD
Naive	.131 ± .18	.135 ± .19	.163 ± .11	.213 ± .04	**.009 ± .01**	**.009 ± .01**
Oracle	.003 ± .01	.002 ± .00	.002 ± .01	.009 ± .02	.001 ± .01	.001 ± .01
Image-adjustment	.168 ± .16	.133 ± .18	.440 ± .24	.183 ± .14	.109 ± .14	.117 ± .15
CBM	.136 ± .18	.136 ± .18	.163 ± .11	.214 ± .04	.011 ± .01	.012 ± .01
Res-CBM	.331 ± .20	.418 ± .09	.171 ± .17	.560 ± .40	.287 ± .24	.253 ± .23
CaCE	.114 ± .14	.087 ± .09	.058 ± .05	**.065 ± .03**	.157 ± .13	.166 ± .07
CEVAE	.058 ± .06	.049 ± .05	.079 ± .08	.112 ± .05	.106 ± .09	.096 ± .06
UnCoVAEr (no $p(X)$)	.113 ± .07	.112 ± .06	.064 ± .06	.119 ± .07	.210 ± .14	.209 ± .14
UnCoVAEr (only Z_C)	.070 ± .11	.089 ± .11	.098 ± .10	.172 ± .08	.080 ± .10	.077 ± .10
UnCoVAEr (Z_S + Z_C)	**.036 ± .04**	.040 ± .03	.055 ± .07	.097 ± .06	.070 ± .05	.065 ± .05
UnCoVAEr (Z_S + Z_{C_i}/C_i)	.041 ± .04	**.037 ± .04**	**.047 ± .03**	.105 ± .04	.136 ± .11	.138 ± .09

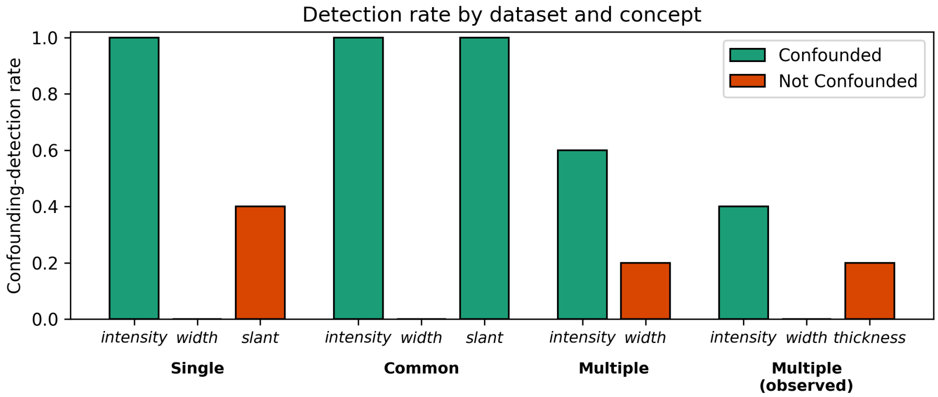

Figure 3: Empirical rate at which each concept was detected as confounded across random seeds (*confounding detection rate*). Results are shown for all three MorphoMNIST variants, plus the multiple-confounders setting where one of the confounders (thickness) is observed.

7 Discussion

Our work addresses a critical gap in concept-based model interpretation: the presence of unobserved visual confounders that bias causal effect estimates. While concept-based methods have gained traction for their interpretability, our results demonstrate that ignoring latent confounding can lead to substantially biased conclusions about which concepts truly drive model predictions.

Our experiments reveal interesting nuances in different confounding scenarios. While UnCoVAEr excels with single or shared confounders, performance degrades when confounders interact in non-linear ways (e.g., XOR). In such cases, direct statistical associations remain easier to capture than the underlying more complex causal structure, and all tested image-based methods fail. Handling complex, interacting confounders remains an open challenge requiring further methodological development. Still, the OOD evaluations are encouraging: robustness to shifts in confounding strength indicates that Z_C captures meaningful causal signals rather than overfitting correlations. This robustness is essential for real-world applications where confounding patterns may vary across datasets or deployment contexts.

