Causal Representation Learning for Generalisable Recommendation

Yorgos Felekis[*]
University of Warwick

Michael O'Riordan[*]
Spotify

Oriol Corcoll
Spotify

Ciarán M. Gilligan-Lee
Spotify & University College London

Abstract

Predictive models trained on observational data often fail to generalise to the distributions they encounter when deployed, especially when the training data is a product of the system being optimised. Recommender systems are a canonical example: they are trained on interaction logs confounded by the deployed policy, past user behaviour, and platform filtering. As a result, the training distribution differs substantially from the candidate distribution scored at serving time, a gap that makes offline metrics unreliable predictors of online performance. We address the distribution shift problem with a method motivated by causal representation learning (CRL). We propose an information-theoretic disentanglement criterion and prove that its optimum excludes confounder-dependent non-causal information from the treatment representation. We then derive a tractable variational lower bound that makes the criterion optimisable from finite observational data alone. The scope of our method is narrower than that of much of the CRL literature, in that we target better generalisation under distribution shift, not full identification of all latent causal factors. This narrower target is what makes the method practical, requiring only the existing confounded logs, applying to any standard supervised model, and adding no inference-time cost. Our headline evaluation is an A/B test with millions of users on the music streaming platform Spotify, applied to a production ranker for personalised playlist generation. A CRL variant matched in capacity to the production baseline performed on par offline but delivered substantial online gains in listener engagement. Complementary evidence on a synthetic dataset with known causal structure and a semi-synthetic variant of the public KuaiRand recommendation dataset shows the same pattern: similar in-distribution performance to the baseline, gains under distribution shift. Across all three settings, adding our causal disentanglement objective yields better out-of-distribution generalisation.

1 Introduction

Predictive models deployed in feedback loops with the systems that produce their training data face a basic problem: the distribution they are trained on diverges from the one they are asked to score at serving time. In this work we focus on recommender systems, where the (user, item, context) tuples appearing in training logs are those a deployed policy already chose to surface, and are further filtered by user engagement. This phenomenon, broadly referred to as *exposure bias* in the recommendation literature Chen et al. [2023], means that at serving time the same model is asked to score a much broader and differently distributed candidate set. The gap between training and

[*]Equal contribution

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

serving-time distributions is widely understood to be a primary reason that offline ranking metrics are often unreliable predictors of online performance. Causal representation learning (CRL) Schölkopf et al. [2021] offers a principled solution. If a model's predictions rely only on causal mechanisms that generate the outcome, they should remain valid under shifts that leave those mechanisms intact, including the policy-induced shift between training and serving distributions. Classical CRL aims to recover latent causal factors with identifiability guarantees [Khemakhem et al., 2020]. Such identification requires additional assumptions or information [Locatello et al., 2019], which can be difficult to justify or obtain from logged observational recommendation data. A second strand applies CRL to recommendation directly, including in production. Wang et al. [2022] treat user feature drift as an intervention via a counterfactual VAE that infers latent user state and performs post-intervention inference at serving time. Corcoll et al. [2026] block-identify causal treatment representations via labelled contrastive pairs that stratify on confounders and outcomes. DCCL Zhao et al. [2023] and Rec4Ad Gao et al. [2023] disentangle specific bias mechanisms (popularity, sample-selection) with components tied to those mechanisms, and have been deployed online at Kuaishou and Taobao.

In this work, we assume that the relevant confounders are observed and present in the training data. In particular, we assume that confounders are encoded in the user, context, and item features that the deployed policy conditions on. In this case, full identification is more than what is needed to address the specific failure mode introduced by exposure bias. The spurious pathway can be suppressed directly by enforcing that the learned representation carries no information about the confounder beyond what the supervised target already requires (Figure 1). This narrower goal does not recover latent factors or enable counterfactual queries, but it does break the dependence that drives the offline-to-online gap, and it does so using only the existing observational training data. We instantiate this idea by learning a treatment representation that remains predictive of the outcome conditional on the observed confounder, while penalising mutual information between the learned representation and the confounder. The resulting criterion keeps the outcome-relevant causal content of the treatment and removes the confounder-dependent non-causal leakage that standard rankers may otherwise absorb. We make the following contributions:

