Proxy-Guided Measurement Calibration

Saketh Vishnubhatla SVISHNU6@ASU.EDU
Arizona State University, Tempe, AZ

Shu Wan SWAN16@ASU.EDU
Arizona State University, Tempe, AZ

Andre Harrison ANDRE.V.HARRISON2.CIV@ARMY.MIL
DEVCOM Army Research Lab, Adelphi, MD

Adrienne Raglin ADRIENNE.RAGLIN2.CIV@ARMY.MIL
DEVCOM Army Research Lab, Adelphi, MD

Huan Liu HUANLIU@ASU.EDU
Arizona State University, Tempe, AZ

Editors: Bijan Mazaheri and Niels Richard Hansen

Abstract

Aggregate outcome variables collected through surveys and administrative records are often subject to systematic measurement error. For instance, in disaster loss databases, county-level losses reported may differ from the true damages due to variations in on-the-ground data collection capacity, reporting practices, and event characteristics. Such miscalibration complicates downstream analysis and decision-making. We study the problem of outcome miscalibration and propose a framework guided by proxy variables for estimating and correcting the systematic errors. We model the data-generating process using a causal graph that separates latent content variables driving the true outcome from the latent bias variables that induce systematic errors. The key insight is that proxy variables that depend on the true outcome but are independent of the bias mechanism provide identifying information for quantifying the bias. Leveraging this structure, we introduce a two-stage approach that utilizes variational autoencoders to disentangle content and bias latents, enabling us to estimate the effect of bias on the outcome of interest. We analyze the assumptions underlying our approach and evaluate it on synthetic data, semi-synthetic datasets derived from randomized trials, and a real-world case study of disaster loss reporting. Our code [1] is made publicly available.

1. Introduction

Measurements obtained in empirical studies or administrative processes often deviate from the true outcome of interest due to systematic measurement errors (Imai and Yamamoto, 2010). These errors arise due to various factors, such as group-specific practices in data collection or a lack of necessary infrastructure (van der Steen et al., 2019). In such settings, quantifying the effect of outcome miscalibration is crucial for adjusting the observed measurements. Common approaches in the presence of measurement bias include running sensitivity tests to assess the validity of the hypothesis or calibration strategies that rely on validation data for which the true outcome is observed (VanderWeele and Li, 2019; Guerdan et al., 2023). Sensitivity tests can help assess the robustness of the findings, although they do not correct the measurement directly. Having access to true outcomes for

1. https://github.com/sak-18/proxy-guided-calibration

a validation subset is a valuable avenue for correcting measurement errors. However, in many real-world scenarios, this may be a strong assumption and often infeasible. In such settings, modeling measurement errors requires additional structure.

A useful signal comes from *proxy variables*: measurements that are correlated with the underlying outcome but not affected by the underlying systematic error mechanism itself. For example, disaster losses are collected on the ground through windshield surveys and on-site inspections, where losses are reported in terms of human lives impacted and property or crop losses in dollars. In this context, using sensor-based measurements, which are entirely independent of the loss data collection mechanism, can function as proxies. When such proxy information is available, it suggests a principled way to separate the latent "content" driving the true outcome from the latent "bias" driving the measurement errors.

Building on this causal perspective, we introduce a proxy-guided measurement calibration framework that utilizes variational autoencoders to disentangle unbiased content latents from bias-specific latents. The learned representation allows us to estimate the magnitude of the bias effect. We specify the assumptions required for identification and, through our empirical analysis, demonstrate that our method reliably recovers the underlying latent factors and accurately estimates the bias effect across a range of settings.

2. Related Work

2.1. Systematic Bias in Outcome Measurement

Several empirical studies demonstrate that measurement processes in crowdsourced urban data introduce systematic biases. Agostini et al. (2024) frame the study in the context of flood reporting across New York City census tracts. They demonstrate that resident-reported incidents exhibit systematic spatial variation in reporting intensity based on the socio-economic and demographic covariates. A Bayesian approach using an Ising model is used to correct for the underreporting, and to predict future event occurrence accurately. Complementing this perspective, Liu et al. (2024a) provide direct empirical evidence of spatial disparities in resident crowdsourcing across New York and Chicago. They propose a method to identify reporting rates solely from the logs of requests filed, without access to ground-truth event incident rates.

Reporting bias also arises in administrative records of public services, where observed usage reflects both demand and access. Liu et al. (2024b) study public library systems and demonstrate that circulation and usage statistics can obscure underlying disparities in service access. Similar concerns have long been raised in the context of disaster loss databases. Gall et al. (2009) document six recurring fallacies in natural hazards loss databases, demonstrating that reported losses depend on thresholds for inclusion, spatial and temporal aggregation choices, and institutional accounting practices, leading to systematic bias across regions and events. Together, these studies underscore that observed outcomes in many applications are shaped by the reporting mechanisms, motivating approaches that explicitly account for systematic reporting biases when drawing conclusions from real-world data.

A related line of work addresses systematic errors from a machine learning perspective. In machine learning, noisy or biased labels are often modeled via a label corruption process, such as a noise transition matrix, and learning is adjusted accordingly; recent results show that such noise models can be estimated even without idealized anchor points by leveraging high-confidence examples (Xia et al., 2019). In many applied settings, however, label noise reflects a deeper form of

outcome mismeasurement, where observed labels serve as proxies for underlying quantities whose relationship to the true outcome varies systematically across contexts or subpopulations (Song et al., 2023). This phenomenon has been formalized as causal label bias, highlighting how reliance on proxy outcomes can mask disparities and lead to misleading conclusions about performance or fairness when evaluated solely on observed labels (Mhasawade et al., 2024). From a causal perspective, outcome mismeasurement can further invalidate standard counterfactual identification; recent work studies counterfactual prediction under outcome measurement error (Guerdan et al., 2023), derives partial identification bounds for causal effects under differential misreporting (Huang and Makar, 2022), and develops sensitivity analysis frameworks to assess how severe outcome mismeasurement would need to be to explain away observed effects (VanderWeele and Li, 2019).

Together, these strands emphasize that outcome miscalibration is often systematic rather than random, motivating approaches that explicitly reason about and correct measurement bias rather than treating noisy labels as incidental corruption.

2.2. Latent Variable Models and Identifiability

Latent variable modeling has become a central tool for causal inference when key variables are unobserved. A prominent example is addressing unobserved confounding in observational studies: Louizos et al. (2017) introduce CEVAE, one of the first works to integrate deep latent variable models with causal effect estimation by using proxy variables to infer latent confounders. Related work applies latent variable models to settings where the outcome of interest is observed only through multiple noisy measurements, showing that causal effects on the underlying outcome can be recovered via an optimally weighted combination of proxies (Fu and Green, 2025).

A central challenge for such models is identifiability. In deep generative models, this issue is closely linked to posterior collapse, where latent representations become uninformative. Wang et al. (2021) show that posterior collapse arises precisely when latent variables are non-identifiable, and propose identifiable VAE constructions that enforce recoverability through structural constraints. Complementary theoretical results establish identifiability for broad classes of VAE-based models under suitable assumptions, even without auxiliary supervision (Kivva et al., 2022), connecting these developments to classical results in statistics and causal discovery where additional assumptions, such as non-Gaussianity, enable identification of latent factors and causal structure (Hyvärinen et al., 2024). Together, this work informs how latent variable models can be designed to recover meaningful latent structure for causal analysis.

3. Preliminaries

3.1. Notation

Throughout the paper, lowercase letters (e.g., k) denote scalars, and lowercase boldface letters (e.g., $\mathbf{v}$) denote vectors in $\mathbb{R}^d$. Uppercase boldface letters (e.g., $\mathbf{W}$) denote matrices, and calligraphic letters (e.g., $\mathcal{S}$) denote sets. Plain uppercase letters such as E denote random variables or nodes in a causal graph. Lowercase variables denote realizations of the corresponding random variables. We write $P(\cdot)$ for probability mass functions, $p(\cdot)$ for densities, $\mathbb{E}[\cdot]$ for expectations, and $do(\cdot)$ for Pearl's intervention operator.

In our model, E denotes observed environment covariates. The latent content vector is written $\mathbf{z} \in \mathbb{R}^{d_z}$, and the latent bias vector is written $\mathbf{a} \in \mathbb{R}^{d_a}$. The unobserved true outcome is Y_{true}, and

the biased observed measurement is Y_{obs}. We observe a vector of m proxy measurements,

$$\mathbf{y}_{\text{proxy}} = \left(y^{(1)}_{\text{proxy}}, \ldots, y^{(m)}_{\text{proxy}}\right) \in \mathbb{R}^m,$$

which depend on the latent content $\mathbf{z}$ but not influenced by the bias mechanism $\mathbf{a}$. When convenient, we write Z and A for the corresponding random variables and $\mathbf{z}$ and $\mathbf{a}$ for their unobserved realized values and $\hat{z}, \hat{a}$ for point estimates.

3.2. Proxy-Guided Measurement Calibration

We define *proxy-guided measurement calibration* as the task of recovering the unobserved true outcome Y_{true} from a biased observed measurement Y_{obs} by leveraging auxiliary proxy variables that are not causally influenced by the bias mechanism. In many real-world settings, systematic reporting errors cause Y_{obs} to deviate from Y_{true}. Such errors arise from latent bias factors $\mathbf{a}$ that may vary across environments indexed by E, while the underlying physical or domain-specific signal is governed by latent content factors $\mathbf{z}$.

Proxy-guided calibration formalizes this setup through a causal model in which the proxies serve as "clean" measurements of $\mathbf{z}$, enabling the disentanglement of the content factors $\mathbf{z}$ from the bias factors $\mathbf{a}$. This separation allows us to conceptually define the de-biased counterfactual outcome $Y_{\text{obs}}(0)$, which corresponds to the true outcome Y_{true} in the model. Recovering $Y_{\text{obs}}(0)$ from observational data would require assumptions we detail in the following sections.

4. Proxy-Guidance for Measurement Calibration

Our framework utilizes proxy measurements to separate true latent factors from the bias mechanism. We present the conditions under which we can identify the unobserved true outcome, describe how they are recovered via a two-stage VAE co-training setup, and demonstrate how the learned representations help quantify the effect of the bias.

4.1. Generative Model

We assume each observation is generated according to the causal graph in Figure 1. We observe an environment variable $E = (E_1, \ldots, E_{d_e})$, which determines the latent factors. The latent content and bias variables are denoted by $Z = (Z_1, \ldots, Z_{d_z})$ and A. The unobserved true outcome is Y_{true}, the biased observed outcome is Y_{obs}, and the proxy measurements are collected in the vector-valued random variable Y_{proxy}. The generative process is defined as below:

$$E \sim p(E), \qquad Z \sim p(Z \mid E), \qquad A \sim p(A \mid E),$$

$$Y_{\text{true}} \sim p(Y_{\text{true}} \mid Z), \qquad Y_{\text{proxy}} \sim p(Y_{\text{proxy}} \mid Z),$$

$$Y_{\text{obs}} \sim p(Y_{\text{obs}} \mid Z, A).$$

The joint distribution factorizes as:

$$p(E, Z, A, Y_{\text{true}}, Y_{\text{proxy}}, Y_{\text{obs}}) = p(E)\, p(Z \mid E)\, p(A \mid E)\, p(Y_{\text{true}} \mid Z)\, p(Y_{\text{proxy}} \mid Z)\, p(Y_{\text{obs}} \mid Z, A),$$

which encodes that the proxy measurements depend only on the latent content variable Z, while the observed outcome depends on both content and bias.

