IDENTIFIABLE ESTIMATION OF CAUSAL CONCEPT EFFECTS UNDER VISUAL LATENT CONFOUNDING

Thomas Melistas[1,2] **Damian Machlanski**[3,4] **Kurt Butler**[3,4] **Angelos Korakitis**[1,2] **Nikos Spyrou**[1,2]
Athanasios Vlontzos[5] **Yannis Panagakis**[1,2] **Sotirios A. Tsaftaris**[2,3,4] **Giorgos Papanastasiou**[2,6]
[1]National and Kapodistrian University of Athens [2]Archimedes, Athena Research Center, Greece
[3]The University of Edinburgh [4]Causality in Healthcare AI Hub (CHAI) [5]Monzo
[6]Mathematics Research Centre, Academy of Athens
th.melistas@athenarc.gr

ABSTRACT

Estimating the causal effect of human-interpretable visual concepts on outcomes is essential for auditing classifiers and assessing bias in image datasets. However, existing estimators typically assume unconfoundedness, a condition rarely met in practice, as concept annotations are seldom exhaustive. We formalize the problem of visual latent confounding, where unannotated factors manifest as high-dimensional visual signatures that jointly influence observed concepts and outcomes. We present UnCoVAEr (Unobserved Confounding Variational AutoEncoder), a latent-variable model that learns identifiable confounder representations from images. By leveraging observed concepts and outcomes as auxiliary variables, we prove that UnCoVAEr identifies representations sufficient for backdoor adjustment under standard assumptions. Empirically, UnCoVAEr achieves lower causal concept effect estimation bias on MorphoMNIST and CelebA benchmarks, outperforming feature-adjustment, counterfactual, and latent-variable baselines.

1 INTRODUCTION

Understanding how different visual concepts causally influence outcome labels is critical for detecting spurious correlations in classifiers and quantifying dataset biases (Jones et al., 2024; Madras et al., 2019; Di Stefano et al., 2020). Goyal et al. (2020) formalized Causal Concept Effect (CaCE) as the expected change in a model's prediction after intervening on a concept. However, their approach yields unbiased estimates only under the assumption that all common causes of a concept and the outcome are observed. This assumption is frequently violated in practice, as concept annotations typically capture only a subset of confounding factors.

In medical imaging, scanner hardware or acquisition protocols often correlate with both anatomical content and clinical diagnoses (Zech et al., 2018; Badgeley et al., 2018; DeGrave et al., 2021). A model may appear to rely on a particular anatomical feature for diagnosis, when in reality this association is confounded by imaging conditions that simultaneously affect feature visibility and diagnostic labels. As illustrated in Figure 1, failing to account for such hidden factors—like age in facial analysis—opens backdoor paths that bias observational estimates. We hypothesize that unobserved confounders, ranging from demographic shifts (Castro et al., 2020) to systematic labeling biases (Lingenfelter et al., 2022), often manifest as high-dimensional visual signatures in images.

Existing deep latent-variable models for causal inference, such as CEVAE (Louizos et al., 2017) and TEDVAE (Zhang et al., 2020), learn representations for adjustment but lack identifiability guarantees, leaving learned representations susceptible to misalignment with true confounders (Rissanen & Marttinen, 2021). Meanwhile, proximal causal inference methods (Tchetgen et al., 2020; Kompa et al., 2022; Xu et al., 2021) require access to two distinct types of proxies that satisfy certain completeness conditions (Miao et al., 2018). Another line of work treats images as observed covariates (Jerzak et al., 2023; Kumar et al., 2023) rather than proxies of latent confounders, potentially leading to weak overlap and unstable estimation (D'Amour et al., 2021). We bridge this gap by leveraging images as a rich high-dimensional proxy, while ensuring identifiable recovery of adjustment-relevant latent factors (see Appendix A for an extended related work discussion).

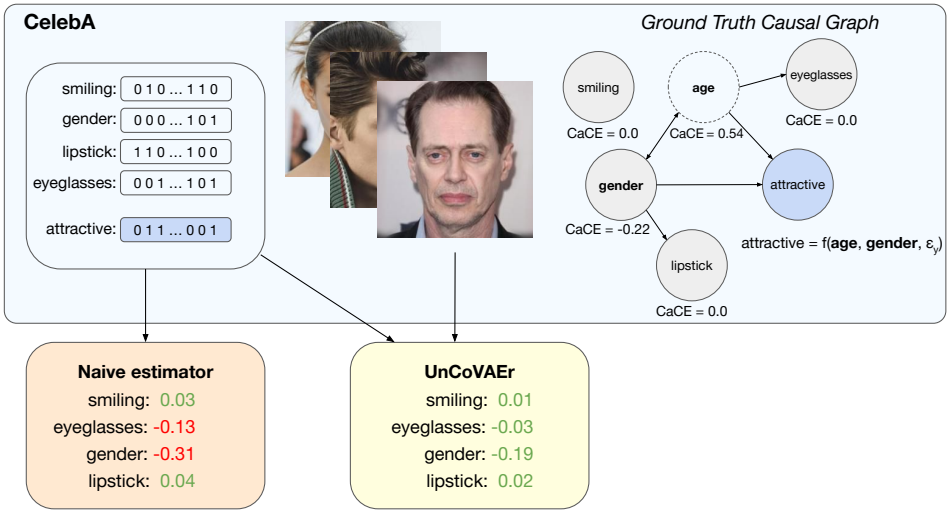


Figure 1: Latent confounders bias causal concept effect (CaCE) estimates. Here, *attractiveness* depends on *age* and *gender*, but *age* is unobserved (dashed node). A naive estimator yields biased CaCE due to open backdoor paths. UnCoVAEr exploits images to learn an *identifiable* latent representation Z that captures confounder-relevant information, substantially reducing bias.

We present UnCoVAEr, integrating TEDVAE's disentanglement of causal roles with identifiability guarantees. Our key insight: observed concepts C and outcomes Y serve as auxiliary variables that enable identifiable recovery of adjustment-sufficient latent factors up to component-wise transformations (Khemakhem et al., 2020; Mita et al., 2021). The factorized conditioning structure enforces disentanglement by acting as an information bottleneck, isolating confounder-relevant variation from other visual factors. We prove that component-wise identifiability suffices for valid backdoor adjustment, yielding a theoretically grounded consistent CaCE estimator under standard regularity conditions.

Contributions: (1) We formalize causal concept effect estimation under latent confounding and adapt latent-variable causal inference for images, by treating them as high-dimensional proxies. (2) We propose UnCoVAEr, learning identifiable confounder representations by conditioning latent priors on observed labels, and offering theoretical guarantees for consistent CaCE estimation. (3) We demonstrate consistent bias reduction across MorphoMNIST and CelebA benchmarks, outperforming feature-adjustment, counterfactual, and non-identifiable latent-variable baselines.

2 Problem Setup

Setting and notation. We observe i.i.d. samples $(X, C, Y) \sim \mathcal{D}$, consisting of an image $X \in \mathcal{X}$, a vector of M binary concept annotations $C = (C_1, \ldots, C_M) \in \{0,1\}^M$, and a binary outcome $Y \in \{0,1\}$. We assume a structural causal model (SCM) with latent factors $Z = (Z_{\text{conf}}, Z_C, Z_Y)$, where Z_{conf} are *confounders* affecting both C and Y, Z_C are concept-specific factors affecting C but not Y directly, and Z_Y are outcome-specific factors affecting Y but not C. All factors influence image appearance. The causal structure is shown in Figure 2. Concepts may exhibit causal relationships captured by a DAG $\mathcal{G}_C$, where $C_j \in \text{pa}(C_i)$ denotes that C_j is a parent of C_i.

Causal concept effect. In line with Goyal et al. (2020), we define the causal effect of C_i on Y as:

$$\text{CaCE}_i = \mathbb{E}\big[Y \mid \text{do}(C_i = 1)\big] - \mathbb{E}\big[Y \mid \text{do}(C_i = 0)\big], \tag{1}$$

where $\text{do}(C_i = c)$ denotes Pearl's do intervention (Pearl, 2009). The *individual* effect conditioned on image x is:

$$\text{ICaCE}_i(x) = \mathbb{E}\big[Y \mid \text{do}(C_i = 1), X = x\big] - \mathbb{E}\big[Y \mid \text{do}(C_i = 0), X = x\big], \tag{2}$$

with the population-level CaCE satisfying $\text{CaCE}_i = \mathbb{E}_X[\text{ICaCE}_i(X)]$.