- We propose an information-theoretic disentanglement criterion and prove that, at the population level, it favors representations that retain the causal content of the treatment and discard confounder-dependent non-causal information. We then derive a tractable variational lower bound that makes the criterion optimisable from finite observational data alone.
- We introduce a deployment-ready CRL recipe for robustness under distribution shift: a straightforward modification of any standard supervised model, trained on the existing confounded log, with no inference-time cost and no requirement for randomised or interventional data.
- We validate the method in three settings of increasing realism: a synthetic SCM, a semi-synthetic variant of the KuaiRand dataset Gao et al. [2022], and a live A/B test on a production personalised playlist generation surface with millions of users on the music streaming platform Spotify. Across all three settings, a capacity-matched CRL variant matches the baseline on in-distribution metrics and outperforms it when evaluation moves off the training distribution.

2 Method

2.1 Setting

We adopt the Structural Causal Model framework of Pearl [2009]. A causal model is a tuple$\mathcal{M} = (\mathbf{V}, \mathbf{U}, \mathcal{F}, P(\mathbf{U}))$, where $\mathbf{V}$ is the set of endogenous variables, $\mathbf{U}$ is the set of exogenous noise variables with joint distribution $P(\mathbf{U})$, and $\mathcal{F} = \{f_V\}_{V \in \mathbf{V}}$ is a collection of structural assignments $V := f_V(\mathrm{Pa}(V), U_V)$. In our setting, $\mathbf{V} = \{T, Y, Z\}$, where Z is an observed confounder (in recommendation, a user representation), T is the treatment to be ranked (an item), and Y is the outcome of interest. As illustrated in Figure 1, the observed treatment is an entangled mixing $T = m(T_C, T_{nC})$ of latent causal components T_C and non-causal components T_{nC}. We make two structural assumptions: **(A1)** $Y \perp T_{nC} \mid T_C, Z$ meaning that the outcome depends on the treatment only through its causal latents, given the confounder; **(A2)** $I(T_{nC}; Z \mid T_C) > 0$: which we call *unique spurious leakage*: the non-causal latents retain information about the confounder that is not already contained in the causal latents. This is the setting that arises in recommender systems whenever the deployed policy has correlated non-causal item-side factors with user-side signal, creating a spurious backdoor path $T \leftarrow T_{nC} \leftarrow Z \rightarrow Y$ that a standard predictor learning $\mathbb{E}[Y \mid T, Z]$ will absorb.