Remarks on the generative process. We observe independent units for $i = 1, \ldots, n$,

$$\big(E_i,\ Z_i,\ A_i,\ Y^{(1)}_{\text{proxy},i}, \ldots, Y^{(m)}_{\text{proxy},i},\ Y_{\text{obs},i} \big),$$

where E_i indexes realized covariates (e.g., county and hazard-specific attributes) and Z_i, A_i are latent. This is *not* a multi-environment framework with repeated samples per environment. Replication arises across units with different covariate values E_i, and the m proxy measurements are distinct measurement channels for the same unit, not repeated draws.

4.2. Identifiability

We study the problem of identifying the effect of reporting bias on the observed outcome. For the purposes of identification and estimation in this work, we focus on the scalar binary case $d_a = 1$, where $A \in \{0, 1\}$ indicates the absence or presence of reporting bias. Our goal is to recover the expected observed outcome under removal of reporting bias. For any observation with environment value e and latent content value z, our causal estimand of interest is:

$$\mu(e, z) := \mathbb{E}[\, Y_{\text{obs}} \mid do(A = 0),\ E = e,\ Z = z \,],$$

which represents the bias-free conditional mean of the observed outcome. Because our representation learning procedure learns the latent variables (Z, A) up to trivial transformations, we treat (E, Z, A) as observed for the identification argument. The DAG in Figure 1 implies two structural conditions: (i) all parents of A are contained in (E, Z), and (ii) E has no parents. These conditions ensure that (E, Z) is a valid adjustment set for the causal effect of A on Y_{obs}.

Proposition 1 (Identification) *For any (e, z) in the support of (E, Z),*

$$\mu(e, z) := \mathbb{E}[\, Y_{\text{obs}} \mid do(A = 0),\ E = e,\ Z = z \,] = \mathbb{E}[\, Y_{\text{obs}} \mid A = 0,\ E = e,\ Z = z \,].$$

Proof [Sketch] By definition,

$$\mu(e, z) = \int y\ p(y \mid do(A = 0),\ E = e,\ Z = z)\ dy.$$

In the DAG of Figure 1, all parents of A are contained in (E, Z) and E has no parents. Thus (E, Z) is an admissible adjustment set for the edge $A \to Y_{\text{obs}}$: conditioning on (E, Z) blocks all backdoor paths from A to Y_{obs} and contains no descendants of A. By the backdoor criterion (a special case of Rule 2 of do-calculus), this implies

$$p(y \mid do(A = 0),\ E = e,\ Z = z) = p(y \mid A = 0,\ E = e,\ Z = z).$$

Substituting into the integral and using the definition of conditional expectation yields

$$\mu(e, z) = \mathbb{E}[\, Y_{\text{obs}} \mid A = 0,\ E = e,\ Z = z \,],$$

as claimed. ■

Identifiability of the latent variables. Proposition 1 reduces the interventional estimand to an observational estimand, but requires access to the latent factors (Z, A). We briefly justify that these are recoverable from the observed data with formal assumptions deferred to Appendix B. For the content latent Z identification argument, we refer to nonlinear ICA with the environment E as an observed auxiliary variable (Hyvarinen et al., 2019). Under conditional independence of $(Y_{\text{proxy}}, Y_{\text{obs}})$ given Z, invertibility of the mixing function generating the proxies, and sufficient variability of $p(Z \mid E)$ across covariate values, the content latent is identifiable from (Y_{proxy}, E) up to component-wise invertible transforms and permutation. For the bias indicator, we restrict $A \in \{0, 1\}$ to be binary, and assume access to *anchor* units for which the bias status is known *a priori*. Together, these two conditions suffice to identify the bias regime from the learned latent score, as detailed in Section 4.4 and Appendix B.

4.3. Latent Recovery with VAEs

We use a two-stage variational autoencoder to learn the content latents Z and the bias latent A. We write $q_\phi(\cdot)$ for encoder distributions (variational posteriors) and $p(\cdot)$ for generative factors (priors and decoders). Throughout, Z is a continuous latent and $A \in \{0, 1\}$ is a binary bias indicator.

Stage 1: Learning Content Latents from Proxies. In the first stage, we learn the content latent Z using only the proxy measurements Y_{proxy} and the environment E. We specify an encoder $q_{\phi_Z}(z \mid y_{\text{proxy}}, E)$, an environment-conditioned prior $p(z \mid E)$, and a decoder $p(y_{\text{proxy}} \mid z)$. The resulting evidence lower bound (ELBO) is

$$\mathcal{L}_Z = \mathbb{E}_{q_{\phi_Z}}\big[\log p(y_{\text{proxy}} \mid z)\big] - \mathrm{KL}(q_{\phi_Z}(z \mid y_{\text{proxy}}, E) \,\|\, p(z \mid E)) \,.$$

Because neither the observed outcome Y_{obs} nor the bias latent A appears in this stage, and because the proxies are assumed not to depend on the bias mechanism, the learned representation Z captures variation associated with the underlying content rather than reporting bias. Let $\hat{z}$ denote a point estimate of Z obtained from encoder for an observation. In the second stage, $\hat{z}$ is treated as fixed.

Stage 2: Learning the Bias Latent. In the second stage, we infer the bias latent A from the observed outcome Y_{obs}, conditional on the frozen content estimate $\hat{z}$ and the environment E. We define an encoder $q_{\phi_A}(a \mid Y_{\text{obs}}, E, \hat{z})$, an environment-conditioned prior $p(a \mid E)$, and a decoder $p(Y_{\text{obs}} \mid \hat{z}, a)$. The corresponding ELBO is

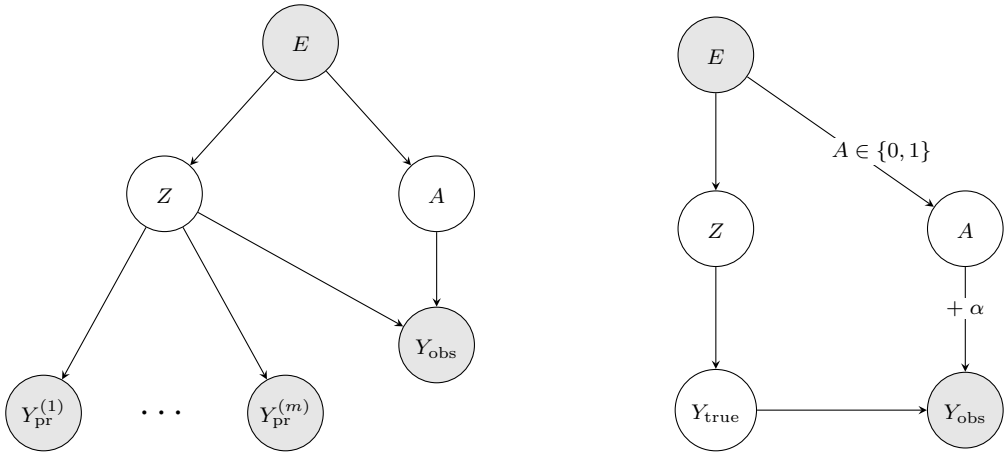
$$\mathcal{L}_A = \mathbb{E}_{q_{\phi_A}}\big[\log p(Y_{\text{obs}} \mid \hat{z}, a)\big] - \mathrm{KL}(q_{\phi_A}(a \mid Y_{\text{obs}}, E, \hat{z}) \,\|\, p(a \mid E)) \,.$$

This stage attributes systematic errors in the observed outcome, relative to the proxy-informed content representation, to the bias latent A, in accordance with the assumed generative structure.

Combined encoder. Together, the two stages define an encoder that produces pointwise estimates of the content and bias latents,

$$(\hat{z}, \hat{a}) = \Big(\hat{z},\ \hat{a}(Y_{\text{obs}}, E, \hat{z})\Big),$$

where $\hat{z}$ is obtained from the Stage 1 encoder and $\hat{a}$ from the Stage 2 encoder. These latent estimates are subsequently used in a post-hoc calibration step to quantify the effect of reporting bias.



(a) Full generative model (b) Error model

Figure 1: **Causal structure for proxy-guided measurement calibration.** (a) Full generative model: environment variables E influence latent content factors Z and latent reporting bias A. Proxy measurements $\{Y_1, \ldots, Y_m\}$ depend only on Z, while the observed outcome Y_{obs} depends on both Z and A. (b) Error model highlighting the measurement bias mechanism: the true outcome Y_{true} is perturbed by an environment-dependent binary bias A with additive magnitude α.

4.4. Bias Model and Estimation

After recovering latent representations for content and bias, we focus on estimating the magnitude of the reporting bias. We adopt a simple additive bias model, illustrated in Figure 1, in which the observed outcome differs systematically from the true outcome when reporting bias is present. The true outcome is generated as a function of the latent content and environment,

$$Y_{\text{true}} = g(Z, E) + \varepsilon, \qquad \mathbb{E}[\varepsilon \mid Z, E] = 0,$$

and the observed outcome incorporates a bias-induced shift,

$$Y_{\text{obs}} = Y_{\text{true}} + \alpha\, A, \qquad A \in \{0, 1\}.$$

When $A = 0$, the observed measurement is unbiased i.e. $Y_{\text{obs}} = Y_{\text{true}}$, while when $A = 1$ the outcome is shifted by α on average. The scalar parameter α therefore quantifies the magnitude of reporting bias.

The two-stage VAE yields pointwise latent representations $(\hat{Z}, \hat{A})$ for each unit, where $\hat{A}$ is a real-valued latent score associated with the binary bias indicator $A \in \{0, 1\}$. Because latent representations are identifiable only up to simple reparameterizations, the scale and orientation of $\hat{A}$ are arbitrary. Consequently, numerical values of $\hat{A}$ (e.g., 0.9 versus 0.1) need not correspond in any fixed way to the presence or absence of bias, and the mapping between $\hat{A}$ and A may be reversed or rescaled without affecting the implied data distribution.

To estimate the bias parameter α, we construct a matched comparison between units inferred to be in the biased regime and units inferred to be in the unbiased regime. Let $\mathcal{I}_1$ denote the set of units with high values of the latent bias score $\hat{A}$ and $\mathcal{I}_0$ the set of units with low values of $\hat{A}$. For

each unit $i \in \mathcal{I}_1$, we select a set $\mathcal{N}_K(i) \subset \mathcal{I}_0$ consisting of its K nearest neighbors in the recovered content space $\hat{Z}$. The bias magnitude is estimated as

$$\hat{\alpha} = \frac{1}{|\mathcal{I}_1|} \sum_{i \in \mathcal{I}_1} \left(Y_{\text{obs},i} - \frac{1}{K} \sum_{j \in \mathcal{N}_K(i)} Y_{\text{obs},j} \right).$$

This estimator contrasts observed outcomes for biased units with the average outcomes of matched unbiased units that share similar latent content, thereby removing variation attributable to the outcome function $g(Z, E)$. Under the additive bias model, and assuming that matching on $\hat{Z}$ (and on E when included) renders the conditional mean of $g(Z, E)$ identical across matched units, the estimator $\hat{\alpha}$ consistently recovers the average reporting bias parameter α.