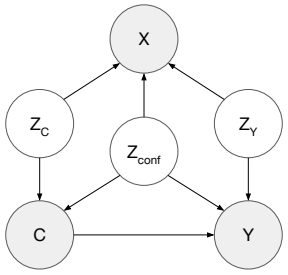


Figure 2: Assumed causal graph. Observed (shaded): image X, concepts C, outcome Y. We assume that all latent factors (white) cause X, while only Z_{conf} confounds the $C \to Y$ relationship.

Confounding bias under unconfoundedness. Causal effects estimators such as regression adjustment (Rosenbaum & Rubin, 1983) yield unbiased estimates under the assumption that all common causes of concepts and outcomes are observed. If C contains all common causes of C_i and Y, an estimator fits $\hat{\mu}(C) = \mathbb{E}[Y \mid C]$ and computes $\widehat{\text{CaCE}}_i = \hat{\mu}(C_i = 1, C_{-i}) - \hat{\mu}(C_i = 0, C_{-i})$. However, when there are latent confounders Z_{conf}, the backdoor path $C_i \leftarrow Z_{\text{conf}} \to Y$ remains open and the observational distribution $P(Y \mid C)$ differs from the interventional $P(Y \mid \text{do}(C))$, yielding systematically biased estimates (for a more detailed analysis see Appendix B).

Leveraging images for identification. We hypothesize that latent confounders manifest visually in X, rendering it a viable proxy to identify and adjust for the confounding structure. If Z_{conf} was observed, we would be able to identify the interventional distribution under $do(C = c)$ with backdoor adjustment: $p(Y \mid \text{do}(C = c)) = \int p(Y \mid C = c, Z_{\text{conf}}) p(Z_{\text{conf}})\, dZ_{\text{conf}}$. Since we cannot estimate its marginal distribution, we follow the framework of Zhang et al. (2020) and estimate the conditional distribution of $Z_{\text{adj}} := (Z_{\text{conf}}, Z_Y)$ given X[1]. The conditional interventional probability becomes:

$$p(Y \mid \text{do}(C = c), X = x) = \int p(Y \mid C = c, Z_{\text{adj}})\, p(Z_{\text{adj}} \mid X = x)\, dZ_{\text{adj}}. \tag{3}$$

To compute CaCE we can plug this into Equation 2 and get the expectation over the image distribution. Our challenge now becomes learning representations $\hat{Z}$ capturing Z_{adj} from (X, C, Y). We show that under standard assumptions, Z_{adj} is identifiable up to component-wise transformations, enabling consistent CaCE estimation.

3 Methodology

3.1 Model architecture

Identifiability via auxiliary conditioning. Khemakhem et al. (2020) show that conditioning latent priors on auxiliary variables u via the exponential family form $p(Z \mid u) = \prod_{i=1}^{d} h_i(z_i) g_i(u) \exp\left[\mathbf{T}_i(z_i)^\top \boldsymbol{\eta}_i(u)\right]$ achieves identifiability up to component-wise transformations, provided parameters $\boldsymbol{\eta}_i(u)$ vary sufficiently with u. We adopt the regularization framework of Mita et al. (2021) and use (C, Y) as auxiliary variables u. Since (C, Y) are causal descendants of latent confounders (Fig. 2), we treat $p_\psi(Z \mid C, Y)$ not as a causal mechanism but as an identification mechanism via auxiliary supervision (Mita et al., 2021). This conditioning provides the statistical variation needed to break latent symmetries and recover adjustment-sufficient representations.

Generative model. We adopt the latent factorization $Z = (Z_{\text{conf}}, Z_C, Z_Y)$ from Section 2. The generative model factorizes as:

$$p_\theta(X, C, Y \mid Z) = p_\theta(X \mid Z)\, p_\theta(C \mid \text{pa}_{\mathcal{G}_C}(C), Z_{\text{conf}}, Z_C)\, p_\theta(Y \mid C, Z_{\text{conf}}, Z_Y), \tag{4}$$

where the image decoder depends on all latent factors, concepts depend on (Z_{conf}, Z_C) and any parents in concept DAG $\mathcal{G}_C$, and outcomes depend on C and (Z_{conf}, Z_Y).

[1]We also add Z_Y to the adjustment set because (i) X is a collider and conditioning on it opens the backdoor path $C \leftarrow Z_{\text{conf}} \to X \leftarrow Z_Y \to Y$, and (ii) Z_Y is a precision variable and including it reduces estimator variance (Rotnitzky & Smucler, 2020).

We parameterize the conditional prior with a structured factorization:

$$p_\psi(Z \mid C, Y) = p_\psi(Z_{\text{conf}} \mid C, Y)\, p_\psi(Z_C \mid C)\, p_\psi(Z_Y \mid Y). \tag{5}$$

Each component is a diagonal Gaussian with neural network parameters satisfying exponential family requirements (Khemakhem et al., 2020). Following Mita et al. (2021), we treat (C, Y) as auxiliary variables whose role is statistical identification, not as conditioning variables in a generative sense. Accordingly, this factorization is a parameterization of the learned prior—it defines how the natural parameters of each latent component vary with auxiliary context—rather than a claim about the true conditional $p(Z \mid C, Y)$. This structured parameterization provides the variation across auxiliary contexts required for identifiability (Section 4), while acting as an information bottleneck that isolates task-relevant variation in each latent component.

3.2 Training objective

Following IDVAE (Mita et al., 2021), we optimize two coupled ELBOs using an amortized inference network $q_\phi(Z \mid X, C, Y)$. The main ELBO trains the encoder and the generative model:

$$\mathcal{L}_1 = \mathbb{E}_{q_\phi(Z|X,C,Y)} \log p_\theta(X, C, Y \mid Z) - \beta\, D_{\text{KL}}\big(q_\phi(Z \mid X, C, Y) \,\|\, p_\psi(Z \mid C, Y)\big). \tag{6}$$

The KL term serves a dual purpose: (1) providing auxiliary supervision for identifiability, and (2) regularizing the latent space via an *information bottleneck* (Tishby et al., 2000). Since $p_\psi(Z_C \mid C)$ excludes Y and $p_\psi(Z_Y \mid Y)$ excludes C, minimizing the KL penalizes dependence of Z_C on Y and of Z_Y on C under the variational distribution, encouraging Z_C to capture concept-relevant variation and Z_Y to capture outcome-relevant variation.

The prior ELBO trains p_ψ by sampling $Z \sim p_\psi(Z \mid C, Y)$ and reconstructing (C, Y):

$$\begin{aligned}\mathcal{L}_2 = \mathbb{E}_{p_\psi(Z|C,Y)}\big[\log p_\theta(C \mid \text{pa}_{\mathcal{G}_C}(C), Z_{\text{conf}}, Z_C) + \log p_\theta(Y \mid C, Z_{\text{conf}}, Z_Y)\big] \\ - D_{\text{KL}}\big(p_\psi(Z \mid C, Y) \,\|\, \mathcal{N}(0, I)\big),\end{aligned} \tag{7}$$

where the KL term regularizes the prior toward a standard normal (Mita et al., 2021).

At test time, only X is observed. We train auxiliary predictors $q_\phi(C \mid X)$ and $q_\phi(Y \mid X, C)$ jointly with the main model to predict unobserved C and Y during inference. The full objective is:

$$\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2 + \lambda_{\text{aux}} \mathbb{E}_{X,C,Y}[\log q_\phi(C \mid X) + \log q_\phi(Y \mid X, C)]. \tag{8}$$

All components are neural networks, parameterized by θ (generative), ϕ (inference), and ψ (conditional prior), trained end-to-end with Bernoulli likelihoods for (C, Y) and Gaussian for X.