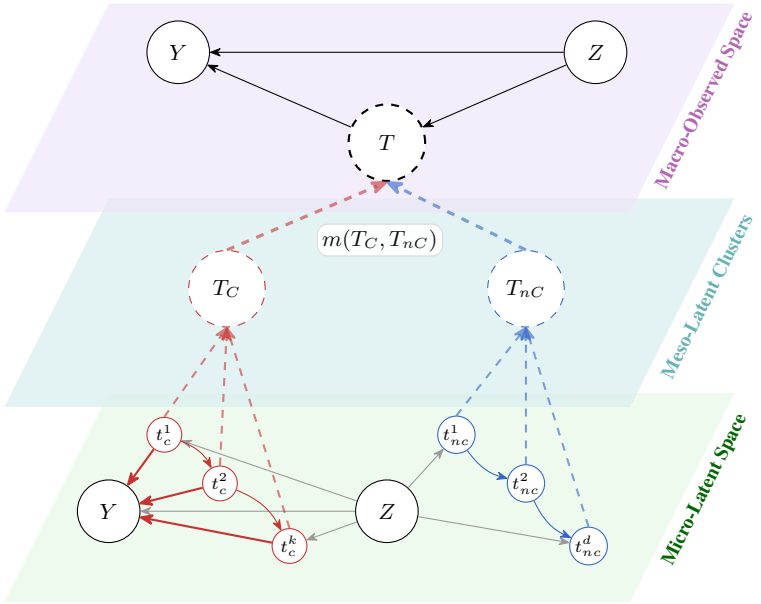

Figure 1: **Hierarchical Causal Entanglement.** *Bottom (Micro-Latent):* Only the red subset $\{t_c^i\}$ causally drives Y; the blue subset $\{t_{nc}^j\}$ contains spurious factors correlated with Z. *Middle (Meso-Latent):* Micro-variables abstract into causal (T_C) and non-causal (T_{nC}) clusters. *Top (Observed):* We observe only the entangled treatment $T = m(T_C, T_{nC})$, which creates the spurious backdoor path $T \leftarrow T_{nC} \leftarrow Z \rightarrow Y$ that standard models will exploit.

Such structures arise also in many causal representation learning settings where spurious correlations between treatment and confounders lead to biased models if not explicitly addressed.

While maximising $I(T; Y \mid Z)$ ensures the learned representation is predictive of the outcome, it may still preserve spurious signals as it does not penalise representations for carrying extra non-causal information. To address this, we introduce a more targeted objective that trades predictive utility against confounder dependence.

2.2 A New Objective for Causal Disentanglement

We propose to learn a representation $g(T)$ that maximizes the following:

$$J(g) = \underbrace{I(g(T); Y \mid Z)}_{\text{Utility}} - \lambda \underbrace{I(g(T); Z)}_{\text{Penalty}} \tag{1}$$

called the *Disentanglement Criterion*, where $\lambda \geq 0$ is a hyperparameter that controls the trade-off between predictive utility and confounder invariance. This objective balances two forces: the utility term rewards representations that are predictive of the outcome Y beyond Z, thereby encouraging the preservation of T_C; the penalty term discourages information shared with Z, thereby suppressing Z-correlated non-causal factors. Note that the penalty constrains the treatment representation $g(T)$, not the predictor's direct use of Z. The predictor still receives both $g(T)$ and Z. The objective discourages redundant confounder information in the treatment representation while rewarding information that predicts Y conditional on Z. Under our structural assumptions, the penalty term is exactly the channel through which T_{nC} leaks into the prediction via the spurious path $T_{nC} \leftarrow Z \rightarrow Y$. An equivalent bounded reparameterization is given in Appendix B. We first verify that the criterion ranks the ideal causal representation above the naive entangled one.

Proposition 1 (Ideal versus naive)**.** *Let $g^\star(T) = T_C$ and let $\tilde{g}(T) = T = (T_C, T_{nC})$ be a naive representation that retains both latent components. Under assumptions **A1** and **A2**, for any $\lambda > 0$,*

$$J(g^\star) > J(\tilde{g}). \tag{2}$$

Proof sketch. By the chain rule for CMI: $I(T_C, T_{nC}; Y \mid Z) = I(T_C; Y \mid Z) + I(T_{nC}; Y \mid T_C, Z)$. The second term is zero by **A1**, so the two representations have equal utility. For the penalty, $I(T_C, T_{nC}; Z) = I(T_C; Z) + I(T_{nC}; Z \mid T_C)$, and the second term is strictly positive by **A2**. Hence the naive representation incurs a strictly larger penalty while achieving the same utility. Therefore $J(g^\star) > J(\tilde{g})$. Full proof in Appendix A. □

Proposition 1 compares two fixed representations. We next formalise the more general mechanism through which the disentanglement criterion removes non-causal information. The observed treatment $T = m(T_C, T_{nC})$ may already lose information about T_C, as does a learned representation $g(T)$. We therefore define $\phi(T_C)$ as the information about T_C recoverable from $g(T)$ via a deterministic map r, i.e. $r(g(T)) = \phi(T_C)$. A *causally purified representation* is then $\bar{g} := r(g) = \phi(T_C)$. Because $\bar{g}$ is a deterministic function of g, the information in g decomposes exactly into the recovered causal content $\phi(T_C)$ and a residual beyond it. We assume that this residual carries no additional predictive information about Y given Z, but remains informative about the confounder, which are the representation-level counterparts of **A1** and **A2**: $I(g(T); Y \mid \phi(T_C), Z) = 0$, $I(g(T); Z \mid \phi(T_C)) > 0$. The following proposition shows that stripping the residual strictly improves J.