Under standard conditions for matching estimators, including sufficient overlap in the support of $\hat{Z}$, bounded covariate support, and smoothness of the conditional mean $g(Z, E)$, classical large-sample theory ensures that the K-nearest-neighbor matching estimator $\hat{\alpha}$ is consistent for the bias parameter α as $n \to \infty$ with fixed K (Abadie and Imbens, 2006, 2011). We also note that the primary estimand of interest is the magnitude of systematic outcome miscalibration rather than pointwise recovery of Y_{true}. Once $\hat{Z}$ and $\hat{A}$ are identified, recovery of Y_{true} follows directly from the additive relation $Y_{\text{obs}} = Y_{\text{true}} + \alpha A$.

5. Experiments

Evaluating outcome miscalibration in real-world settings is challenging. The true data-generating mechanisms are often unknown and calibration errors are rarely documented. We therefore begin with a controlled synthetic setting, where the full latent structure is observable, and then validate our approach on semi-synthetic datasets built from randomized trials. Finally, we provide a real-world case study motivated by biases in recording natural disaster losses.

5.1. Experiments on Synthetic Data

We first assess our framework under different synthetic data-generating processes (DGPs) that vary in latent dimensionality, sample size, and noise type. This allows us to examine conditions such as functional form, proxy informativeness, and bias strength that cannot be controlled in real-world observational data. Synthetic data is sampled from the following DGP:

Table 1: Assumptions across experimental settings. ✓ indicates the assumption is satisfied, × indicates it is not assumed, and ∘ indicates partial or indirect availability.

	Synthetic	Semi-synthetic	Real-world
(A1) Proxy exclusion	✓	✓	✓
(A2) Bias model: constant α	✓	✓	×
(A3) Sign restriction: $\alpha \geq 0$	✓	✓	×
(A4) Linear structural mechanisms	✓	×	×
(A5) Availability of ground-truth (A, Z)	✓	∘	×
(A6) Overlap in A	✓	✓	∘

$$
\begin{aligned}
\mathbf{e} &\sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{d_e}), \qquad \mathbf{e} \in \mathbb{R}^{d_e}, \\
\mathbf{z} &= \mathbf{W}_z^\top \mathbf{e} + \boldsymbol{\varepsilon}_z, \qquad \mathbf{W}_z \in \mathbb{R}^{d_e \times d_z},\ \boldsymbol{\varepsilon}_z \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{d_z}), \\
A &= \mathbf{1}\Big\{\sigma\Big(\mathbf{w}_a^\top \mathbf{e} + \varepsilon_a\Big) > \tfrac{1}{2}\Big\}, \qquad \mathbf{w}_a \in \mathbb{R}^{d_e}, \\
Y_{\text{true}} &= \mathbf{w}_y^\top \mathbf{z} + \varepsilon_y, \qquad \mathbf{w}_y \in \mathbb{R}^{d_z}, \\
Y_{\text{obs}} &= Y_{\text{true}} + \alpha A + \varepsilon_{\text{obs}}, \qquad \alpha \in \mathbb{R}, \\
Y_{\text{proxy}}^{(k)} &= c_k\, Y_{\text{true}} + \varepsilon_{\text{proxy}}^{(k)}, \qquad k = 1, \ldots, m.
\end{aligned}
$$

In this synthetic setup, environment covariates $\mathbf{e} = (E_1, \ldots, E_{d_e})$ jointly induce latent content variables $\mathbf{z} = (Z_1, \ldots, Z_{d_z})$ and a scalar bias indicator A through linear functions of $\mathbf{e}$. The true outcome Y_{true} depends only on the content latents, while the observed measurement Y_{obs} incorporates an additive bias of magnitude α whenever the bias mechanism is active. Proxy measurements $Y_{\text{proxy}}^{(1)}, \ldots, Y_{\text{proxy}}^{(m)}$ are generated as independent noisy linear scalings of Y_{true} and are assumed to be unaffected by the bias variable A, enforcing a strict exclusion restriction. This DGP makes several strong structural assumptions. In particular, both the content and bias mechanisms are linear in the environment, the bias enters the outcome additively, and all proxy variables are conditionally independent given Y_{true} and depend on it only through fixed linear coefficients.

Setup. Unless stated otherwise, we use $d_e = 10$ environment variables. The dataset is split into train/validation/test in an 80/10/10 ratio. Both VAE stages use Adam with learning rate 10^{-3}. Following latent recovery, we estimate treatment effects via a matching estimator between treated units (high A) and controls (low A) stratified by Z, where the threshold is chosen on the validation set. Evaluation focuses on estimating α and recovering (Z, A) upto permutation and scale.

Results. Table 2 reports $\hat{\alpha}$ across sample sizes, latent dimensions, and noise models. Across all settings, our method recovers α accurately, with performance improving as sample size grows. Bias magnitudes of $\alpha = 5$ show the clearest trend with sample size, while extreme values ($\alpha = 1$ and $\alpha = 10$) exhibit broader variance. Gaussian and Poisson noise exhibit nearly identical behavior, indicating that the type of noise has no effect.

5.2. Experiments on Semi-Synthetic Data

To evaluate performance in more realistic settings, we next construct semi-synthetic datasets from two randomized controlled trials. In these datasets, the environment variables and proxy measurements, and consequently the latents, arise naturally from the real study context, while reporting bias is introduced artificially. This setting allows us to study real-world latent mechanisms while having access to the induced bias level α.

Data. We evaluate our method using two standard causal inference benchmarks. The Oregon Health Insurance Experiment (OHIE) is a randomized Medicaid lottery in which treatment corresponds to insurance enrollment. The outcome is log-transformed total hospital spending, and we use five biomedical and mental health indicators as proxy measurements that reflect underlying health

Table 2: Bias estimation with synthetic data under different settings.

			Estimated α					
			$\alpha = 1$		$\alpha = 5$		$\alpha = 10$	
n	d_z	d_e	Gaussian	Poisson (scaled)	Gaussian	Poisson (scaled)	Gaussian	Poisson (scaled)
500	1	10	1.20±0.23	1.26±0.23	4.16±0.65	4.17±0.64	8.59±1.48	8.61±1.47
500	2	10	1.07±0.28	1.06±0.36	3.88±0.51	3.63±0.48	8.23±1.47	7.82±1.65
500	5	10	1.08±0.22	0.88±0.33	3.95±0.48	3.71±0.65	8.61±0.93	8.19±0.93
500	10	10	1.01±0.28	1.06±0.15	3.81±1.37	4.03±1.07	7.58±2.31	8.16±1.56
500	20	10	1.28±0.32	1.21±0.34	3.90±0.30	3.72±0.41	8.26±2.38	8.33±2.34
1000	1	10	1.18±0.21	1.16±0.22	4.34±0.34	4.37±0.50	9.02±0.97	8.93±1.21
1000	2	10	0.89±0.12	0.86±0.13	3.94±0.47	4.03±0.50	9.23±0.97	8.89±1.07
1000	5	10	0.76±0.17	0.76±0.16	4.03±0.41	3.95±0.37	8.86±1.22	8.69±1.01
1000	10	10	1.04±0.16	1.05±0.15	4.41±0.47	4.31±0.47	8.04±1.78	7.91±1.48
1000	20	10	1.29±0.17	1.28±0.18	3.64±0.25	3.66±0.30	8.34±1.20	8.36±1.20
2500	1	10	0.97±0.11	0.98±0.13	4.01±0.43	4.02±0.46	7.93±1.04	7.97±1.07
2500	2	10	0.80±0.18	0.83±0.22	4.31±0.44	4.18±0.56	8.78±0.78	8.74±0.80
2500	5	10	0.74±0.15	0.77±0.12	4.43±0.35	4.43±0.35	8.83±0.70	8.81±0.59
2500	10	10	0.96±0.07	0.97±0.10	4.44±0.44	4.33±0.52	8.90±1.07	8.77±1.26
2500	20	10	1.13±0.25	1.09±0.25	4.01±0.26	3.98±0.24	8.87±1.10	9.01±0.98
5000	1	10	1.03±0.16	1.02±0.14	4.34±0.50	4.24±0.55	8.46±1.31	8.33±1.43
5000	2	10	0.84±0.16	0.84±0.13	4.28±0.54	4.32±0.51	8.81±1.28	8.81±1.18
5000	5	10	0.74±0.09	0.75±0.07	4.50±0.69	4.50±0.68	9.08±1.35	8.99±1.38
5000	10	10	0.95±0.06	0.93±0.06	4.54±0.39	4.41±0.42	9.04±0.86	8.83±0.79
5000	20	10	0.98±0.15	1.02±0.18	4.19±0.45	4.20±0.42	8.14±1.26	8.20±1.16
10000	1	10	1.11±0.08	1.10±0.06	4.39±0.57	4.43±0.56	8.44±1.04	8.70±1.11
10000	2	10	0.98±0.06	0.97±0.05	4.10±0.60	4.14±0.52	8.69±1.17	8.72±1.09
10000	5	10	0.97±0.17	0.96±0.09	4.29±0.64	4.46±0.58	8.78±1.08	9.20±0.88
10000	10	10	1.05±0.08	1.05±0.06	3.91±0.67	4.14±0.62	7.89±1.54	8.24±1.28
10000	20	10	1.18±0.12	1.20±0.12	4.81±0.09	4.65±0.35	8.79±1.33	8.90±1.11

status but are not subject to reporting bias. Environment variables capture pre-lottery utilization and hospitalization history. The JOBS dataset studies the impact of a job training program, with treatment defined by program participation. The outcome is 1978 earnings, while pre-treatment earnings from 1974 and 1975 serve as unbiased proxies. Environment covariates include standard demographic attributes such as age, education, race, and marital status. Both datasets provide clean treatment assignment with auxiliary information that can be used as proxies, making them well-suited for evaluating proxy-guided measurement calibration.

Experimental Setup. We compare our method against several baselines that do not explicitly leverage proxies, all evaluated under k-fold cross-validation over 10 folds. In all settings, folds are constructed at the unit level, models are trained exclusively on training folds, and all latent representations and estimators are frozen before evaluation on held-out test folds.

The *proxy-only* baseline uses the proxy variables to learn a predictor $\widehat{Y}$ of the observed outcome, treating proxies as unbiased measurements of the latent signal. Bias is then estimated by contrasting residuals $Y_{\text{obs}} - \widehat{Y}$ across inferred treatment groups, attributing systematic residual differences to reporting distortion. The *environment-only* baseline models Y_{obs} directly as a function of environmental covariates, attributing systematic variation across treatment groups to environment-driven reporting effects. Differences predicted by this model are used to estimate the bias magnitude $\hat{\alpha}$, implicitly assuming that environmental variables capture the dominant sources of reporting heterogeneity.

We also include TEDVAE Zhang et al. (2021), a variational autoencoder designed to disentangle latent factors relevant for heterogeneous treatment effect estimation using proxies for unobserved

Table 3: Estimated $\hat{\alpha}$ for JOBS and OHIE across baselines and latent dimensions.