3.3 Causal effect estimation

To estimate CaCE_i via backdoor adjustment (Eq. 3), we require $p_\theta(Y \mid C, Z_{\text{adj}})$ and the marginal posterior $q_\phi(Z_{\text{adj}} \mid X)$. The model provides the former by construction. For the latter, we marginalize over (C, Y), following Louizos et al. (2017):

$$q_\phi(Z \mid X) = \sum_{c,y} q_\phi(c \mid X)\, q_\phi(y \mid X, c)\, q_\phi(Z \mid X, c, y). \tag{9}$$

We implement this via Monte Carlo: sample $\tilde{c} \sim q_\phi(C \mid X)$, then $\tilde{y} \sim q_\phi(Y \mid X, \tilde{c})$, then $z \sim q_\phi(Z \mid X, \tilde{c}, \tilde{y})$.

For each image, we (i) sample latents from the marginal posterior, (ii) intervene on target concept C_i by setting it to 0 and 1, propagating effects through concept DAG $\mathcal{G}_C$ if needed, (iii) evaluate outcome probabilities $\mathbb{E}_{p_\theta(Y|C,z)}[Y]$ under each intervention, and (iv) compute the individual causal effect as the difference. The population estimate averages over all images: $\widehat{\text{CaCE}}_i = \frac{1}{N}\sum_{n=1}^{N} \widehat{\text{ICaCE}}_i(x_n)$. The detailed estimation procedure is depicted in Algorithm 1 (Appendix C).

4 IDENTIFIABILITY ANALYSIS

We now establish that, under certain assumptions, $Z_{\text{adj}} = (Z_{\text{conf}}, Z_Y)$ is identifiable up to component-wise invertible transformations (Khemakhem et al., 2020) and show that this suffices for consistent CaCE estimation.

Assumption 4.1 (Causal factorization)**.** The data-generating process admits the factorization:

$$p(X, C, Y, Z) = p(X \mid Z)\, p(C \mid \text{pa}_{\mathcal{G}_C}(C), Z_{\text{conf}}, Z_C)\, p(Y \mid C, Z_{\text{conf}}, Z_Y)\, p(Z_{\text{conf}})\, p(Z_C)\, p(Z_Y),$$

with mutually independent latent factors and known concept DAG $\mathcal{G}_C$.

Assumption 4.2 (Identifiable auxiliary supervision)**.** Each conditional prior $p_\psi(Z_{\text{conf}} \mid C, Y)$, $p_\psi(Z_C \mid C)$, $p_\psi(Z_Y \mid Y)$ belongs to an exponential family with twice differentiable sufficient statistics $T_i(z_i)$ and natural parameters $\eta_i(u)$ that vary sufficiently across conditioning contexts u (Khemakhem et al., 2020). The decoder $p_\theta(X \mid Z)$ is injective almost everywhere and has all second-order cross derivatives.

Assumption 4.3 (Overlap)**.** For all x in the support of X, the posterior $q_\phi(Z_{\text{adj}} \mid x)$ has positive density over values needed for counterfactual evaluation.

Lemma 4.4 (Backdoor adjustment under component-wise transformations)**.** *Under Assumption 4.1, let $\hat{Z}_{\text{adj}} = \tau(Z_{\text{adj}})$ for a component-wise invertible function τ. Then the backdoor formula holds with transformed variables:*

$$\mathbb{E}[Y \mid \text{do}(C = c), X = x] = \int \mathbb{E}[Y \mid C{=}c, \hat{Z}_{\text{adj}}]\, p(\hat{Z}_{\text{adj}} \mid X{=}x)\, d\hat{Z}_{\text{adj}}. \tag{10}$$

The key insight is that $\mathbb{E}[Y \mid C, Z_{\text{adj}}]$ is a function of Z_{adj}, and any invertible transformation of Z_{adj} preserves this conditional expectation (proof in Appendix E).

Theorem 4.5 (Identifiability of adjustment set)**.** *Under Assumptions 4.1–4.2, if two models yield the same observed distribution $p_{\theta^*}(X, C, Y) = p_\theta(X, C, Y)$, their adjustment latents satisfy $\hat{z}_{\text{adj},i} = f_{\pi(i)}(z^*_{\text{adj},i})$ for some permutation π and invertible scalar functions $f_i : \mathbb{R} \to \mathbb{R}$.*

Proof sketch. The factorized prior structure provides distinct conditioning contexts for each latent component. By Theorem 2 of Khemakhem et al. (2020) (applied as in Assumption 4.2), sufficient variation of natural parameters $\eta_i(u)$ across contexts recovers each component up to permutation and scalar invertible transformation. We treat (C, Y) as auxiliary variables for identification (not as causal mechanisms), following IDVAE (Mita et al., 2021). Full proof in Appendix E. □

Corollary 4.6 (Consistency)**.** *Under Assumptions 4.1–4.3, as $N, S \to \infty$ and model capacity grows, $\widehat{\text{CaCE}}_i \xrightarrow{p} \text{CaCE}_i$.*

Proof sketch. By Theorem 4.5 and Lemma 4.4, backdoor adjustment is valid with learned $\hat{Z}_{\text{adj}}$. The estimation error decomposes into model approximation error (vanishes with capacity and data), finite sample error (vanishes as $N \to \infty$), and Monte Carlo error (vanishes as $S \to \infty$). Full proof in Appendix E. □

5 EXPERIMENTS

We evaluate our method on two datasets with known ground-truth causal concept effects. We build a semi-synthetic benchmark on **MorphoMNIST** (Castro et al., 2019), where we control both the causal effect of concepts C on a synthetic outcome Y, as well as the causal relationships between concepts. We define four binary concepts by thresholding morphological measurements: *thickness* (t), *intensity* (i), *slant* (s) and *width* (w) and generate $Y \sim \text{Bernoulli}(\beta^\top C)$. We systematically evaluate four latent confounding structures: **(i) single confounder** (t unobserved, $t \to i$), **(ii) multiple confounders** (s and t unobserved, $s \to i \leftarrow t$), **(iii) common confounder** (t unobserved, $s \leftarrow t \to i$), and **(iv) causally related observed concepts** ($i \to s$, t (unobserved)$\to s$) (full details on the data generating process in Appendix F). Across variants, we set C_i to the confounder value with probability α, controlling the confounding strength. We evaluate on an in-distribution

Table 1: Mean CaCE estimation error (lower is better) on in-distribution test sets, averaged across concepts. Mean ± std over 5 seeds. Best non-oracle per column in **bold**.

	MorphoMNIST				CelebA	
Method	Single	Multiple	Common	Causal	Unknown age	Unknown gender
Naive/CBM	.066 ± .08	.145 ± .14	.069 ± .05	.143 ± .12	.072 ± .04	.048 ± .07
Oracle	.016 ± .02	.020 ± .01	.018 ± .02	.012 ± .01	.002 ± .00	.004 ± .00
Image adjustment	.076 ± .08	.164 ± .14	.293 ± .19	.120 ± .09	.180 ± .11	.160 ± .13
Res-CBM	.074 ± .04	.145 ± .06	.193 ± .09	.099 ± .06	.111 ± .08	.103 ± .11
CaCE-VAE	.080 ± .08	.084 ± .07	.067 ± .07	.102 ± .06	.062 ± .06	.179 ± .16
CEVAE	.102 ± .05	.121 ± .05	.121 ± .06	.120 ± .10	.058 ± .06	.047 ± .05
TEDVAE	.060 ± .06	.096 ± .09	.068 ± .06	.158 ± .11	.063 ± .05	.043 ± .05
UnCoVAEr$_{\mathbf{Z}}$	**.047 ± .03**	.071 ± .03	.075 ± .04	.100 ± .07	.056 ± .06	.041 ± .05
UnCoVAEr	.048 ± .03	**.030 ± .02**	**.059 ± .04**	**.077 ± .07**	**.052 ± .03**	**.037 ± .04**

test set with strong confounding ($\alpha = 0.9$) (Table 1) and an out-of-distribution test set with weaker confounding ($\alpha = 0.6$) (Table 2 in Appendix H).