Proposition 2 (Causal purification)**.** *Let $g(T)$ be a representation satisfying the recoverability assumption, and define its causal purification as $\bar{g}(T) := r(g(T)) = \phi(T_C)$. Then, for any $\lambda > 0$,*

$$J(\bar{g}) > J(g). \tag{3}$$

Proof sketch. Since $\bar{g}$ is a deterministic function of g, we have $I(g; Z) = I(\bar{g}; Z) + I(g; Z \mid \bar{g})$.

By assumption, $I(g; Y \mid Z) = I(\phi(T_C); Y \mid Z) = I(\bar{g}; Y \mid Z)$, so the utility is preserved. The penalty gap is $I(g; Z) - I(\bar{g}; Z) = I(g; Z \mid \phi(T_C)) > 0$, which yields $J(\bar{g}) - J(g) = \lambda I(g; Z \mid \phi(T_C)) > 0$. Full proof in Appendix A. □

Remark 1. *Proposition 2 shows that the disentanglement criterion preserves the causal content available in the treatment representation while removing non-causal confounder-dependent information that is not needed for predicting Y. Hence, a maximising representation is purified with respect to the leakage targeted by the criterion and can be written as $g^\star(T) = \phi(T_C)$.*

However, the form of ϕ remains unspecified. Should it preserve all information in T_C (i.e. be invertible), or could a lossy compression be preferable? Since T_C is correlated with both Y (good for utility) and Z (costly for the penalty), any lossy ϕ reduces the penalty at the cost of some utility, with the trade-off governed by λ. To formalise this, let ϕ_{id} denote the identity (uncompressed) and let $\phi \in \mathcal{H}$ denote any non-invertible alternative transformation, where $\mathcal{H}$ is the class of all non-invertible lossy functions applicable to T_C. For each $\phi \in \mathcal{H}$ we define

$$\Delta U(\phi) := I(T_C; Y \mid Z) - I(\phi(T_C); Y \mid Z), \qquad \Delta P(\phi) := I(T_C; Z) - I(\phi(T_C); Z). \tag{4}$$

where $\Delta U(\phi)$ is the utility lost by compressing the causal component, while $\Delta P(\phi)$ is the reduction in confounding penalty gained by that compression. We assume a *no-free-lunch condition* according to which every nontrivial lossy transformation $\phi \in \mathcal{H}$ under consideration incurs a genuine utility loss and yields a genuine penalty reduction:

$$\Delta U(\phi) > 0 \text{ and } \Delta P(\phi) > 0. \tag{5}$$

We further define the *critical penalty weight* as the ratio of the two:

$$\Lambda(\phi) := \frac{\Delta U(\phi)}{\Delta P(\phi)}. \tag{6}$$

Proposition 3 (Lossless regime)**.** *Let $\lambda^\star := \inf_{\phi \in \mathcal{H}} \Lambda(\phi)$. Under the no-free-lunch condition* (5)*, for any $\lambda \in (0, \lambda^\star)$, the uncompressed representation T_C strictly dominates every lossy alternative $\phi(T_C)$ with $\phi \in \mathcal{H}$ under J; i.e. $J(T_C) > J(\phi(T_C))$, $\forall \phi \in \mathcal{H}$.*

Proof. Fix any $\phi \in \mathcal{H}$. Expanding the definition of J gives

$$I(T_C; Y \mid Z) - \lambda I(T_C; Z) > I(\phi(T_C); Y \mid Z) - \lambda I(\phi(T_C); Z) \iff \lambda < \frac{\Delta U(\phi)}{\Delta P(\phi)} = \Lambda(\phi)$$