	JOBS			OHIE		
	$\alpha = 1$	$\alpha = 5$	$\alpha = 10$	$\alpha = 1$	$\alpha = 5$	$\alpha = 10$
Baseline (Proxy only)	4.718 ± 0.03	8.382 ± 0.04	13.382 ± 0.04	22.418 ± 0.05	24.073 ± 0.06	27.973 ± 0.07
Baseline (Env only)	3.526 ± 0.13	7.010 ± 0.26	11.979 ± 0.26	15.204 ± 0.55	15.200 ± 0.47	16.114 ± 0.29
Baseline (TEDVAE)	0.505 ± 0.32	1.304 ± 0.71	2.275 ± 1.37	0.225 ± 0.18	0.279 ± 0.18	0.769 ± 0.43
Ours dim. (Z=5)	1.694 ± 0.42	3.807 ± 1.14	7.580 ± 1.30	1.807 ± 0.37	4.415 ± 0.30	7.700 ± 0.36
Ours dim. (Z=10)	1.505 ± 0.35	3.529 ± 0.89	6.518 ± 1.63	1.941 ± 0.36	4.320 ± 0.18	7.548 ± 0.28
Ours dim. (Z=15)	1.498 ± 0.44	3.583 ± 0.71	6.773 ± 1.69	1.690 ± 0.37	4.311 ± 0.31	7.594 ± 0.50
Ours dim. (Z=20)	1.556 ± 0.46	3.296 ± 0.47	7.647 ± 1.85	1.863 ± 0.36	4.294 ± 0.14	7.832 ± 0.39

confounders. In contrast, our method uses a two-stage VAE to disentangle content and bias latents by combining proxy exclusion with environment-conditioned bias modeling prior to estimating α.

Results. Table 3 shows differences across datasets. On OHIE, our proxy-guided model accurately recovers the true bias magnitude α across all regimes and substantially outperforms all baselines, with particularly strong performance for $\alpha = 5$ and $\alpha = 10$. On JOBS, our method consistently improves over proxy-only, environment-only, and TEDVAE baselines, but underestimates α in moderate and high bias settings, reflecting the increased difficulty of calibration in this dataset. Proxy-only and environment-only baselines substantially overestimate bias in both datasets, while TEDVAE often yields near-zero estimates across regimes. TEDVAE optimizes a latent noise component to support treatment effect estimation rather than to identify or preserve the magnitude of systematic measurement bias, which leads to attenuated bias estimates in calibration-focused evaluations. Across both datasets, our performance is moderately stable with respect to the latent dimension d_z, indicating robustness to the choice of representation size.

5.3. Case Study on Real-World Data: SHELDUS Disaster Loss Data

For a real-world case study, we use SHELDUS, a disaster loss database (ASU Center for Emergency Management and Homeland Security, 2025). SHELDUS reports the county-level damages in terms of property and crop damages, along with injury and fatality counts for natural disasters.

Data. Our units of interest include counties in SHELDUS that have been affected by wildfires, hurricanes/tropical storms, flooding, and tornadoes. As our outcome variable of interest, we chose the loss for property damage in the county. The observed outcome Y_{obs} is log property damage for each county–hazard–year. Remote sensing indicators, which provide information on land cover change from one class to another (e.g., area change from built-up to water indicative during floods), form a set of proxy variables. We limit the time period of interest from 2016 to 2023, as the remote sensing product we rely on, Dynamic World, is available only from 2015 (Brown et al., 2022). We provide more information on the dataset in the appendix A.2. Environment variables E includes demographic and socioeconomic attributes (e.g., population, income, poverty rate, median age) and the category of the disaster.

Estimating effects of reporting bias. Our framework enables counterfactual comparisons between counties that are physically similar but differ in their degree of reporting bias. Counties are partitioned into "treated" (high bias) and "control" (low bias) groups using a threshold on the estimated latent $\hat{A}$. For each treated county i, we estimate a counterfactual outcome by matching to its

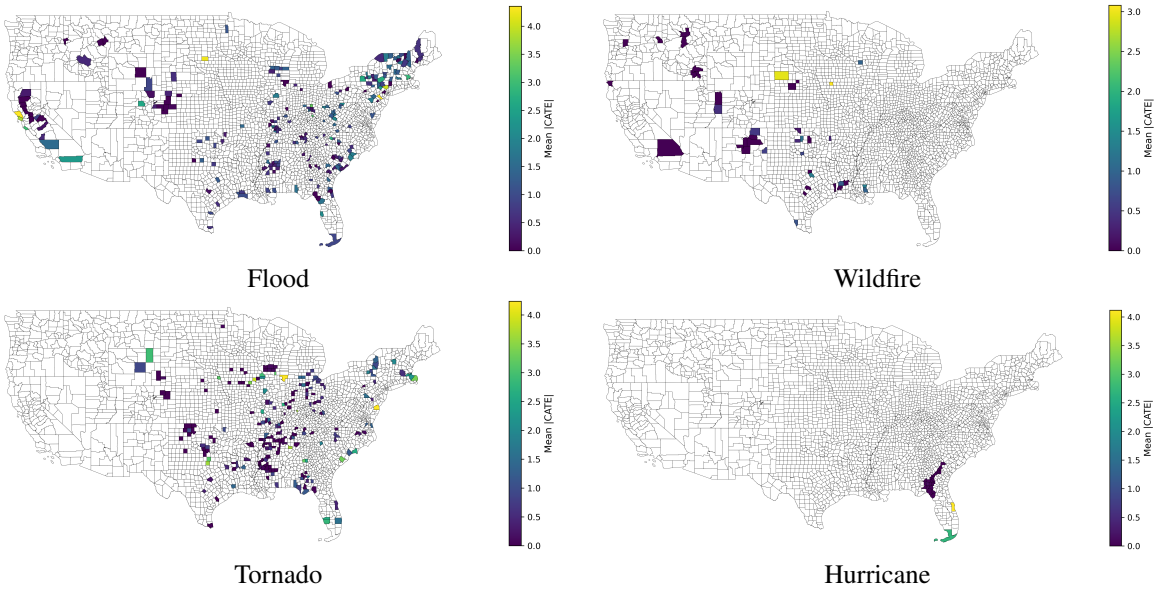


Figure 2: County-level mean absolute CATE estimates $|\widehat{\tau}_i|$ for 2023 across four hazard types. White indicates counties with no event; lighter colors represent higher estimated reporting bias.

K nearest control counties in the latent space Z. The resulting conditional average treatment effect is estimated as

$$\widehat{\tau}_i = \left| Y_{\text{obs},i} - \frac{1}{K} \sum_{j \in \mathcal{N}_K(i)} Y_{\text{ctrl},j} \right|.$$

This quantity represents a local estimate of the conditional average treatment effect evaluated at the county's latent content representation Z_i, and quantifies the magnitude of reporting bias while holding underlying physical characteristics fixed.

Results. Figure 2 reveals pronounced geographic heterogeneity in reporting bias across counties. Counties with no qualifying events appear in white. The county-level maps in Figure 2 indicate that hurricane-related reporting bias is concentrated along the coastal regions, particularly in Florida, where we might see a direct landfall. In contrast, we do not see a clear cluster of biased counties for other events. However, patches of hotspots, such as California for wildfires and the Tornado Alley in the central U.S for tornadoes, appear to yield low estimates for bias.

Figure 3 aggregates local CATE magnitudes by hazard type. Floods exhibit the largest average magnitude of reporting bias, followed by tornadoes, whereas wildfires and hurricanes show comparatively lower effects of reporting bias. This ordering is consistent with prior analyses based on SHELDUS and related loss databases, which document greater uncertainty in flood loss reporting relative to other events (Gall et al., 2009; Zhou et al., 2025).

6. Discussion

This work introduces proxy-guided measurement calibration, a framework for recovering true outcomes from systematically biased measurements. We formalize this problem through an explicit

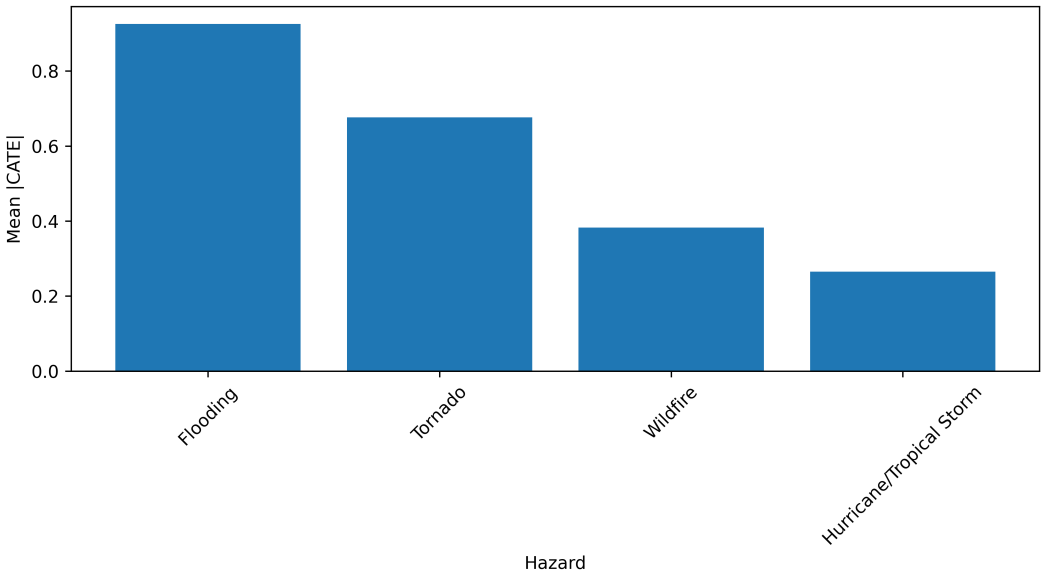


Figure 3: Mean absolute CATE values grouped by hazard type.

causal model in which measurement distortion is driven by latent bias factors that act on the observed outcome but are excluded from proxy variables. By separating the latent space into content-specific latents and bias-specific latents, where proxy variables depend only on the former, we provide a principled approach to identifying the magnitude of systematic miscalibration. Across different synthetic and semi-synthetic settings, we empirically show the performance of our framework, followed by a real-world case study.

Our semi-parametric approach, which combines co-trained latent representations with a matching-based non-parametric estimator, demonstrates a decent performance across a range of settings. Although the learned latent variables are identifiable only up to scale and permutation, this indeterminacy does not affect estimation because matching is performed exclusively in the latent space, while outcomes are compared on their original scale. As a result, the causal estimand remains invariant to latent rescaling or rotation. In contrast to parametric methods that propagate latent rescaling into outcome predictions, our estimator avoids outcome rescaling issues, leading to improved empirical performance.

Several limitations provide an avenue for future work. Firstly, the assumed error model remains restrictive. Although the real-world application relaxes several assumptions imposed in the synthetic and semi-synthetic settings, it still relies on the monotonicity assumption. Relaxing this assumption for identifying the magnitude of the bias is an important direction for future work. Moreover, our framework identifies conditional average treatment effects at the unit level, rather than individual treatment effects, reflecting fundamental limits imposed by the available data and assumptions. Our work finds many use cases where we can easily obtain proxy variables to correct for the data-generating bias. Extending this framework to other domains where outcomes are systematically mismeasured, such as public health surveillance, administrative records, and environmental monitoring, remains an important direction for future research.