We use **CelebA** (Liu et al., 2015) at 64×64 resolution with five facial attributes as binary concepts: *Smiling*, *Eyeglasses*, *Lipstick*, *Age* and *Gender*. To enable quantitative evaluation, we replace the original *Attractiveness* label with a synthetic outcome $Y \sim \text{Bernoulli}(\beta_0 + \beta_{\text{age}}\text{Age} + \beta_{\text{gender}}\text{Gender} + \beta_{\text{int}}\text{Age} \cdot \text{Gender})$. We evaluate two complementary settings: **(i) Unknown age** and **(ii) Unknown gender**. *Age* and *gender* are correlated in CelebA due to dataset collection bias (Torfason et al., 2017), and both influence other attributes (e.g., *Age→Eyeglasses*, *Gender→Lipstick*), making this a challenging testbed for latent confounding.

Baselines. We compare against: (1) **Naive/CBM** (Koh et al., 2020): predicts Y from observed concepts C via logistic regression, ignoring latent confounding, (2) **Oracle**: has access to all concepts, including latent confounders (upper bound), (3) **Image adjustment** (Jerzak et al., 2023): estimates concept effects via inverse probability weighting (IPW) computing propensity scores from images, (4) **Res-CBM**: trains a CBM with residuals (Yuksekgonul et al., 2023) and uses the residuals for IPW adjustment, (5) **CaCE-VAE** (Goyal et al., 2020): trains a conditional VAE to generate images x' under concept interventions and compares classifier predictions $q_\phi(Y|x')$, assuming unconfoundedness, (6-7) **CEVAE** (Louizos et al., 2017) and **TEDVAE** (Zhang et al., 2020): latent-variable adjustment without conditional priors, and (8) **UnCoVAEr$_{\mathbf{Z}}$**: our method with merged latents (analogous to CEVAE with conditional prior). More details on the baselines in Appendix G.

Results. We report $\text{CaCE}_{\text{error}} = \frac{1}{M}\sum_{i=1}^{M} |\widehat{\text{CaCE}}_i - \text{CaCE}_{\text{true},i}|$ in Table 1. UnCoVAEr outperforms CEVAE and TEDVAE, which learn latent representations for adjustment in the same manner, but lack identifiability guarantees. This validates our core contribution: leveraging (C, Y) as auxiliary variables recovers representations aligned with true confounders up to component-wise transformations, sufficient for consistent estimation. UnCoVAEr's gains over UnCoVAEr$_{\mathbf{Z}}$ increase with confounding complexity, demonstrating that explicitly separating latents provides benefits beyond generic latent-variable modeling. This confirms that incorporating causal structure into latent architecture is crucial for reliable inference. Image adjustment and Res-CBM often perform worse than concept-only estimation, as conditioning on high-dimensional images can lead to weak overlap, resulting in unstable estimation, especially when confounders influence multiple concepts. On CelebA, UnCoVAEr achieves the lowest error in both Unknown Age and Unknown Gender settings, validating that identifiable representations generalize beyond controlled synthetic data. Unknown Age proves more challenging across methods, since age manifests through subtler visual patterns, while gender is highly correlated with known concepts, like lipstick. CaCE-VAE shows high variance when unconfoundedness is violated, confirming the necessity of explicit confounder modeling.

6 Conclusion

We presented UnCoVAEr, a framework for identifiable causal concept effect estimation under visual latent confounding. By leveraging observed concepts and outcomes as auxiliary variables within a factorized exponential family prior, our model recovers identifiable confounder representations

sufficient for consistent backdoor adjustment. Empirically, UnCoVAEr outperforms concept-based, counterfactual and non-identifiable baselines across diverse confounding structures.

Limitations. Confounders that do not manifest visually remain unidentifiable, a fundamental constraint of image-based methods. While our synthetic outcomes enable quantitative evaluation, validation on purely experimental interventions would strengthen claims. Our identifiability guarantees are asymptotic, with finite-sample performance depending on model capacity and practical validity of exponential family assumptions. Future work should also explore settings with continuous concepts and outcomes, and further investigate the minimum number of labels needed for reliable identification of confounders.

ACKNOWLEDGMENTS

This work has been partially supported by project MIS 5154714 of the National Recovery and Resilience Plan Greece 2.0 funded by the European Union under the NextGenerationEU Program. We acknowledge the support of the UKRI AI programme, and the Engineering and Physical Sciences Research Council, for CHAI - Causality in Healthcare AI Hub [grant number EP/Y028856/1].

REFERENCES

Marcus A. Badgeley, John R. Zech, Luke Oakden-Rayner, Benjamin Scott Glicksberg, Manway Liu, William Gale, Michael V. McConnell, Bethany Percha, Thomas M. Snyder, and Joel T. Dudley. Deep learning predicts hip fracture using confounding patient and healthcare variables. *NPJ Digital Medicine*, 2, 2018. URL https://api.semanticscholar.org/CorpusID:53250171.

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.

Alex J DeGrave, Joseph D Janizek, and Su-In Lee. Ai for radiographic covid-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3(7):610–619, 2021.

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

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

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

Luigi Gresele, Paul K Rubenstein, Arash Mehrjou, Francesco Locatello, and Bernhard Schölkopf. The incomplete rosetta stone problem: Identifiability results for multi-view nonlinear ica. In *Uncertainty in Artificial Intelligence*, pp. 217–227. PMLR, 2020.

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.

Charles Jones, Daniel C Castro, Fabio De Sousa Ribeiro, Ozan Oktay, Melissa McCradden, and Ben Glocker. A causal perspective on dataset bias in machine learning for medical imaging. *Nature Machine Intelligence*, 6(2):138–146, 2024.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International conference on artificial intelligence and statistics*, pp. 2207–2217. PMLR, 2020.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

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.

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.

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.

Sébastien Lachapelle, Pau Rodriguez, Yash Sharma, Katie E Everett, Rémi Le Priol, Alexandre Lacoste, and Simon Lacoste-Julien. Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ica. In *Conference on Causal Learning and Reasoning*, pp. 428–484. PMLR, 2022.

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.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pp. 4114–4124. PMLR, 2019.

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.

Graziano Mita, Maurizio Filippone, and Pietro Michiardi. An identifiable double vae for disentangled representations. In *International Conference on Machine Learning*, pp. 7769–7779. PMLR, 2021.

Gemma Elyse Moran, Dhanya Sridhar, Yixin Wang, and David Blei. Identifiable deep generative models via sparse decoding. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=vd0onGWZbE.

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. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2009.

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.

Paul Rosenbaum and Donald Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70:41–55, 04 1983. doi: 10.1093/biomet/70.1.41.

Andrea Rotnitzky and Ezequiel Smucler. Efficient adjustment sets for population average causal treatment effect estimation in graphical models. *Journal of Machine Learning Research*, 21(188): 1–86, 2020. URL http://jmlr.org/papers/v21/19-1026.html.

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.

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.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

Robert Torfason, Eirikur Agustsson, Rasmus Rothe, and Radu Timofte. From face images and attributes to attributes. In Shang-Hong Lai, Vincent Lepetit, Ko Nishino, and Yoichi Sato (eds.), *Computer Vision – ACCV 2016*, pp. 313–329, Cham, 2017. Springer International Publishing. ISBN 978-3-319-54187-7.

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.

Ziqi Xu, Debo Cheng, Jiuyong Li, Jixue Liu, Lin Liu, and Kui Yu. Causal inference with conditional front-door adjustment and identifiable variational autoencoder. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=wFf9m4v7oC.

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.

John R Zech, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS Medicine*, 15(11):e1002683, 2018.

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.

Yaochen Zhu, Jing Ma, Liang Wu, Qi Guo, Liangjie Hong, and Jundong Li. Causal effect estimation with mixed latent confounders and post-treatment variables. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=qe1CsfnN1W.

A Extended related work

Concept-based models and causal concept effects. Concept Bottleneck Models (CBMs) (Koh et al., 2020) achieve interpretability by predicting human-interpretable concepts as intermediate representations. Extensions address incomplete concept sets (Yuksekgonul et al., 2023; Oikarinen et al., 2023; Sawada & Nakamura, 2022), though questions remain about whether learned concepts capture true semantic content (Mahinpei et al., 2021; Margeloiu et al., 2021). Goyal et al. (2020) quantify the effect of concepts on a classifier by generating counterfactuals, but assume that all confounders are observed, which rarely holds in real-world settings.