Now suppose $\lambda \in (0, \lambda^\star)$. Since $\lambda < \lambda^\star = \inf_{\phi \in \mathcal{H}} \Lambda(\phi) \leq \Lambda(\phi)$ for every $\phi \in \mathcal{H}$, the condition $\lambda < \Lambda(\phi)$ holds uniformly, so $J(T_C) > J(\phi(T_C))$ for every $\phi \in \mathcal{H}$. Hence, throughout the regime $(0, \lambda^\star)$, the optimal representation preserves the full causal component. □

Proposition 3 shows that after non-causal leakage has been removed, sufficiently small penalty weights λ preserve the full causal component, and ϕ_{id} is optimal; compression of T_C becomes preferable only once the penalty reduction outweighs the corresponding utility loss. Finally, we provide a concrete sanity check for the preceding claims in Appendix C for a linear Gaussian SCM where all mutual-information terms can be computed in closed form: the appendix verifies that J strictly prefers the causal representation T_C over the naive entangled representation (T_C, T_{nC}), and illustrates the lossless regime by varying the amount of stochastic compression applied to T_C.

2.3 A Tractable Objective For Causal Disentanglement

The criterion in (1) cannot be optimised directly since both $I(g(T); Y \mid Z)$ and $I(g(T); Z)$ are unknown and intractable to compute from finite data. We therefore derive a tractable surrogate by replacing each term with a computable bound: a lower bound on the utility and an upper bound on the penalty so that the surrogate is a valid lower bound on $J(g)$.

Lower bound on the utility. For any conditional distribution $q(y \mid g(T), z)$, the standard variational bound on conditional mutual information (CMI) Barber and Agakov [2003] gives

$$I(g(T);\, Y \mid Z) \;\geq\; \mathbb{E}_{p(g(T),y,z)}\big[\log q(y \mid g(T), z)\big] + H(Y \mid Z), \tag{7}$$

which motivates learning q to approximate the true posterior $p(y \mid g(T), z)$. We instantiate this via a contrastive InfoNCE estimator van den Oord et al. [2018]: for each anchor $(g(T), y^+, z)$ we draw K negative outcomes $y_1^-, \ldots, y_K^-$ from $p(y \mid z)$ and define

$$\mathcal{L}_{\text{utility}}(\theta) := -\mathbb{E}\left[\log \frac{e^{f_\theta(g(T),\, y^+)/\tau}}{e^{f_\theta(g(T),\, y^+)/\tau} + \sum_{k=1}^{K} e^{f_\theta(g(T),\, y_k^-)/\tau}}\right]. \tag{8}$$

The InfoNCE loss directly lower-bounds the CMI van den Oord et al. [2018]:

$$I(g(T);\, Y \mid Z) \;\geq\; \log(K{+}1) - \mathcal{L}_{\text{utility}}(\theta) =: -\mathcal{L}_{\text{utility}}(\theta) + C_Y, \quad C_Y := \log(K{+}1). \tag{9}$$

Upper bound on the penalty. For any representation $g(T)$, and optimal InfoNCE critic $d_{\phi^\star}$, the NCE-CLUB estimator Liang et al. [2023] upper-bounds the mutual information between $g(T)$ and Z:

$$I(g(T);\, Z) \;\leq\; \underbrace{\mathbb{E}_{p(g(T),z)}\big[d_{\phi^\star}(g(T), z)\big] - \mathbb{E}_{p(g(T))p(z)}\big[d_{\phi^\star}(g(T), z)\big]}_{=:\, I_{\text{NCE-CLUB}}(g(T);\, Z;\, d_{\phi^\star})}. \tag{10}$$

The first term scores *positive* pairs $(g(T), z) \sim p(g, z)$, while the second scores *negative* pairs drawn independently from $p(g)p(z)$. The guarantee assumes an optimal critic recovering the log-density ratio $d_{\phi^\star}(g, z) = \log p(z \mid g) - \log p(z)$. In practice, we approximate this with a learnable critic $d_\phi(g(T), z)$, trained to discriminate joint samples $(g, z) \sim p(g, z)$ from independent samples $(g, z) \sim p(g)p(z)$. Further details are provided in Appendix D.