Limitations and Future Work UnCoVAEr's primary limitation is its reliance on the assumption that confounders manifest visually in the image. Our experiments also highlight that complex causal structures or interactions remain challenging for current latent-variable approaches. However, the most critical challenge—and the most important direction for future work—is validating UnCoVAEr on complex real-world datasets. The true test of our model's practical utility lies in its ability to perform robustly in settings like medical imaging or model auditing, where concepts interact in unpredictable ways and the ground-truth confounding variables are fundamentally unknown. Successfully demonstrating effectiveness in these noisy, high-stakes environments is essential for moving from a theoretical proof of concept to a reliable tool for causal interpretability in applied domains.

8 Conclusions

We introduced UnCoVAEr, a deep latent-variable model for estimating causal concept effects under visual latent confounding. By partitioning the latent space into confounder-related and residual components, our method recovers proxy variables that enable valid backdoor adjustment even when key visual concepts remain unannotated. On controlled benchmarks, UnCoVAEr substantially reduces bias in causal effect estimates compared to existing concept-based and latent-variable approaches, while maintaining robustness under distribution shift. Our work highlights a critical consideration for practitioners that rely on concept-based explanations: incomplete concept annotations can severely bias causal conclusions. UnCoVAEr provides a practical tool for detecting and correcting such biases, enabling more trustworthy concept-level causal inference in partially annotated image datasets.

REFERENCES

Mohammad Taha Bahadori and David Heckerman. Debiasing concept-based explanations with causal analysis. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=6puUoArESGp.

Samuel Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. In *Proceedings of the 20th SIGNLL conference on computational natural language learning*, pp. 10–21, 2016.

Daniel C Castro, Jeremy Tan, Bernhard Kainz, Ender Konukoglu, and Ben Glocker. Morpho-mnist: Quantitative assessment and diagnostics for representation learning. *Journal of Machine Learning Research*, 20(178):1–29, 2019.

Daniel C Castro, Ian Walker, and Ben Glocker. Causality matters in medical imaging. *Nature Communications*, 11(1):3673, 2020.

Pengyu Cheng, Zhijie Hao, Wenyue Dai, Xinyuan Liu, and Weinan Chen. Club: A contrastive log-ratio upper bound of mutual information. In *International Conference on Machine Learning (ICML)*, 2020.

Pietro G Di Stefano, James M Hickey, and Vlasios Vasileiou. Counterfactual fairness: removing direct effects through regularization. *arXiv preprint arXiv:2002.10774*, 2020.

Gabriele Dominici, Pietro Barbiero, Francesco Giannini, Martin Gjoreski, Giuseppe Marra, and Marc Langheinrich. Counterfactual concept bottleneck models. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=w7pMjyjsKN.

Gabriele Dominici, Pietro Barbiero, Mateo Espinosa Zarlenga, Alberto Termine, Martin Gjoreski, Giuseppe Marra, and Marc Langheinrich. Causal concept graph models: Beyond causal opacity in deep learning. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=lmKJ1b6PaL.

Jifan Gao and Guanhua Chen. Mcce: Missingness-aware causal concept explainer, 2024. URL https://arxiv.org/abs/2411.09639.

Yash Goyal, Amir Feder, Uri Shalit, and Been Kim. Explaining classifiers with causal concept effect (cace), 2020. URL https://arxiv.org/abs/1907.07165.

Marton Havasi, Sonali Parbhoo, and Finale Doshi-Velez. Addressing leakage in concept bottleneck models. *Advances in Neural Information Processing Systems*, 35:23386–23397, 2022.

Daniel Israel, Aditya Grover, and Guy Van den Broeck. High dimensional causal inference with variational backdoor adjustment, 2023. URL https://arxiv.org/abs/2310.06100.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=rkE3y85ee.