Latent-variable methods for causal inference. CEVAE (Louizos et al., 2017) trains a modified VAE (Kingma & Welling, 2013) to learn latent confounders and then performs backdoor adjustment, while TEDVAE (Zhang et al., 2020) disentangles the latent representations for treatment-inducing, outcome-inducing, and confounding factors. However, the standard VAEs that they utilize are non-identifiable: many latent representations can explain the same observational distributions (Locatello et al., 2019). Both CEVAE and TEDVAE lack identifiability guarantees, leading to potentially inconsistent estimates under model misspecification (Rissanen & Marttinen, 2021). Khemakhem et al. (2020) address identifiability by conditioning the prior on auxiliary variables, recovering latents up to permutation and component-wise transformations. Mita et al. (2021) extend this by learning a regularized optimal representation for the conditional prior, while other works leverage sparsity (Lachapelle et al., 2022; Moran et al., 2022), or multi-view observations (Gresele et al., 2020). Another line of work tackles identifiability without auxiliary variables by using structured priors such as Gaussian mixtures (Kivva et al., 2022).

Recent work combines identifiable VAEs with causal estimation. CFDiVAE (Xu et al., 2024) learns conditional front-door adjustment variables from proxies, extending the classical front-door criterion. CiVAE (Zhu et al., 2025) assumes only mixed confounders and post-treatment variables, using identifiable VAE to avoid post-treatment bias. Wu & Fukumizu (2022) model prognostic scores for treatment effect estimation under limited overlap. These methods vary in their structural assumptions and how they leverage auxiliary information for causal adjustment.

Proximal and image-based causal inference. Proximal causal inference (Tchetgen et al., 2020) provides formal identification under latent confounding when treatment- and outcome-inducing proxies are available and satisfy certain *completeness* conditions (Miao et al., 2018; Wang & Blei, 2021). Deep learning extensions can handle high-dimensional proxies (Xu et al., 2021; Kompa et al., 2022), but also require *two distinct proxy types*. On the other hand, Jerzak et al. (2023); Kumar et al. (2023); Schulte et al. (2025) use image features directly for adjustment, treating images as observed covariates, rather than proxies of latent confounders. Our approach differs by treating the image as a single rich proxy and leveraging auxiliary supervision to achieve identifiability without requiring two proxies.

B Formal bias characterization

We provide a formal characterization of the bias in naive concept-based estimation. Consider the naive estimator:

$$\widehat{\mathrm{CaCE}}_i^{\mathrm{naive}} = \mathbb{E}[Y \mid C_i = 1, C_{-i}] - \mathbb{E}[Y \mid C_i = 0, C_{-i}]. \tag{11}$$

The bias is:

$$\begin{aligned}
\mathrm{Bias} &= \widehat{\mathrm{CaCE}}_i^{\mathrm{naive}} - \mathrm{CaCE}_i && (12)\\
&= \mathbb{E}[Y \mid C_i = 1, C_{-i}] - \mathbb{E}[Y \mid C_i = 0, C_{-i}] && (13)\\
&\quad - \big(\mathbb{E}[Y \mid \mathrm{do}(C_i = 1), C_{-i}] - \mathbb{E}[Y \mid \mathrm{do}(C_i = 0), C_{-i}]\big). && (14)
\end{aligned}$$

Under our causal model, we can expand the observational expectation:

$$\mathbb{E}[Y \mid C_i = c, C_{-i}] = \int \mathbb{E}[Y \mid C_i = c, C_{-i}, Z_{\mathrm{conf}}] p(Z_{\mathrm{conf}} \mid C_i = c, C_{-i}) dZ_{\mathrm{conf}}, \tag{15}$$

while the interventional expectation is:

$$\mathbb{E}[Y \mid \mathrm{do}(C_i = c), C_{-i}] = \int \mathbb{E}[Y \mid C_i = c, C_{-i}, Z_{\mathrm{conf}}] p(Z_{\mathrm{conf}} \mid C_{-i}) dZ_{\mathrm{conf}}. \tag{16}$$

The difference arises because $p(Z_{\text{conf}} \mid C_i = c, C_{-i}) \neq p(Z_{\text{conf}} \mid C_{-i})$ whenever Z_{conf} causally influences C_i. This dependency means that observing $C_i = c$ provides information about the confounder distribution, creating spurious associations between C_i and Y through the backdoor path $C_i \leftarrow Z_{\text{conf}} \rightarrow Y$.

The magnitude of bias depends on:

1. The strength of confounding: $Z_{\text{conf}} \rightarrow C_i$ and $Z_{\text{conf}} \rightarrow Y$
2. The degree of conditional dependence between Z_{conf} and C_i given C_{-i}
3. The variance of Z_{conf} in the population

This bias is non-zero and systematic whenever latent confounders exist, justifying our approach of learning and adjusting for these hidden factors.

C CAUSAL CONCEPT EFFECT ESTIMATION ALGORITHM

Algorithm 1 Causal concept effect estimation with UnCoVAEr

Require: Images $\{x_n\}_{n=1}^N$, trained model $(q_\phi, p_\theta, p_\psi)$, target concept i, MC samples S
Ensure: $\widehat{\text{CaCE}}_i$

1: $\tau \leftarrow 0$
2: **for** $n = 1, \ldots, N$ **do**
3: **for** $s = 1, \ldots, S$ **do**
4: Sample $\tilde{c} \sim q_\phi(C \mid x_n)$, $\tilde{y} \sim q_\phi(Y \mid x_n, \tilde{c})$
5: Sample $z \sim q_\phi(Z \mid x_n, \tilde{c}, \tilde{y})$
6: **for** $c \in \{0, 1\}$ **do**
7: Set $C_i \leftarrow c$; $\mathcal{D} \leftarrow \{j : C_i \in \text{ancestors}(C_j)\}$
8: **for** $j \in \mathcal{D}$ in topological order of $\mathcal{G}_C$ **do**
9: Sample $C_j \sim p_\theta(C_j \mid \text{pa}(C_j), z)$
10: **end for**
11: $\hat{y}_c \leftarrow \mathbb{E}_{p_\theta(Y \mid C, z)}[Y]$
12: **end for**
13: $\tau \leftarrow \tau + (\hat{y}_1 - \hat{y}_0)$
14: **end for**
15: **end for**
16:
17: **return** $\widehat{\text{CaCE}}_i \leftarrow \tau / (N \cdot S)$

The algorithm implements the backdoor adjustment formula from Equation 3. For each image, we:

Step 1 (Lines 4-5): Obtain the marginal posterior $q_\phi(Z \mid X)$ by first predicting unobserved labels via auxiliary networks, then sampling latents conditioned on these predictions.

Step 2 (Lines 6-10): For each intervention value $c \in \{0, 1\}$, we set $C_i = c$ and propagate effects through the concept DAG. If concepts have causal dependencies (captured by $\mathcal{G}_C$), we sample descendants of C_i in topological order from their conditional distributions. This respects the causal structure: intervening on C_i affects its descendants but not its parents or non-descendants.

Step 3 (Line 11): Evaluate the expected outcome under intervention $\text{do}(C_i = c)$ by computing $\mathbb{E}_{p_\theta(Y \mid C, z)}[Y]$. The latent z controls for confounding while the concept values reflect the intervention.

Step 4 (Line 13): Compute the individual causal effect $\widehat{\text{ICaCE}}_i(x_n)$ as the difference in expected outcomes.

The outer loops average over multiple Monte Carlo samples per image (S) and over all images (N) to obtain the population-level estimate $\widehat{\text{CaCE}}_i$.

D CODE AND IMPLEMENTATION DETAILS

All code, datasets, model architectures and configuration files (including hyperparameters) required to replicate our results are provided in `https://github.com/gulnazaki/uncovaer`.

E IDENTIFIABILITY ANALYSIS: FULL PROOFS

E.1 CONDITIONAL INDEPENDENCE FROM CAUSAL FACTORIZATION

We first establish that Assumption 4.1 implies the required conditional independence.

Proposition E.1. *Under Assumption 4.1,* $Y \perp\!\!\!\perp Z_C \mid C, Z_{\text{adj}}$.