Acknowledgements

This material is based upon work supported by, or in part by the U.S. Army Materiel Command under Grant Award Number W911NF24-2-0175 and by the U.S. Army Research Laboratory under Grant Award Number W911NF2020124. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies of the U.S. Army Materiel Command or the U.S. Army Research Laboratory.

References

Alberto Abadie and Guido W Imbens. Large sample properties of matching estimators for average treatment effects. *econometrica*, 74(1):235–267, 2006.

Alberto Abadie and Guido W Imbens. Bias-corrected matching estimators for average treatment effects. *Journal of business & economic statistics*, 29(1):1–11, 2011.

Gabriel Agostini, Emma Pierson, and Nikhil Garg. A bayesian spatial model to correct under-reporting in urban crowdsourcing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 21888–21896, 2024.

ASU Center for Emergency Management and Homeland Security. The Spatial Hazard Events and Losses Database for the United States, Version 23.0 [Online Database]. https://sheldus.org, 2025. Accessed: 2025-12-21.

Christopher F Brown, Steven P Brumby, Brookie Guzder-Williams, Tanya Birch, Samantha Brooks Hyde, Joseph Mazzariello, Wanda Czerwinski, Valerie J Pasquarella, Robert Haertel, Simon Ilyushchenko, et al. Dynamic world, near real-time global 10 m land use land cover mapping. *Scientific data*, 9(1):251, 2022.

Jiawei Fu and Donald P Green. Causal inference for experiments with latent outcomes: Key results and their implications for design and analysis. *arXiv preprint arXiv:2505.21909*, 2025.

Melanie Gall, Kevin A Borden, and Susan L Cutter. When do losses count? six fallacies of natural hazards loss data. *Bulletin of the American Meteorological Society*, 90(6):799–810, 2009.

Luke Guerdan, Amanda Coston, Kenneth Holstein, and Zhiwei Steven Wu. Counterfactual prediction under outcome measurement error. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1584–1598, 2023.

Pengrun Huang and Maggie Makar. Conditional differential measurement error: partial identifiability and estimation. In *NeurIPS workshop on causal machine learning for real world impact*, 2022.

Aapo Hyvarinen, Hiroaki Sasaki, and Richard Turner. Nonlinear ica using auxiliary variables and generalized contrastive learning. In *The 22nd international conference on artificial intelligence and statistics*, pages 859–868. PMLR, 2019.

Aapo Hyvärinen, Ilyes Khemakhem, and Ricardo Monti. Identifiability of latent-variable and structural-equation models: from linear to nonlinear. *Annals of the Institute of Statistical Mathematics*, 76(1):1–33, 2024.

Kosuke Imai and Teppei Yamamoto. Causal inference with differential measurement error: Nonparametric identification and sensitivity analysis. *American Journal of Political Science*, 54(2): 543–560, 2010.

Bohdan Kivva, Goutham Rajendran, Pradeep Ravikumar, and Bryon Aragam. Identifiability of deep generative models without auxiliary information. *Advances in Neural Information Processing Systems*, 35:15687–15701, 2022.

Zhi Liu, Uma Bhandaram, and Nikhil Garg. Quantifying spatial under-reporting disparities in resident crowdsourcing. *Nature Computational Science*, 4(1):57–65, 2024a.

Zhi Liu, Sarah Rankin, and Nikhil Garg. Identifying and addressing disparities in public libraries with bayesian latent variable modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 22258–22265, 2024b.

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.

Vishwali Mhasawade, Alexander D'Amour, and Stephen R Pfohl. A causal perspective on label bias. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 1282–1294, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704505. doi: 10.1145/3630106.3658972. URL https://doi.org/10.1145/3630106.3658972.

Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 34(11):8135–8153, 2023. doi: 10.1109/TNNLS.2022.3152527.

Jenny T van der Steen, Gerben Ter Riet, Cornelis A van den Bogert, and Lex M Bouter. Causes of reporting bias: a theoretical framework. *F1000Research*, 8:280, 2019.

Tyler J VanderWeele and Yige Li. Simple sensitivity analysis for differential measurement error. *American journal of epidemiology*, 188(10):1823–1829, 2019.

Yixin Wang, David Blei, and John P Cunningham. Posterior collapse and latent variable nonidentifiability. *Advances in neural information processing systems*, 34:5443–5455, 2021.

Xiaobo Xia, Tongliang Liu, Nannan Wang, Bo Han, Chen Gong, Gang Niu, and Masashi Sugiyama. Are anchor points really indispensable in label-noise learning? *Advances in neural information processing systems*, 32, 2019.

Dingling Yao, Shimeng Huang, Riccardo Cadei, Kun Zhang, and Francesco Locatello. The third pillar of causal analysis? a measurement perspective on causal representations. *arXiv preprint arXiv:2505.17708*, 2025.

Weijia Zhang, Lin Liu, and Jiuyong Li. Treatment effect estimation with disentangled latent factors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 10923–10930, 2021.

Yao Zhou, Christopher T Emrich, and Melanie Gall. Knowing more about losing more: Investigating spatial-temporal patterns of damage from storm-related hazards in the contiguous united states. *Annals of the American Association of Geographers*, pages 1–20, 2025.

Contents

Appendix A. Additional Experimental Discussion

A.1. Experiments on Synthetic Data

Before evaluating proxy-guided measurement calibration, we validate that the synthetic datasets conform to the intended causal data-generating process. These checks ensure that (i) the imposed structural assumptions are respected, (ii) the injected bias behaves as designed, and (iii) the latent variables are recoverable up to the expected indeterminacies. Together, these diagnostics establish that performance results in the main experiments reflect properties of the method rather than artifacts of data construction.

Tables 4–6 and Figures 4–5 summarize the synthetic data validation results. The bias latent error in Table 4 increases with the injected bias strength α, reflecting that the inferred bias representation is unconstrained and captures bias magnitude rather than acting as a bounded classifier. In contrast, Table 5 shows that proxy reconstruction error remains stable across α and noise models, confirming that proxies depend only on the content latents and are unaffected by bias. Table 6 demonstrates consistent recovery of the content latents up to scale and permutation. These quantitative results are complemented by distributional diagnostics in Figures 5 and 6, as well as qualitative latent alignment visualizations in Figure 4.

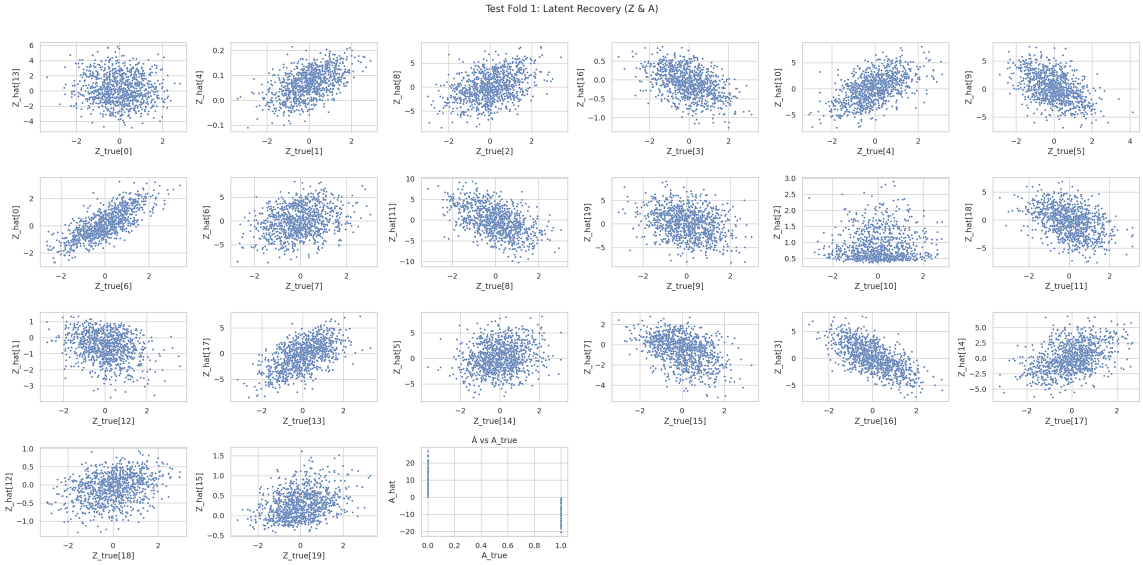


Figure 4: Scatter plot of the recovered content latents $\hat{Z}$ versus the true latents Z for a representative fold, after aligning dimensions using the closest permutation that maximizes permuted R^2.

Table 4: Mean±Std of a_{mae} grouped by (n, d_z, d_e) rows and (α, noise) columns.

			a_{mae}					
			$\alpha = 1$		$\alpha = 5$		$\alpha = 10$	
n	d_z	d_e	Gaussian	Poisson (scaled)	Gaussian	Poisson (scaled)	Gaussian	Poisson (scaled)
500	1	10	1.29±0.33	1.29±0.32	1.39±0.47	1.38±0.49	1.48±0.50	1.46±0.51
500	2	10	1.52±0.22	1.54±0.21	1.63±0.34	1.64±0.34	1.81±0.39	1.85±0.43
500	5	10	1.37±0.27	1.29±0.38	1.09±0.31	1.18±0.28	1.19±0.29	1.13±0.37
500	10	10	1.68±0.34	1.68±0.35	1.72±0.47	1.82±0.40	1.83±0.41	1.83±0.44
500	20	10	1.64±0.32	1.63±0.22	1.57±0.39	1.57±0.39	1.51±0.61	1.66±0.43
1000	1	10	1.49±0.18	1.49±0.19	1.60±0.36	1.61±0.35	1.83±0.36	1.82±0.36
1000	2	10	1.63±0.16	1.64±0.16	1.79±0.39	1.80±0.38	2.24±0.46	2.25±0.45
1000	5	10	1.63±0.05	1.65±0.08	1.51±0.53	1.39±0.49	1.77±0.41	1.66±0.53
1000	10	10	1.93±0.35	1.75±0.36	1.46±0.43	2.12±0.44	2.10±0.64	2.23±0.45
1000	20	10	1.88±0.20	1.88±0.19	1.61±0.33	1.63±0.33	1.95±0.45	1.96±0.44
2500	1	10	1.75±0.29	1.77±0.31	2.12±0.48	2.10±0.47	2.59±0.48	2.60±0.49
2500	2	10	1.90±0.13	1.89±0.14	2.73±0.35	2.67±0.38	3.30±0.41	3.22±0.40
2500	5	10	1.70±0.50	1.87±0.19	2.74±0.55	2.18±1.00	2.96±0.53	2.76±0.96
2500	10	10	2.22±0.37	2.21±0.39	3.05±0.53	3.09±0.46	3.24±0.53	3.10±0.88
2500	20	10	2.16±0.22	2.17±0.20	2.07±0.46	2.09±0.46	3.04±0.58	3.03±0.58
5000	1	10	1.95±0.26	1.91±0.23	2.50±0.49	2.51±0.51	3.66±0.50	3.64±0.49
5000	2	10	1.91±0.10	1.92±0.10	3.74±0.33	3.72±0.36	4.42±0.42	4.42±0.43
5000	5	10	2.29±0.91	2.19±0.96	3.49±0.43	3.61±0.54	4.18±0.55	4.18±0.53
5000	10	10	2.29±0.60	2.22±0.51	4.01±0.54	4.10±0.56	4.41±0.55	4.50±0.55
5000	20	10	2.50±0.16	2.49±0.16	1.89±0.96	1.93±1.08	4.28±0.57	4.27±0.56
10000	1	10	1.13±0.17	1.14±0.18	1.86±0.72	1.85±0.71	4.23±0.58	4.20±0.61
10000	2	10	1.24±0.07	1.22±0.06	4.07±0.53	4.15±0.49	5.44±0.31	5.40±0.33
10000	5	10	1.58±1.50	1.14±0.09	3.92±1.13	4.17±1.18	5.40±0.61	4.90±1.38
10000	10	10	1.30±0.35	1.41±0.45	5.12±0.55	5.32±0.78	6.46±1.06	6.13±0.58
10000	20	10	1.83±0.15	1.82±0.15	1.02±0.59	1.13±0.72	5.42±0.52	5.41±0.58

Table 5: Mean±Std of $rmse_{proxies}$ grouped by (n, d_z, d_e) rows and (α, noise) columns.