Connor T. Jerzak, Fredrik Johansson, and Adel Daoud. Estimating causal effects under image confounding bias with an application to poverty in africa, 2023. URL https://arxiv.org/abs/2206.06410.

Jean Kaddour, Yuchen Zhu, Qi Liu, Matt Kusner, and Ricardo Silva. Causal effect inference for structured treatments. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=0v9EPJGc10.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2668–2677. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/kim18d.html.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International conference on machine learning*, pp. 5338–5348. PMLR, 2020.

Benjamin Kompa, David Bellamy, Tom Kolokotrones, james m robins, and Andrew Beam. Deep learning methods for proximal inference via maximum moment restriction. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 11189–11201. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/487c9d6ef55e73aa9dfd4b48fe3713a6-Paper-Conference.pdf.

Abhinav Kumar, Amit Deshpande, and Amit Sharma. Causal effect regularization: Automated detection and removal of spurious correlations. *Advances in Neural Information Processing Systems*, 36:20942–20984, 2023.

Sören R. Künzel, Jasjeet S. Sekhon, Peter J. Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165, 2019. doi: 10.1073/pnas.1804597116. URL https://www.pnas.org/doi/abs/10.1073/pnas.1804597116.

Bryson Lingenfelter, Sara R. Davis, and Emily M. Hand. A quantitative analysis of labeling issues in the celeba dataset. In George Bebis, Bo Li, Angela Yao, Yang Liu, Ye Duan, Manfred Lau, Rajiv Khadka, Ana Crisan, and Remco Chang (eds.), *Advances in Visual Computing*, pp. 129–141, Cham, 2022. Springer International Publishing. ISBN 978-3-031-20713-6.

Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. *Advances in Neural Information Processing Systems*, 30, 2017.

David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Fairness through causal awareness: Learning causal latent-variable models for biased data. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pp. 349–358, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450361255. doi: 10.1145/3287560.3287564. URL https://doi.org/10.1145/3287560.3287564.

Anita Mahinpei, Justin Clark, Isaac Lage, Finale Doshi-Velez, and Weiwei Pan. Promises and pitfalls of black-box concept learning models. *arXiv preprint arXiv:2106.13314*, 2021.

Andrei Margeloiu, Matthew Ashman, Umang Bhatt, Yanzhi Chen, Mateja Jamnik, and Adrian Weller. Do concept bottleneck models learn as intended? *arXiv preprint arXiv:2105.04289*, 2021.

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.

Ricardo Miguel de Oliveira Moreira, Jacopo Bono, Mário Cardoso, Pedro Saleiro, Mário A. T. Figueiredo, and Pedro Bizarro. Diconstruct: Causal concept-based explanations through black-box distillation. In Francesco Locatello and Vanessa Didelez (eds.), *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pp. 740–768. PMLR, 01–03 Apr 2024. URL https://proceedings.mlr.press/v236/moreira24a.html.

Tuomas Oikarinen, Subhro Das, Lam M Nguyen, and Tsui-Wei Weng. Label-free concept bottleneck models. In *The Eleventh International Conference on Learning Representations*, 2023.

Judea Pearl. [bayesian analysis in expert systems]: Comment: Graphical models, causality and intervention. *Statistical Science*, 8(3):266–269, 1993. ISSN 08834237. URL http://www.jstor.org/stable/2245965.

Sukrut Rao, Sweta Mahajan, Moritz Böhle, and Bernt Schiele. Discover-then-name: Task-agnostic concept bottlenecks via automated concept discovery. In *Computer Vision – ECCV*

2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LXXVII, pp. 444–461, Berlin, Heidelberg, 2024. Springer-Verlag. ISBN 978-3-031-72979-9. doi: 10.1007/978-3-031-72980-5_26. URL https://doi.org/10.1007/978-3-031-72980-5_26.

Severi Rissanen and Pekka Marttinen. A critical look at the consistency of causal estimation with deep latent variable models. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=vU96vWPrWL.