Proof. By Assumption 4.1, Y depends on Z only through $(C, Z_{\text{conf}}, Z_Y)$, i.e. $p(Y \mid C, Z) = p(Y \mid C, Z_{\text{conf}}, Z_Y)$. Combined with the mutual independence of latent factors $(Z_{\text{conf}}, Z_C, Z_Y)$, we have:

$$\begin{aligned} p(Y, Z_C \mid C, Z_{\text{adj}}) &= p(Y \mid C, Z_{\text{adj}}, Z_C) \cdot p(Z_C \mid C, Z_{\text{adj}}) \\ &= p(Y \mid C, Z_{\text{adj}}) \cdot p(Z_C \mid C), \end{aligned}$$

where the second equality uses that Y does not depend on Z_C given (C, Z_{adj}), and $Z_C \perp\!\!\!\perp Z_{\text{adj}}$ by mutual independence.

This factorization establishes $Y \perp\!\!\!\perp Z_C \mid C, Z_{\text{adj}}$. □

E.2 PROOF OF LEMMA 4.4

Lemma E.2 (Restatement)**.** *Under Assumption 4.1, let* $\hat{Z}_{\text{adj}} = \tau(Z_{\text{adj}})$ *for a component-wise invertible function* τ*. Then the backdoor formula holds with transformed variables:*

$$\mathbb{E}[Y \mid \text{do}(C = c), X = x] = \int \mathbb{E}[Y \mid C{=}c, \hat{Z}_{\text{adj}}]\, p(\hat{Z}_{\text{adj}} \mid X{=}x)\, d\hat{Z}_{\text{adj}}.$$

Proof. We proceed in two steps.

Step 1: Conditional expectation is preserved under invertible transformations.

Define $g(z_{\text{adj}}) := \mathbb{E}[Y \mid C = c, Z_{\text{adj}} = z_{\text{adj}}]$. Since τ is invertible, conditioning on $\hat{Z}_{\text{adj}} = \hat{z}_{\text{adj}}$ is equivalent to conditioning on $Z_{\text{adj}} = \tau^{-1}(\hat{z}_{\text{adj}})$:

$$\mathbb{E}[Y \mid C = c, \hat{Z}_{\text{adj}} = \hat{z}_{\text{adj}}] = \mathbb{E}[Y \mid C = c, Z_{\text{adj}} = \tau^{-1}(\hat{z}_{\text{adj}})] = g(\tau^{-1}(\hat{z}_{\text{adj}})).$$

Step 2: The backdoor integral is invariant under change of variables.

For a component-wise transformation $\hat{z}_{\text{adj}} = \tau(z_{\text{adj}})$, the Jacobian is diagonal:

$$J_\tau(z_{\text{adj}}) = \text{diag}\big(\tau_1'(z_{\text{adj},1}), \ldots, \tau_d'(z_{\text{adj},d})\big),$$

with determinant $|J_\tau| = \prod_{i=1}^{d} |\tau_i'(z_{\text{adj},i})|$.

By the standard backdoor adjustment formula:

$$\mathbb{E}[Y \mid \text{do}(C = c), X = x] = \int g(z_{\text{adj}})\, p(z_{\text{adj}} \mid X = x)\, dz_{\text{adj}}.$$

Applying the change of variables $\hat{z}_{\text{adj}} = \tau(z_{\text{adj}})$:

$$\int g(z_{\text{adj}})\, p(z_{\text{adj}} \mid X{=}x)\, dz_{\text{adj}} = \int g(\tau^{-1}(\hat{z}_{\text{adj}}))\, p(\tau^{-1}(\hat{z}_{\text{adj}}) \mid X{=}x)\, |J_{\tau^{-1}}|\, d\hat{z}_{\text{adj}}.$$

By the change-of-variables formula for densities, $p(\hat{z}_{\text{adj}} \mid X) = p(\tau^{-1}(\hat{z}_{\text{adj}}) \mid X) \cdot |J_{\tau^{-1}}|$, so:

$$\begin{aligned} &= \int g(\tau^{-1}(\hat{z}_{\text{adj}}))\, p(\hat{z}_{\text{adj}} \mid X{=}x)\, d\hat{z}_{\text{adj}} \\ &= \int \mathbb{E}[Y \mid C{=}c, \hat{Z}_{\text{adj}}{=}\hat{z}_{\text{adj}}]\, p(\hat{z}_{\text{adj}} \mid X{=}x)\, d\hat{z}_{\text{adj}}, \end{aligned}$$

where the Jacobians cancel. This shows the causal estimand can be equivalently computed using $\hat{Z}_{\text{adj}}$. □

E.3 PROOF OF THEOREM 4.5

Theorem E.3 (Restatement)**.** *Under Assumptions 4.1–4.2, if two models yield the same observed distribution $p_{\theta^*}(X, C, Y) = p_\theta(X, C, Y)$, their adjustment latents satisfy $\hat{z}_{\mathrm{adj},i} = f_{\pi(i)}(z^*_{\mathrm{adj},i})$ for some permutation π and invertible scalar functions $f_i : \mathbb{R} \to \mathbb{R}$.*

Proof. We adapt the identifiability framework of iVAE (Khemakhem et al., 2020) and IDVAE (Mita et al., 2021). The key insight is that although (C, Y) are causal descendants of Z in the generative model, we use them as auxiliary variables for *statistical identification* following IDVAE.

Step 1: Equivalence of conditional distributions.

Suppose two models with parameters (θ^*, ψ^*) and (θ, ψ) yield the same observed distribution $p_{\theta^*}(X, C, Y) = p_\theta(X, C, Y)$. Then all derived conditional distributions must also match:

$$p_{\theta^*}(X \mid C{=}c, Y{=}y) = p_\theta(X \mid C{=}c, Y{=}y) \quad \forall (c, y). \tag{17}$$

Both models generate X by marginalizing over latents:

$$p_{\theta^*}(X \mid c, y) = \int p_{\theta^*}(X \mid Z^*)\, p_{\psi^*}(Z^* \mid c, y)\, dZ^*, \tag{18}$$

$$p_\theta(X \mid c, y) = \int p_\theta(X \mid Z)\, p_\psi(Z \mid c, y)\, dZ. \tag{19}$$

Step 2: Factorized prior structure.

By construction, the conditional prior is parameterized with a structured factorization:

$$p_\psi(Z \mid c, y) = p_\psi(Z_{\mathrm{conf}} \mid c, y)\, p_\psi(Z_C \mid c)\, p_\psi(Z_Y \mid y). \tag{20}$$

This is a parameterization of the learned prior (Section 3), structured to reflect the causal roles of each component: Z_{conf} is conditioned on both C and Y, Z_C on C only, and Z_Y on Y only. Within this model class, the natural parameters of each component vary with a distinct subset of auxiliaries, enabling the component-wise identification argument below.

Step 3: Applying iVAE identifiability to Z_{conf}.

The iVAE argument is applied within the model class defined by the factorized parameterization of Eq. (5). Within this class, the natural parameters of each latent component vary with a distinct subset of auxiliaries, and the identification argument proceeds component-wise as follows.

The prior $p_\psi(Z_{\mathrm{conf}} \mid c, y)$ is conditioned on (C, Y), providing 2^{M+1} distinct contexts for M binary concepts and binary outcome. By Assumption 4.2, this prior belongs to an exponential family:

$$p_\psi(Z_{\mathrm{conf}} \mid c, y) = h(Z_{\mathrm{conf}}) \exp\big(\langle T(Z_{\mathrm{conf}}), \eta(c, y)\rangle - A(\eta(c, y))\big), \tag{21}$$

with natural parameters $\eta(c, y)$ that vary sufficiently across contexts. By Theorem 2 of Khemakhem et al. (2020), since (i) the decoder is injective almost everywhere and has all second-order cross derivatives, (ii) sufficient statistics $T_{i,1}(z) = z$ and $T_{i,2}(z) = z^2$ are twice differentiable, and (iii) the rank condition of Assumption 4.2 holds (feasible since 2^{M+1} contexts are available and $2^{M+1} \geq 2d_{\mathrm{conf}} + 1$ for reasonable d_{conf}), matching conditional distributions $p_{\theta^*}(X \mid c, y) = p_\theta(X \mid c, y)$ implies Z_{conf} is identifiable up to permutation and component-wise invertible scalar transformations:

$$\hat{z}_{\mathrm{conf},i} = f^{\mathrm{conf}}_{\pi(i)}(z^*_{\mathrm{conf},i}), \tag{22}$$

where π is a permutation and each f^{conf}_i is invertible.