			$rmse_{proxies}$					
			$\alpha = 1$		$\alpha = 5$		$\alpha = 10$	
n	d_z	d_e	Gaussian	Poisson (scaled)	Gaussian	Poisson (scaled)	Gaussian	Poisson (scaled)
500	1	10	0.20±0.03	0.20±0.03	0.20±0.03	0.20±0.03	0.20±0.03	0.20±0.03
500	2	10	0.21±0.03	0.20±0.02	0.21±0.03	0.20±0.02	0.21±0.03	0.20±0.02
500	5	10	0.19±0.02	0.19±0.02	0.19±0.02	0.19±0.02	0.19±0.02	0.19±0.02
500	10	10	0.15±0.02	0.15±0.02	0.15±0.02	0.15±0.02	0.15±0.02	0.15±0.02
500	20	10	0.26±0.03	0.25±0.03	0.26±0.03	0.25±0.03	0.26±0.03	0.25±0.03
1000	1	10	0.19±0.06	0.19±0.09	0.19±0.06	0.19±0.09	0.19±0.06	0.19±0.09
1000	2	10	0.18±0.03	0.17±0.02	0.18±0.03	0.17±0.02	0.18±0.03	0.17±0.02
1000	5	10	0.16±0.02	0.17±0.01	0.16±0.02	0.17±0.01	0.16±0.02	0.17±0.01
1000	10	10	0.13±0.01	0.13±0.01	0.17±0.01	0.13±0.01	0.13±0.01	0.13±0.01
1000	20	10	0.20±0.01	0.20±0.01	0.20±0.01	0.20±0.01	0.20±0.01	0.20±0.01
2500	1	10	0.14±0.01	0.14±0.01	0.14±0.01	0.14±0.01	0.14±0.01	0.14±0.01
2500	2	10	0.14±0.01	0.14±0.01	0.14±0.01	0.15±0.01	0.14±0.01	0.15±0.01
2500	5	10	0.13±0.01	0.14±0.01	0.13±0.01	0.14±0.01	0.14±0.01	0.14±0.01
2500	10	10	0.12±0.00	0.12±0.00	0.12±0.00	0.12±0.00	0.12±0.01	0.12±0.00
2500	20	10	0.17±0.01	0.16±0.01	0.17±0.01	0.17±0.01	0.17±0.01	0.17±0.01
5000	1	10	0.12±0.01	0.13±0.01	0.12±0.01	0.13±0.01	0.12±0.01	0.13±0.01
5000	2	10	0.13±0.01	0.12±0.01	0.13±0.01	0.12±0.01	0.13±0.01	0.12±0.01
5000	5	10	0.12±0.00	0.12±0.01	0.12±0.00	0.12±0.01	0.12±0.00	0.12±0.01
5000	10	10	0.11±0.00	0.11±0.00	0.11±0.00	0.11±0.00	0.11±0.00	0.11±0.00
5000	20	10	0.14±0.01	0.14±0.01	0.14±0.01	0.14±0.01	0.14±0.01	0.14±0.01
10000	1	10	0.11±0.01	0.11±0.01	0.11±0.01	0.11±0.01	0.11±0.01	0.11±0.01
10000	2	10	0.11±0.01	0.11±0.01	0.11±0.01	0.11±0.01	0.11±0.01	0.11±0.01
10000	5	10	0.10±0.00	0.10±0.00	0.10±0.00	0.10±0.00	0.10±0.00	0.10±0.00
10000	10	10	0.09±0.00	0.09±0.00	0.09±0.00	0.09±0.00	0.09±0.00	0.09±0.00
10000	20	10	0.11±0.00	0.11±0.00	0.11±0.00	0.11±0.00	0.11±0.00	0.11±0.00

Table 6: Mean±Std of z_{mae} grouped by (n, d_z, d_e) rows and (α, noise) columns.

			z_{mae}					
			$\alpha = 1$		$\alpha = 5$		$\alpha = 10$	
n	d_z	d_e	Gaussian	Poisson (scaled)	Gaussian	Poisson (scaled)	Gaussian	Poisson (scaled)
500	1	10	1.84±0.86	1.82±0.86	1.84±0.86	1.82±0.86	1.84±0.86	1.82±0.86
500	2	10	1.42±0.41	1.50±0.37	1.42±0.41	1.50±0.37	1.42±0.41	1.50±0.37
500	5	10	1.22±0.19	1.23±0.19	1.22±0.19	1.23±0.19	1.22±0.19	1.23±0.19
500	10	10	1.14±0.06	1.14±0.07	1.14±0.06	1.14±0.07	1.14±0.06	1.14±0.07
500	20	10	1.09±0.10	1.10±0.11	1.09±0.10	1.10±0.11	1.09±0.10	1.10±0.11
1000	1	10	2.07±0.92	2.08±0.94	2.07±0.92	2.08±0.94	2.07±0.92	2.08±0.94
1000	2	10	1.59±0.48	1.59±0.48	1.59±0.48	1.59±0.48	1.59±0.48	1.59±0.48
1000	5	10	1.29±0.20	1.29±0.21	1.29±0.20	1.29±0.21	1.29±0.20	1.29±0.21
1000	10	10	1.22±0.08	1.22±0.08	1.12±0.06	1.22±0.08	1.22±0.08	1.22±0.08
1000	20	10	1.18±0.10	1.17±0.11	1.18±0.10	1.17±0.11	1.18±0.10	1.17±0.11
2500	1	10	1.94±0.95	1.95±0.93	1.94±0.95	1.95±0.93	1.94±0.95	1.95±0.93
2500	2	10	1.66±0.41	1.64±0.37	1.66±0.41	1.64±0.38	1.67±0.41	1.64±0.38
2500	5	10	1.40±0.20	1.40±0.21	1.40±0.20	1.40±0.21	1.40±0.21	1.40±0.21
2500	10	10	1.32±0.06	1.32±0.06	1.32±0.06	1.32±0.06	1.32±0.06	1.32±0.06
2500	20	10	1.28±0.09	1.28±0.09	1.28±0.09	1.28±0.09	1.27±0.09	1.28±0.09
5000	1	10	2.24±0.82	2.23±0.83	2.24±0.82	2.23±0.83	2.24±0.82	2.23±0.83
5000	2	10	1.75±0.48	1.75±0.48	1.75±0.48	1.75±0.48	1.75±0.48	1.75±0.48
5000	5	10	1.49±0.23	1.49±0.23	1.49±0.23	1.49±0.23	1.49±0.23	1.49±0.23
5000	10	10	1.44±0.16	1.44±0.16	1.44±0.16	1.44±0.16	1.44±0.16	1.44±0.16
5000	20	10	1.46±0.10	1.46±0.09	1.46±0.10	1.46±0.09	1.46±0.10	1.46±0.09
10000	1	10	2.20±0.86	2.21±0.84	2.20±0.86	2.21±0.84	2.20±0.86	2.21±0.84
10000	2	10	1.77±0.34	1.76±0.35	1.77±0.34	1.76±0.35	1.77±0.34	1.76±0.35
10000	5	10	1.57±0.32	1.56±0.33	1.57±0.32	1.56±0.33	1.57±0.32	1.56±0.33
10000	10	10	1.48±0.16	1.47±0.15	1.48±0.16	1.47±0.15	1.48±0.16	1.47±0.15
10000	20	10	1.66±0.14	1.67±0.13	1.66±0.14	1.67±0.13	1.66±0.14	1.67±0.13

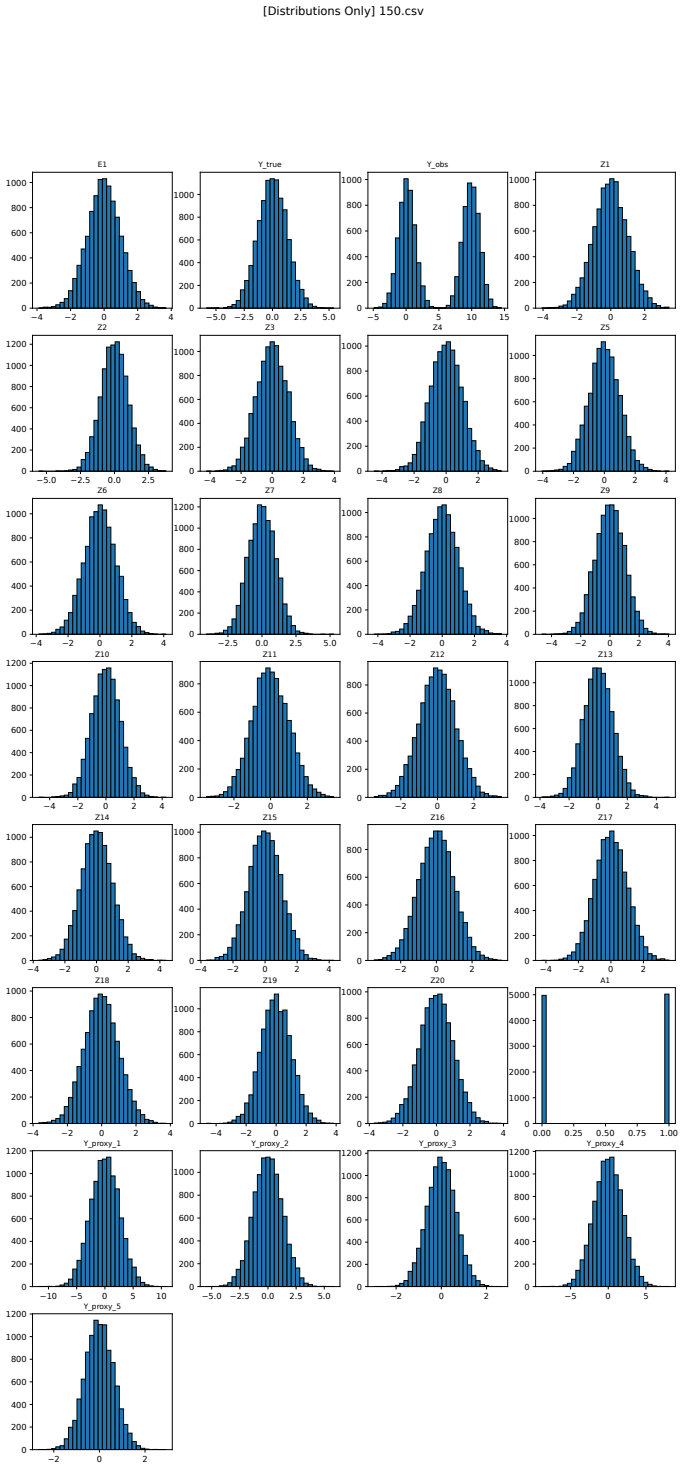


Figure 5: **Sanity Check for the Distribution.** Histograms for all key variables: E_1, Z, A, Y_{true}, Y_{obs}, and proxies for a given configuration. Ensures data follows the intended generative model.