Substituting (9) and (10) into (1), yields a tractable variational lower bound for $J(g)$.

Theorem 1. *For any encoder g_θ and optimal InfoNCE critic $d_{\phi^\star}$,*

$$J(g_\theta) \;\geq\; -\mathcal{L}_{\text{utility}}(\theta) - \lambda\, I_{\text{NCE-CLUB}}(g_\theta(T);\, Z;\, d_{\phi^\star}) + C_Y. \tag{11}$$

Since $C_Y = \log(K{+}1)$ does not depend on $(\theta, \phi^\star)$, maximising this bound is equivalent to minimising

$$\mathcal{L}_{\text{final}}(\theta, \phi^\star) := \mathcal{L}_{\text{utility}}(\theta) + \lambda\, \mathcal{L}_{\text{penalty}}(\theta, \phi^\star), \quad \mathcal{L}_{\text{penalty}}(\theta, \phi^\star) := I_{\text{NCE-CLUB}}(g_\theta(T);\, Z;\, d_{\phi^\star}). \tag{12}$$

The resulting objective augments the standard ranking loss with an explicit mutual-information penalty that discourages dependence between the learned representation and the confounder, unlike IPM-based methods Shalit et al. [2017] which require explicit population matching.

2.4 Practical implementation

The tractable surrogate objective in Theorem 1 uses NCE-CLUB to upper-bound the mutual-information penalty. As described in Appendix D, direct optimisation alternates between fitting the critic with InfoNCE and updating the encoder against the resulting NCE-CLUB estimate, together

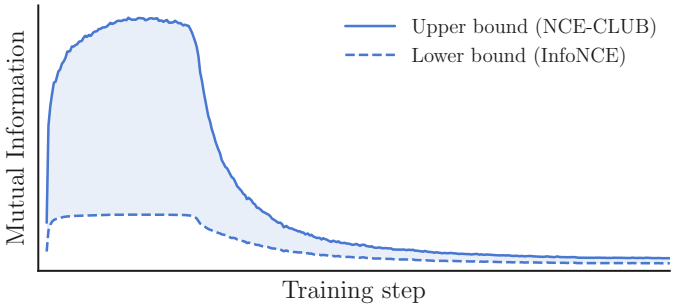

Figure 2: InfoNCE lower bound (dashed) and NCE-CLUB upper bound (solid) on $I(g(T); Z)$ during a production CRL training run. Gradient reversal acts on the InfoNCE bound. NCE-CLUB is tracked only as a diagnostic. Both rise during the critic warm-up, then decay as gradient reversal drives MI down. Axis values are omitted for confidentiality. As discussed in Section 2.4, a conservative reading of the bounds gives a reduction in MI of at least a factor of 5.

with the supervised utility loss. In our experiments, we instead find it convenient to apply gradient reversal [Ganin and Lempitsky, 2015] on the InfoNCE objective [van den Oord et al., 2018]. This allows the encoder and critic to be updated through a single backward pass, which was simpler to integrate into our existing production training pipeline. We emphasise that this choice is motivated by implementation convenience and is not a requirement of the proposed disentanglement criterion.

Of course, minimising an InfoNCE lower bound does not guarantee a decrease in the mutual information. We therefore track both the InfoNCE lower-bound estimate and the NCE-CLUB upper-bound estimate during training. Figure 2 shows a production run in which both estimates rise during critic warm-up and then decrease once gradient reversal becomes active. In particular, the NCE-CLUB estimate decreases even though it is not directly optimised. To quantify the MI reduction relative to the non-CRL baseline model, we compare the final CRL upper bound, U_{CRL}, with the baseline lower bound, L_{base}. Writing I_{CRL} and I_{base} for their respective actual MI, the validity conditions of the bounds give $I_{\mathrm{CRL}} \leq U_{\mathrm{CRL}}$ and $L_{\mathrm{base}} \leq I_{\mathrm{base}}$. In our production setup, this diagnostic gives $U_{\mathrm{CRL}} \leq L_{\mathrm{base}}/5$, which implies $I_{\mathrm{CRL}} \leq U_{\mathrm{CRL}} \leq \frac{1}{5} L_{\mathrm{base}} \leq \frac{1}{5} I_{\mathrm{base}}$. Therefore, assuming valid bounds, we can conclude that the MI has decreased by at least a factor of 5 relative to the non-CRL model. If this NCE-CLUB diagnostic does not support a reduction in MI, we recommend direct optimisation of the NCE-CLUB penalty following Appendix D.