Step 4: Identification of Z_Y.

The prior $p_\psi(Z_Y \mid y)$ conditions on binary Y, providing two contexts ($y = 0$ and $y = 1$). The iVAE argument applies with $u = y$, yielding identification of Z_Y up to component-wise invertible transformations, provided $\dim(Z_Y)$ is chosen such that the rank condition of Assumption 4.2 is satisfied—in practice we keep $\dim(Z_Y)$ small accordingly:

$$\hat{z}_{Y,i} = f^Y_{\rho(i)}(z^*_{Y,i}), \tag{23}$$

for some permutation ρ and invertible functions f_i^Y.

Step 5: Combining results for the adjustment set.

Since $Z_{\text{adj}} = (Z_{\text{conf}}, Z_Y)$ and both components are identifiable up to component-wise transformations, the full adjustment set is identifiable up to component-wise transformations (with a combined permutation over all components).

Note on Z_C: We do not require identifiability of Z_C for valid backdoor adjustment. By Proposition E.1, $Y \perp\!\!\!\perp Z_C \mid C, Z_{\text{adj}}$, so Z_C is irrelevant for predicting Y once we condition on C and Z_{adj}. The information bottleneck induced by $p_\psi(Z_C \mid C)$ (which excludes Y) combined with KL minimization encourages Z_C to capture only concept-specific variation, but this is for training stability rather than a theoretical necessity. □

E.4 Proof of Corollary 4.6

Corollary E.4 (Restatement). *Under Assumptions 4.1–4.3, as $N, S \to \infty$ and model capacity grows, $\widehat{\text{CaCE}}_i \xrightarrow{p} \text{CaCE}_i$.*

Proof. The proof proceeds in three stages: establishing validity of backdoor adjustment with learned latents, describing the algorithmic implementation, and analyzing asymptotic convergence.

Stage 1: Validity of backdoor adjustment with transformed latents.

Theorem 4.5 is a population-level identifiability result: it states that any two models achieving the same observed distribution must have adjustment latents related by component-wise invertible transformations. To connect this to convergence of the learned model, we invoke Theorem 4 of Khemakhem et al. (2020): if the variational family $q_\phi(Z \mid X, C, Y)$ contains the true posterior and the ELBO is maximized, then in the limit of infinite data the VAE converges to the true parameters up to the equivalence class of Theorem 4.5. Formally, as model capacity and data grow, the learned $\hat{Z}_{\text{adj}}$ satisfies:

$$\hat{z}_{\text{adj},i} = \tau_i(z^*_{\text{adj},i}), \quad i = 1, \ldots, \dim(Z_{\text{adj}}), \tag{24}$$

for invertible scalar functions $\{\tau_i\}$.

By Lemma 4.4, the backdoor formula holds exactly with transformed latents:

$$\mathbb{E}[Y \mid \text{do}(C = c), X = x] = \int \mathbb{E}[Y \mid C = c, \hat{Z}_{\text{adj}}]\, p(\hat{Z}_{\text{adj}} \mid X = x)\, d\hat{Z}_{\text{adj}}. \tag{25}$$

Stage 2: Algorithmic implementation.

The estimation algorithm (Algorithm 1) implements backdoor adjustment via Monte Carlo integration:

(a) *Latent inference:* For each image x_n, sample from the marginal posterior by first sampling $\tilde{c} \sim q_\phi(C \mid x_n)$, then $\tilde{y} \sim q_\phi(Y \mid x_n, \tilde{c})$, then $\hat{z} \sim q_\phi(Z \mid x_n, \tilde{c}, \tilde{y})$.
(b) *Interventional outcomes:* For each intervention $c \in \{0, 1\}$, set $C_i = c$, propagate through concept DAG $\mathcal{G}_C$, and compute $\mathbb{E}_{p_\theta(Y \mid C, \hat{z})}[Y]$.
(c) *Averaging:* Compute $\widehat{\text{CaCE}}_i = \frac{1}{NS} \sum_{n,s} (\hat{y}^1_{n,s} - \hat{y}^0_{n,s})$.

Stage 3: Asymptotic convergence analysis.

The estimation error decomposes as:

$$
\begin{aligned}
|\widehat{\mathrm{CaCE}}_i - \mathrm{CaCE}_i| \leq & \underbrace{\mathbb{E}_X \left| \mathbb{E}_{q_\phi(\hat{Z}|X)}[\Delta] - \mathbb{E}_{p(Z^*|X)}[\Delta^*] \right|}_{\text{(I) Model approximation}} \\
& + \underbrace{\left| \frac{1}{N} \sum_n \mathbb{E}_{q_\phi}[\Delta \mid x_n] - \mathbb{E}_X[\mathbb{E}_{q_\phi}[\Delta \mid X]] \right|}_{\text{(II) Finite sample}} \\
& + \underbrace{\left| \frac{1}{S} \sum_s \Delta_s - \mathbb{E}_{q_\phi}[\Delta \mid x] \right|}_{\text{(III) Monte Carlo}},
\end{aligned} \tag{26}
$$

where $\Delta = \mathbb{E}[Y \mid C_i{=}1, \hat{Z}_{\mathrm{adj}}] - \mathbb{E}[Y \mid C_i{=}0, \hat{Z}_{\mathrm{adj}}]$.

(I) Model approximation error: By Assumption 4.2, neural networks have universal approximation capacity. As capacity and data increase, learned parameters approach the equivalence class of Theorem 4.5. Crucially, by Lemma 4.4, component-wise transformations introduce *no bias*—the backdoor formula holds exactly. Thus (I) $\to 0$.

(II) Finite sample error: Since $Y \in \{0, 1\}$, Δ is bounded in $[-1, 1]$. By the strong law of large numbers, as $N \to \infty$: (II) $\xrightarrow{\text{a.s.}} 0$.

(III) Monte Carlo error: For fixed x and model, MC samples are i.i.d. from $q_\phi(\hat{Z} \mid x)$. By the law of large numbers, as $S \to \infty$: (III) $\xrightarrow{\text{a.s.}} 0$.

Role of overlap (Assumption 4.3): This ensures the posterior $q_\phi(Z_{\mathrm{adj}} \mid x)$ has positive density over values needed for counterfactual evaluation, guaranteeing that:

- Backdoor integrals are well-defined (no division by zero),
- Outcome regression $\mathbb{E}[Y \mid C, Z_{\mathrm{adj}}]$ can be estimated for all relevant values,
- The learned model captures sufficient latent variation for counterfactual queries.

Combining bounds: For any $\epsilon > 0$, choose capacity, N, and S sufficiently large such that each error term is bounded by $\epsilon/3$. By the triangle inequality:

$$
\mathbb{P}(|\widehat{\mathrm{CaCE}}_i - \mathrm{CaCE}_i| < \epsilon) \to 1 \quad \text{as } N, S \to \infty, \text{ capacity} \to \infty. \tag{27}
$$

The same argument applies to individual effects $\widehat{\mathrm{ICaCE}}_i(x) \xrightarrow{p} \mathrm{ICaCE}_i(x)$ without term (II). □

F Data generating processes

F.1 MorphoMNIST details

We construct four confounding scenarios using MorphoMNIST morphological features. We define four binary concepts by thresholding morphological measurements: *thickness* (t), *intensity* (i), *slant* (s), and *width* (w).