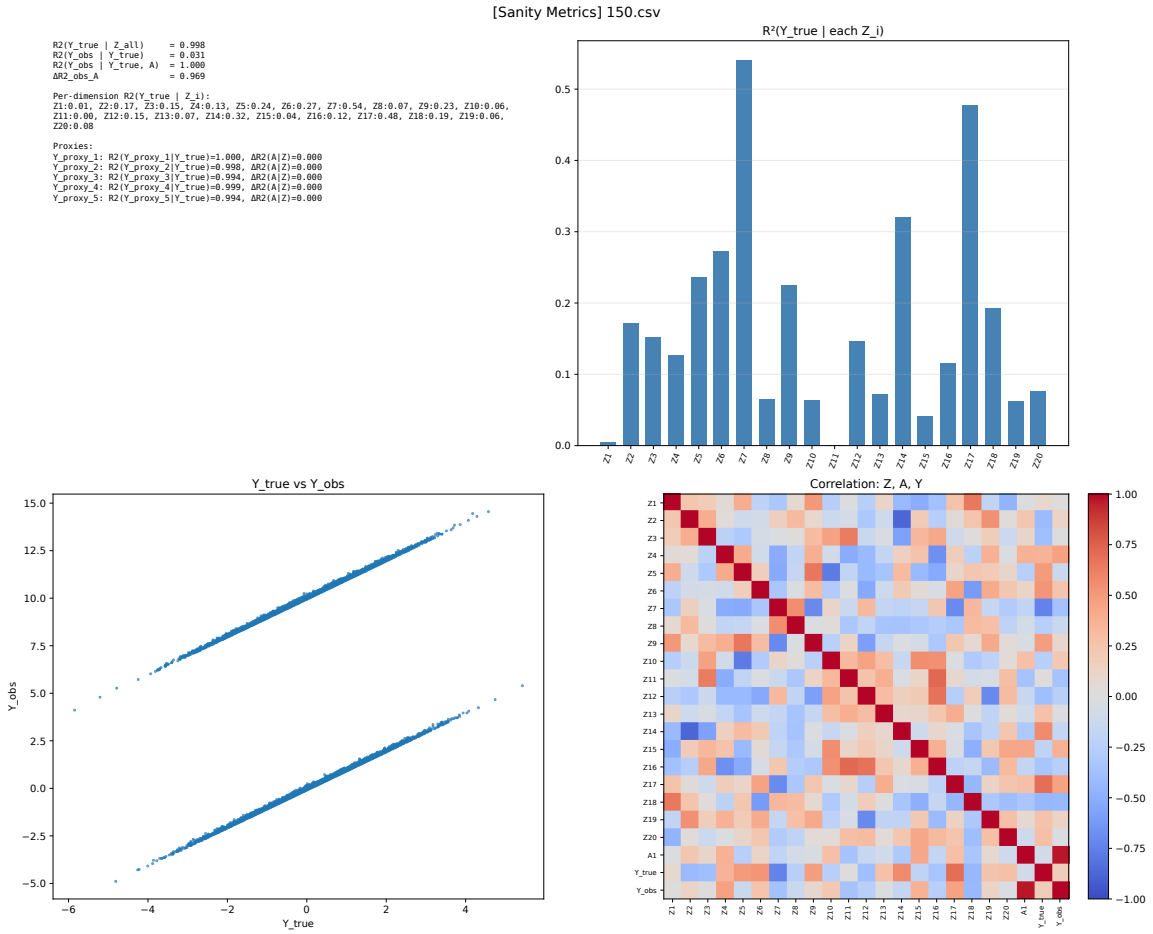


Figure 6: **Sanity Check - Core Metrics and Correlations.** This page summarizes the main structural sanity metrics: $R^2(Y_{\text{true}} \mid Z)$, $R^2(Y_{\text{obs}} \mid Y_{\text{true}})$, the effect of bias latents via ΔR^2, per-dimension R^2 for each Z_i, proxy alignment metrics, as well as the correlation matrix over $\{Z, A, Y\}$. These diagnostics confirm that the dataset correctly encodes both latent content and latent bias structure.

A.2. Experiments on Semi-Synthetic Datasets

To evaluate proxy-guided calibration in settings that more closely resemble real-world predictive tasks, we construct semi-synthetic datasets from two randomized controlled trials: the JOBS job training experiment and the Oregon Health Insurance Experiment (OHIE). In both datasets, the environment variables and proxy measurements arise naturally from high-quality experimental data, while a controlled amount of reporting bias is injected synthetically. This design preserves the advantages of real-world covariate structure and proxy behavior, while providing access to known ground-truth bias levels for benchmarking.

JOBS (Employment Training). The JOBS dataset originates from the National Supported Work (NSW) program, later merged with the PSID comparison sample to create a rich set of demographic and economic covariates. Following the preprocessing in our pipeline, the outcome is 1978 earnings (log-transformed), and the proxies consist of pretreatment earnings from 1974 and 1975, which are unaffected by reporting bias and thus serve as clean measurements of underlying economic status. Environment variables include treatment assignment and demographic attributes such as age, education, race, marital status, and prior earnings. After preprocessing, the dataset contains a fully numeric table with the structure shown in Table 8. Bias is injected by adding αA to the standardized outcome, where A is a synthetically generated latent bias variable derived only from environment covariates.

OHIE (Medicaid Lottery). The OHIE dataset is derived from the Oregon Medicaid expansion via a lottery-based allocation mechanism. Treatment is the randomized selection to receive Medicaid coverage. The outcome is log-transformed emergency department spending. Proxies represent biomedical and mental-health measurements obtained during in-person assessments (blood pressure, HDL, A1C, and depression score), all of which serve as clinical proxies for underlying health status. Environment covariates include variables from the pre-randomization descriptive survey, state-level program participation, prior emergency department visits, and selected in-person clinical measures. The preprocessing pipeline filters environment variables by variance and removes collinear features, producing the unified schema shown in Table 7. As with JOBS, bias is introduced by adding αA to the standardized observed outcome, where A depends only on selected environment variables.

Together, JOBS and OHIE provide complementary settings: an employment-focused economic context with proxies reflecting income history, and a healthcare context with proxies reflecting physiological health markers. These semi-synthetic datasets enable controlled evaluation of measurement error methods while retaining realistic feature distributions, proxy relationships, and treatment assignment from real experimental studies.

Table 7: Schema of the OHIE semi-synthetic dataset after preprocessing. Environment variables E include treatment, state-program participation, prior emergency-department utilization, and in-person clinical measures. Proxies are clean biomedical measurements and the outcome is log-transformed ED spending.

Variable Type	Name(s)	Description
Outcome	Y_{obs}	Log-transformed emergency department spending (0–11.13).
Proxies	$Y_{\text{proxy},1}$–$Y_{\text{proxy},5}$	Biomedical measurements: systolic BP, diastolic BP, HDL, A1C, PHQ depression score.
Environment E: Treatment	E_1 (T)	Lottery selection indicator (0/1).
Environment E: State program participation	E_2: SNAP household (prenotify) E_3: TANF household (prenotify)	SNAP/TANF participation prior to randomization (large variation: 286–2952 unique values).
Environment E: Prior ED utilization	E_4: num_visit_pre_cens_ed E_5: num_on_pre_cens_ed E_6: num_off_pre_cens_ed E_7: num_chron_pre_cens_ed E_8: num_inj_pre_cens_ed E_9: num_hiun_pre_cens_ed E_{10}: num_loun_pre_cens_ed E_{11}: ed_charg_tot_pre_ed E_{12}: charg_tot_pre_ed	Utilization counts and total pre-period ED charges. Reflects burden of chronic, injury-related, heart, head, and psychiatric emergency visits. Many exhibit wide ranges (e.g., \$0–\$180,054 in total charges).
Environment E: In-person clinical measures	E_{13}: tot_med_spend_other_inp	Supplementary in-person medical expenditure measure (0–92,475 across subjects).
Bias latent	A	Synthetic binary latent generated from a logistic function of all environment variables.

Table 8: Schema of the JOBS semi-synthetic dataset after preprocessing. Environment variables consist of treatment and demographic covariates; proxies are clean pretreatment earnings; the outcome is log-transformed 1978 earnings.

Variable Type	Name(s)	Description
Outcome	Y_{obs}	Log-transformed 1978 earnings ($\log(1 + \text{re78})$), range 0–11.01, with 457 unique values.
Proxies	$Y_{\text{proxy},1}$, $Y_{\text{proxy},2}$	Clean pretreatment earnings: • $Y_{\text{proxy},1}$ (re74): range 0–35,040, 358 unique values • $Y_{\text{proxy},2}$ (re75): range 0–25,142, 356 unique values These proxies are unaffected by the injected measurement bias.
Environment E: Treatment	E_1 (treat)	Random assignment to the job-training intervention (binary 0/1).
Environment E: Demographics	E_2: age (16–55; 40 unique values) E_3: educ (0–18; 19 unique values) E_4: black (0/1) E_5: hispan (0/1) E_6: married (0/1) E_7: nodegree (0/1)	Baseline demographic variables from the NSW/JOBS dataset, representing socioeconomic and educational background.
Bias latent	A	Synthetic binary latent generated from a logistic function of the environment variables, used to inject controlled bias into the observed outcome.

A.3. More Details on Real-world Case Study: SHELDUS

The Spatial Hazard Events and Losses Database for the United States (SHELDUS) provides county-level records of direct economic losses attributed to natural hazards, including thunderstorms, floods, hurricanes, wildfires, and winter storms. Each entry aggregates property damage, crop damage, and casualty counts reported by local agencies, making the dataset a rich source for studying disaster impacts but also one in which measurement error is well documented. Loss estimates often depend on heterogeneous reporting practices, varying administrative capacity across counties, and inconsistent assessment procedures, all of which introduce systematic biases into the observed outcomes. To support our calibration task, we augment SHELDUS with environment variables capturing socioeconomic characteristics (e.g., income, population, educational attainment), hazard-type indicators, and exposure features, while also incorporating proxy variables such as remote-sensing damage indicators or alternative loss metrics when available. This setting provides a realistic testbed in which the true underlying losses are unobserved, reporting bias is substantial, and proxy-guided calibration has the potential to correct distortions in disaster loss estimation across heterogeneous counties.

Table 9: Variables used in the SHELDUS-based real-world case study after preprocessing.