Yoshihide Sawada and Keigo Nakamura. Concept bottleneck model with additional unsupervised concepts. *IEEE Access*, 10:41758–41765, 2022.

Rickmer Schulte, David Rügamer, and Thomas Nagler. Adjustment for confounding using pretrained representations. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=D2cDJzotb8.

Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International conference on machine learning*, pp. 3076–3085. PMLR, 2017.

Chenming Shang, Shiji Zhou, Hengyuan Zhang, Xinzhe Ni, Yujiu Yang, and Yuwang Wang. Incremental residual concept bottleneck models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11030–11040, 2024.

Eric J Tchetgen Tchetgen, Andrew Ying, Yifan Cui, Xu Shi, and Wang Miao. An introduction to proximal causal learning. *arXiv preprint arXiv:2009.10982*, 2020.

Yixin Wang and David Blei. A proxy variable view of shared confounding. In *International Conference on Machine Learning*, pp. 10697–10707. PMLR, 2021.

Pengzhou Abel Wu and Kenji Fukumizu. β-intact-VAE: Identifying and estimating causal effects under limited overlap. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=q7n2RngwOM.

Liyuan Xu, Heishiro Kanagawa, and Arthur Gretton. Deep proxy causal learning and its application to confounded bandit policy evaluation. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=0FDxsIEv9G.

Chih-Kuan Yeh, Been Kim, Sercan Arik, Chun-Liang Li, Tomas Pfister, and Pradeep Ravikumar. On completeness-aware concept-based explanations in deep neural networks. *Advances in neural information processing systems*, 33:20554–20565, 2020.

Mert Yuksekgonul, Maggie Wang, and James Zou. Post-hoc concept bottleneck models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=nA5AZ8CEyow.

Weijia Zhang, Lin Liu, and Jiuyong Li. Treatment effect estimation with disentangled latent factors. In *AAAI Conference on Artificial Intelligence*, 2020. URL https://api.semanticscholar.org/CorpusID:210943075.

Xuan Zhao, Klaus Broelemann, Salvatore Ruggieri, and Gjergji Kasneci. Causal fairness-guided dataset reweighting using neural networks. In *2023 IEEE International Conference on Big Data (BigData)*, pp. 1386–1394. IEEE, 2023.

A Code and Implementation Details

All code, configuration files, and instructions required to reproduce our experiments are available at `https://anonymous.4open.science/r/causal-residual-concepts-346E/README.md`. The repository includes full implementations of UnCoVAEr and all baselines, as well as scripts for dataset preparation, training, and evaluation. We provide detailed configuration files specifying model architectures, optimizer settings, training schedules, and hyperparameter choices. Additional results, including json files and qualitative figures (e.g., counterfactual visualizations for benchmarked methods), are also included in the repository.

Reproducibility checklist

- **Datasets:** MorphoMNIST variants described in Section 5, with generation scripts included in the repository.
- **Evaluation metrics:** mean absolute error (MAE) of ATE estimates, bootstrap uncertainty test for confounding assessment, as described in Section 6.
- **Code availability:** full training/evaluation code and pre-trained model checkpoints are provided.
- **Hyperparameters:** all hyperparameters (learning rate, optimizer type, batch size, KL-annealing schedules, Gumbel-Softmax temperature annealing) are specified in configuration files.
- **Compute:** experiments were run on a single NVIDIA A10 GPU (24GB memory); training a model typically takes around 15 minutes.
- **Randomness:** results are averaged over 5 seeds, with random seeds fixed and logged for reproducibility.

B Use of Large Language Models

Large Language Models (LLMs) were used as assistive tools during the preparation of this paper. Their role was limited to improving readability and presentation: for example, rephrasing paragraphs for smoother academic flow, standardizing LaTeX formatting, and polishing grammar. In some cases, LLMs were also used to suggest more concise ways of summarizing experimental findings. They were not involved in research ideation, experimental design, implementation, or interpretation of results. All scientific contributions are the sole responsibility of the authors, who take full responsibility for the final content.