An excessively strong penalty can remove information useful for prediction. Proposition 3 characterises the ideal lossless regime, but its population threshold is not directly observable. Moreover, the scale of λ depends on the parameterisation of the utility and penalty losses, the critic, and the application. We therefore treat λ as a standard training hyperparameter and suggest using in-distribution validation performance as a safeguard against over-regularisation. We recommend increasing λ from zero and selecting the strongest regularisation that keeps the validation metric within an application-specific tolerance of the $\lambda = 0$ baseline. The MI diagnostics provide additional guidance on whether the penalty is adequately reducing confounder dependence. Consistent with this approach, our CRL and non-CRL models remain close on in-distribution metrics, while the CRL models improve under distribution shift.

3 Related work

Treatment-effect estimation with neural networks. CFR Shalit et al. [2017] augments TARNet with an integral-probability-metric penalty that balances covariate representations across treatment groups. CEVAE Louizos et al. [2017] addresses unobserved confounders with a deep generative model using proxies. See also Jeunen et al. [2022], Van Goffrier et al. [2023], Fawkes et al. [2025] for alternate methods of dealing with unobserved confounders. Dragonnet Shi et al. [2019] adds a propensity-prediction head to encourage representations sufficient for treatment-effect identification. Deep Twin Networks Vlontzos et al. [2023, 2026], Reynaud et al. [2022] adapt neural networks to answer counterfactual questions Richens et al. [2020], Pearl [2009] by ensuring their architecture matches the twin network Pearl [2009], Graham et al. [2019] of the causal structure in question.

Related work studies structured treatments Kaddour et al. [2021], Harada and Kashima [2021] and effects of linguistic properties Pryzant et al. [2021]. GraphITE penalises dependence between treatment and individual representations using HSIC. We instead use an information-theoretic criterion with a causal-purification analysis. Corcoll et al. [2026] learn causal treatment representations using contrastive pairs stratified on confounders and outcomes, whereas our approach does not require constructing stratified contrastive pairs.

Debiasing causal recommendation. Wang et al. [2022] address shifts in user attributes with a causal variational autoencoder that predicts preferences under updated attributes while reducing the influence of outdated interaction history. Debiasing approaches include IPS reweighting Schnabel et al. [2016], Joachims et al. [2017], pseudo-rating-based asymmetric tri-training Saito [2020], instrumental-variable methods such as IViDR Deng et al. [2024], proxy-based latent-confounder identification in iDCF Zhang et al. [2023], and causal adjustment Wang et al. [2020], Zhang et al. [2021]. CausE Bonner and Vasile [2018] couples embeddings learned from biased and unbiased exposure data, while DICE Zheng et al. [2021] and DCCL Zhao et al. [2023] disentangle interest and conformity. Rec4Ad Gao et al. [2023] augments advertising logs with organic recommendation samples and learns aligned, decorrelated representations. Our method instead regularises treatment-confounder dependence within the existing training log.

Adversarial MI estimation and invariance. The practical estimator combines gradient reversal Ganin and Lempitsky [2015] with the InfoNCE lower bound van den Oord et al. [2018]. Alternative MI estimators include MINE Belghazi et al. [2018] and the CLUB upper bound Cheng et al. [2020]; NCE-CLUB Liang et al. [2023] provides the upper-bound machinery used in our diagnostics (Appendix D). MI-based representation objectives are also at the core of Deep InfoMax Hjelm et al. [2019]. The disentanglement criterion $J(g)$ is complementary to invariance-based approaches, including invariant causal prediction Peters et al. [2016] and invariant risk minimisation Arjovsky et al. [2019]. These exploit multiple environments, whereas we penalise dependence on an observed confounder without requiring an environment partition.