Category	**Variables and description**
Identifiers	FIPS (county identifier), Year, Month
Hazard indicators	Hazard type (Flooding, Hurricane/Tropical Storm, Tornado, Wildfire), one-hot encoded as $E_{\text{hazard},k}$
Observed outcome	Y_{obs}: reported county-level economic loss (property damage, inflation-adjusted to 2023 USD), log-transformed and standardized
Proxy variables	$Y_{\text{proxy},1}$: fraction of land transitioning from bare ground to water $Y_{\text{proxy},2}$: fraction of built area transitioning to damage-related classes $Y_{\text{proxy},3}$: fraction of built area transitioning to water $Y_{\text{proxy},4}$: fraction of vegetation transitioning to damage-related classes $Y_{\text{proxy},5}$: fraction of vegetation transitioning to water (all proxies log-scaled with factor 10^6 and standardized)
Socioeconomic environment (E)	Population, median age, median household income, per-capita income, median home value, median rent, poverty count and rate, education counts (high school, bachelor's) and percentages, education skew
Housing and labor characteristics	Housing units, household count, housing unit density, household size, labor force size, employment and unemployment counts and rates, labor force participation rate

Appendix B. Remarks on Identifiability

Our approach combines latent-variable modeling with causal adjustment to estimate bias-adjusted outcomes. This raises two distinct, but often conflated questions of identifiability: (i) identifiability of the latent representations learned by variational autoencoders, and (ii) identifiability of the causal estimand of interest once an appropriate representation is available. We first state the formal identification conditions for the content latent Z (Section B.1), the bias indicator A (Section B.2), and then discuss how these relate to broader identifiability results in the VAE and causal representation learning literatures.

B.1. Identification of the Content Latent Z

We justify identifiability of the latent content variable Z using the nonlinear ICA framework of Hyvarinen et al. (2019), which establishes conditions under which latent sources can be recovered from nonlinear mixtures given an observed auxiliary variable.

In our setup, the proxy measurements are generated as $Y_{\text{proxy}} = f(Z, \epsilon_{\text{proxy}})$, where f is a nonlinear mixing function and ϵ_{proxy} is independent noise. The environment variable E serves as the observed auxiliary variable. The required assumptions are:

1. **Conditional independence given Z:** $Y_{\text{proxy}} \perp\!\!\!\perp Y_{\text{obs}} \mid Z$. This follows directly from the generative model (Figure 1a), in which the proxy depends on Z alone.

2. **Invertible mixing:** $Y_{\text{proxy}} = f(Z, \epsilon_{\text{proxy}})$ with f invertible.

3. **Auxiliary variability:** The environment variable E induces non-degenerate changes in the conditional distribution $p(Z \mid E)$ across units. Concretely, for sufficiently many distinct values of E, the family $\{p(Z \mid E = e)\}_e$ is not related by trivial (e.g., location-only) shifts.

Under the above assumptions, the results on nonlinear ICA identifiability from Hyvarinen et al. (2019) implies that Z is identifiable from (Y_{proxy}, E) up to component-wise invertible transformations and permutation. That is, any two models satisfying these assumptions yield latent representations related by an element-wise reparameterization and reordering of components.

This level of identifiability is sufficient for our downstream estimation, because all causal estimands considered in this work are invariant to such transformations.

B.2. Identification of the Bias Indicator A

We identify the bias magnitude α under the additive error model introduced in Section 4.4. The required assumptions are:

1. **Additive error model:** $Y_{\text{obs}} = Y_{\text{true}} + \alpha\, A$, where $Y_{\text{true}} = g(Z, E) + \epsilon$ with $\mathbb{E}[\epsilon \mid Z, E] = 0$, and $A \in \{0, 1\}$.

2. **Overlap:** $0 < P(A = 1 \mid E = e, Z = z) < 1$ for all (e, z) in the support. This ensures that both biased and unbiased units exist across the covariate space, so that matched comparisons are well-defined.

3. **Anchors:** A subset of units have known labels for bias presence (i.e., known A_i). Anchors fix the mapping between the real-valued latent score $\hat{A}$ produced by the VAE encoder and the binary regime labels $\{0, 1\}$.

Without anchors, the binary latent A is identifiable only up to label swapping: the likelihood under (A, α) is identical to that under $(1 - A, -\alpha)$. Anchors resolve this ambiguity by defining which regime corresponds to $A = 0$ (unbiased) and which to $A = 1$ (biased). In practice, the threshold on $\hat{A}$ that separates the two regimes is selected on the validation set using anchor information when available, or by leveraging domain knowledge about the expected direction of bias (e.g., the sign restriction $\alpha \geq 0$ in Table 1).

We note that our identification target is the bias magnitude α, not individual treatment effects. The estimand is a conditional mean (CATE-style) and its aggregation (ATE-style) under the restricted error model above. Without additional measurement information, causal effects are not identifiable when the treatment or outcome is latent; our approach therefore relies on the explicit measurement assumptions stated here.

Latent identifiability with VAEs. It is well known that latent variables in deep generative models are not identifiable without additional structure. Recent work formalizes this limitation and provides a hierarchy of identifiability results under increasingly strong assumptions on the prior and decoder. In particular, under mixture priors over latents and piecewise affine decoders (such as ReLU networks), latent variables are identifiable up to an invertible affine transformation. This is the weakest but most general identifiability guarantee in this hierarchy (Table 1 in (Kivva et al. (2022)), and it applies directly to standard VAE architectures.

Importantly, this form of identifiability does not imply semantic disentanglement or recovery of a canonical coordinate system for the latent variables. Rather, it guarantees that the learned representation is equivalent to the true latent variable up to affine reparameterization. Throughout this work, we explicitly adopt this weakest guarantee and do not assume any stronger notion of latent identifiability.

Why affine identifiability is sufficient for our setting. Our method uses the learned latent Z exclusively as a *content or adjustment representation*. All downstream operations: nearest-neighbor matching in latent space, conditioning in regression, and computation of absolute conditional average treatment effects are invariant under affine transformations of Z. Consequently, the affine identifiability result is sufficient for our purposes: the latent representation need not be uniquely defined, only stable up to transformations that preserve relative geometry (Kivva et al. (2022)).

This distinction is crucial. We do not require recovery of the true causal variables in a structural sense, nor do we require disentanglement of all generative factors. Instead, we require that the representation preserve the information necessary to block spurious associations between the bias variable and the observed outcome.

A measurement-model perspective on learned representations. A complementary view is provided by the measurement-model framework for causal representation learning. From this perspective, the learned latent $\widehat{Z}$ is interpreted as a measurement variable generated from the true (unobserved) content variable Z via an unknown measurement function. A representation is said to be *causally valid* for a downstream estimand if it can be used as a drop-in replacement without altering the estimand's value (Yao et al. (2025)).

Crucially, standard adjustment and matching estimands are invariant to invertible reparameterizations of adjustment variables. If $\widehat{Z} = h(Z)$ for a bijective (or affine) function h, then backdoor-style functionals expressed in terms of Z can equivalently be expressed in terms of $\widehat{Z}$. By contrast, when the treatment or outcome itself is only identified up to an unknown reparameterization, average treatment effects are generally not identified without additional scale-setting information. In our setting, Z is used strictly as an adjustment representation, while Y_{obs} is directly observed, placing us in the invariant regime.

Appendix C. Model Architecture and Training Details

This appendix provides additional details on the representation-learning model used in our proxy-guided calibration framework. The model consists of two coordinated variational autoencoders, a proxy VAE (ZVAE) that learns latent content factors and a bias VAE (AVAE) that captures latent reporting bias affecting the observed outcome. The architecture, objectives, and training protocol described below apply uniformly across synthetic, semi-synthetic, and real-world datasets.

C.1. Two-Stage Latent Variable Architecture

Proxy VAE (ZVAE). The first stage models the relationship between proxies $\{Y_{\text{proxy},k}\}_{k=1}^{K}$ and the latent content variable Z. The encoder receives the proxy vector and (optionally) environment covariates E, producing mean and variance parameters for a Gaussian posterior $q_\phi(Z \mid Y_{\text{proxy}}, E)$. The decoder reconstructs the proxies independently via a diagonal Gaussian likelihood. The ZVAE thereby isolates a low-dimensional content representation that is predictive of both Y_{true} (in synthetic and semi-synthetic setups) and the systematic components of the proxies in real data. All encoders and decoders are two-layer MLPs with ReLU activations, batch normalization, and hidden dimension 256.

Bias VAE (AVAE). The second stage models the observed outcome Y_{obs} as depending on both the content latent Z and an additional bias latent A. The AVAE encoder takes (Y_{obs}, Z, E) as input and infers a posterior distribution $q_\psi(A \mid Y_{\text{obs}}, Z, E)$. The decoder reconstructs Y_{obs} given (Z, A) through a Gaussian likelihood with learned variance. This stage forces A to capture variation in Y_{obs} not explained by Z alone, thus corresponding to systematic measurement error or reporting bias. The architecture matches that of ZVAE, using two-layer MLPs with hidden dimension 256.

C.2. Training Objective

Each VAE is trained with a standard evidence lower bound (ELBO):

$$\mathcal{L}_{\text{VAE}} = \mathbb{E}_{q(z)}[-\log p(x \mid z)] \; + \; \beta \, D_{\text{KL}}(q(z) \,\|\, p(z)),$$

where β is a tunable KL weight. For ZVAE, x corresponds to the proxy vector; for AVAE, it corresponds to Y_{obs}. In all experiments we use a spherical standard normal prior for Z and A. To ensure numerical stability across datasets with different scales, all features except Y_{obs} are standardized; Y_{obs} is optionally log-transformed according to dataset rules.

C.3. Training and Evaluation Protocol

All experiments are conducted using k-fold cross-validation with $k = 10$. Models are reinitialized and trained independently on each fold. During training, reconstruction losses, latent recovery metrics, and diagnostic curves are logged for each fold. Evaluation is performed on held-out splits using the following metrics:

- Permuted R^2 for reconstruction of Z and A,
- RMSE for proxy reconstruction and Y_{obs} reconstruction,
- Absolute error in the estimated bias parameter α (synthetic and semi-synthetic),

Within each fold, model selection is based on minimizing validation reconstruction loss.

Models are trained using the Adam optimizer with learning rate 10^{-3} and batch size 512. Training is performed for a fixed number of epochs (50 for synthetic datasets and up to 100 for real-world datasets), without early stopping. The variational objective is optimized via the standard ELBO with a fixed KL weight, which was sufficient to prevent latent collapse in all settings considered due to the low-dimensional latent structure and proxy-based supervision.

C.4. Implementation Details

All experiments use PyTorch models implemented with modular encoder/decoder components. Tabular datasets follow a unified interface that partitions covariates into $(E, Y_{\text{proxy}}, Y_{\text{obs}})$ and standardizes them according to dataset-specific rules. The training pipeline handles fold splits, metric computation, checkpointing, ELBO logging, and PDF report generation. Figures in the main text and appendix are produced using seaborn and matplotlib, and performance summaries are exported as both CSV and LaTeX tables.

Appendix D. Generative AI Disclosure

Portions of this manuscript were prepared with the assistance of generative AI tools (e.g., OpenAI's ChatGPT) for tasks such as coding, figure generation, and help with drafting. All AI-assisted outputs were carefully reviewed and validated by the authors. The conceptual framework, experimental design, and scientific conclusions are solely the responsibility of the authors.