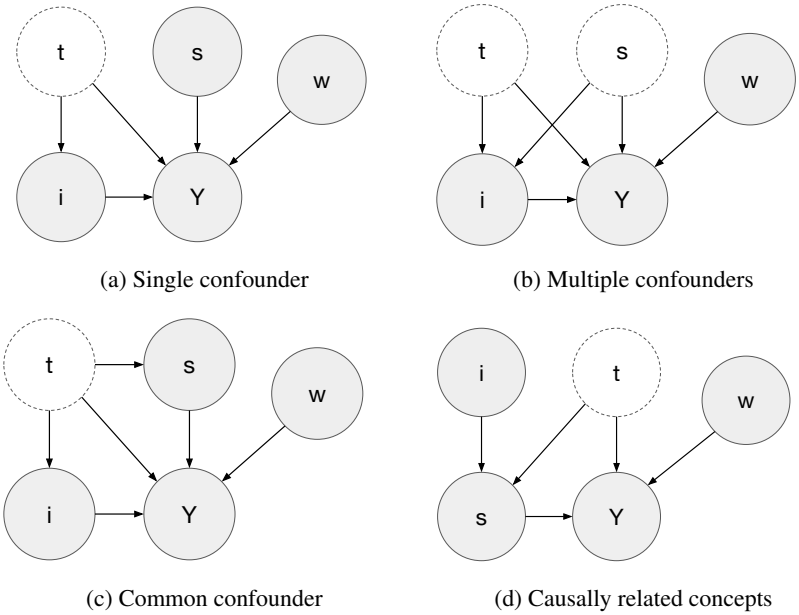
Single confounder: Thickness t is latent and causes intensity i. Specifically, we set $i = t$ with probability α and sample i independently with probability $1 - \alpha$. The outcome is $Y \sim \mathrm{Bernoulli}(0.3 \cdot i + 0.2 \cdot w + 0.3 \cdot s + 0.2 \cdot t)$ where only intensity, slant, and width are observed concepts during training (thickness remains latent).

Multiple confounders: Both thickness t and slant s are latent. With probability α, we set $i = t \vee s$, otherwise we sample i independently. $Y \sim \mathrm{Bernoulli}(0.3 \cdot i + 0.2 \cdot w + 0.3 \cdot s + 0.2 \cdot t)$ and only intensity and width are observed.

Common confounder: Thickness t is latent and influences both intensity and slant: $i = t$ with probability α and $s = t$ also with probability α (independently). The outcome is $Y \sim \mathrm{Bernoulli}(0.3 \cdot i + 0.2 \cdot w + 0.3 \cdot s + 0.2 \cdot t)$, and only intensity, slant, and width are observed.

Causally related concepts: Thickness t is latent and causes slant s jointly with intensity i (observed). With probability α, we set $s = t \vee i$, otherwise we sample s independently. Outcome is

given by $Y \sim \text{Bernoulli}(0.3 \cdot w + 0.3 \cdot s + 0.4 \cdot t)$. This tests how methods handle concept-to-concept causal relationships, while also accounting for latent confounders.



(a) Single confounder (b) Multiple confounders

(c) Common confounder (d) Causally related concepts

Figure 3: Data generating causal graphs for MorphoMNIST benchmark

In Figure 3 we depict the data generating processes for MorphoMNIST using causal DAGs. Across all scenarios, we use $\alpha = 0.9$ for in-distribution test sets and $\alpha = 0.6$ for out-of-distribution sets.

F.2 CelebA Details

We use the standard CelebA train/val/test splits with images resized to 64×64. To enable quantitative evaluation with known ground-truth effects, we replace the original *Attractiveness* label with a synthetic outcome:

$$Y \sim \text{Bernoulli}(0.5 \cdot \text{Age} - 0.3 \cdot \text{Gender} + 0.4 \cdot \text{Age} \cdot \text{Gender}), \tag{28}$$

where Age $\in \{0, 1\}$ (1=Young), and Gender $\in \{0, 1\}$ (1=Male). The interaction term creates non-additive dependence on both confounders, testing whether methods can recover complex causal structures.

G Baseline implementation details

Naive/CBM: We train concept predictors $q_\phi(C_i \mid X)$ using a CNN encoder ϕ, then train logistic regression $g : C \to Y$ on predicted concepts. For CaCE estimation, we compute $\widehat{\text{CaCE}}_i = \frac{1}{N} \sum_n [g(C_i{=}1, C_{-i,n}) - g(C_i{=}0, C_{-i,n})]$ where $C_{-i,n}$ are the predicted concepts from $q_\phi(C_{-i} \mid x_n)$.

Oracle: We augment observed concepts with true latent confounders and train logistic regression on the full set. This represents an upper bound on achievable performance if confounders were observable.

Image adjustment: Following Jerzak et al. (2023), we train propensity score models $e_i(x) = P(C_i{=}1 \mid X{=}x)$ using CNN encoders, then estimate CaCE via Inverse Probability Weighting. Stabilized weights are used with weight trimming to handle extreme propensities.

Res-CBM: We implement the post-hoc residual approach of Yuksekgonul et al. (2023). After training a standard CBM, we extract residual features by computing the difference between the encoder output and concept-predicted features. These residuals are then used as auxiliary adjustment variables in a regression framework.

CaCE-VAE: Following Goyal et al. (2020), we train a conditional VAE $p(X \mid C)$ with per-concept and style latents to generate counterfactual images $\tilde{x}_{i,c} \sim p(X \mid C_i{=}c, C_{-i})$. We then estimate effects by comparing outcome predictions from an auxiliary network: $\widehat{\mathrm{CaCE}}_i = \frac{1}{N}\sum_n [q(Y{=}1 \mid \tilde{x}_{i,1,n}) - q(Y{=}1 \mid \tilde{x}_{i,0,n})]$. This approach assumes unconfoundedness, treating concepts as fully observed.

CEVAE: We adapt the original tabular CEVAE (Louizos et al., 2017) to images by using CNN encoder-decoder architectures. The model learns a single latent variable Z without factorization, using a standard Gaussian prior $p(Z) = \mathcal{N}(0, I)$. Backdoor adjustment is performed using the learned latent Z to control for confounding.

TEDVAE: We implement the three-way factorization $(Z_C, Z_Y, Z_{\mathrm{conf}})$ as in Zhang et al. (2020), using CNN encoder-decoder architectures adapted to images. Unlike UnCoVAEr, TEDVAE uses standard Gaussian priors without auxiliary conditioning: $p(Z_{\mathrm{conf}}) = p(Z_C) = p(Z_Y) = \mathcal{N}(0, I)$, providing no identifiability guarantees beyond architectural inductive bias.

All methods use CNN encoder-decoder architectures with similar capacity: convolutional encoders with 4 layers downsampling to a feature dimension of 128 for MorphoMNIST and 256 for CelebA, and symmetric decoders with transposed convolutions. MLP components use 2 hidden layers with 256 hidden units.

H Extra results

Table 2 reports CaCE estimation error on the MorphoMNIST out-of-distribution test set, where confounding strength is reduced $\alpha = 0.6$ relative to training. This setting evaluates how well methods generalize under shifts in the confounding mechanism. Across all confounding structures, UnCoVAEr consistently achieves the lowest or near-lowest error among non-oracle methods, with particularly strong performance in the Multiple and Single confounder settings. The advantage over UnCoVAEr$_Z$ is most pronounced when multiple independent confounders are present, highlighting the benefit of explicitly separating confounders from concept- and outcome-specific latent factors. In contrast, methods that adjust directly on images or residual features exhibit unstable behavior under confounding shifts, while non-identifiable latent-variable approaches (CEVAE, TEDVAE) show limited robustness. Overall, these results demonstrate that identifiability and structured latent factorization are key to reliable causal effect estimation under distribution shift.

Table 2: Mean CaCE estimation error (lower is better) across methods for **MorphoMNIST** out-of-distribution test set ($\alpha = 0.6$). Results are averaged across concepts and reported as mean ± std over 5 seeds. Best non-oracle baseline per column is in **bold**

Method	Single	Multiple	Common	Causal
Naive	.064 ± .08	.146 ± .14	.094 ± .02	.145 ± .11
Oracle	.015 ± .02	.020 ± .01	.045 ± .03	.018 ± .01
Image-adjustment	.057 ± .05	.082 ± .07	.163 ± .10	.079 ± .05
Res-CBM	.217 ± .22	.226 ± .18	.146 ± .14	.106 ± .09
CaCE-VAE	.067 ± .07	.074 ± .06	.069 ± .06	.101 ± .05
CEVAE	.103 ± .05	.120 ± .05	.097 ± .05	.122 ± .11
TEDVAE	.054 ± .06	.092 ± .08	.058 ± .05	.172 ± .11
UnCoVAEr$_Z$	.046 ± .03	.088 ± .04	**.052 ± .03**	.093 ± .07
UnCoVAEr	**.045 ± .06**	**.033 ± .02**	.059 ± .04	**.077 ± .06**