Causal bottlenecks and abstractions. Our objective is related to causal information-bottleneck methods, which retain task-relevant causal information while discarding irrelevant variation. The Causal Information Bottleneck Simoes et al. [2025] learns causal variable abstractions by compressing an input while preserving causal control over a target, and Structural Causal Bottleneck Models Bing et al. [2026] assume that causal effects between high-dimensional SCM variables are mediated by lower-dimensional bottlenecks of their causes. These perspectives connect to causal abstraction Rubenstein et al. [2017], which seeks coarser variables or models that are interventionally consistent with the original system, together with methods that learn such maps Zennaro et al. [2023], Felekis et al. [2024], Kekić et al. [2024], Xia and Bareinboim [2024], Felekis et al. [2026]. In the language of Figure 1, such methods would ask whether the entangled $T = m(T_C, T_{nC})$ admits a coarser description, ideally T_C, that preserves interventional consistency. Our setting is adjacent but deliberately weaker: with m fixed and unobserved, we learn a map $T \mapsto g(T)$ sufficient for the recoverable causal content $\phi(T_C)$, while removing the Z-correlated non-causal leakage carried by T_{nC}. We do not identify T_C, learn m, or verify model-level consistency; we learn a representation that retains the causal information needed for prediction.

4 Experiments

We evaluate the proposed method in three settings of increasing realism. The synthetic SCM (§4.1) shows that the method successfully isolates the causal components of the treatment representation in a simple setup where the full data generation process is known. The semi-synthetic KuaiRand experiment (§4.2) uses real user and video features, with a synthetic nonlinear outcome function to test robustness to a controlled change in non-causal feature assignment. The online A/B test (§4.3) verifies that this robustness translates into measurable online gains on a deployed production model at Spotify.

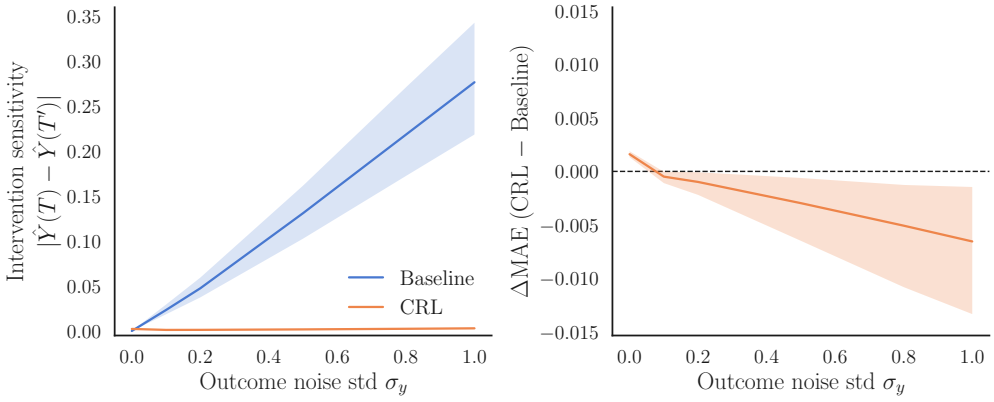

Figure 3: Synthetic SCM, sweep over outcome-noise scale $\sigma_Y \in [0, 1]$ across 2,000 seeds per noise level. Lines are per-seed medians; shaded bands are the inter-quartile range across seeds. **Left:** intervention sensitivity $|\hat{Y}(T) - \hat{Y}(T')|$, where T' resamples the non-causal coordinates of T; lower is better, with 0 indicating that the model has not absorbed any signal from T_{nC}. **Right:** paired $\Delta\text{MAE} = \text{MAE}_{\text{CRL}} - \text{MAE}_{\text{baseline}}$ per seed, centered on the zero line. The CRL penalty does not degrade predictive performance, and is in fact slightly better at higher noise levels.

4.1 Synthetic SCM

We instantiate a linear additive SCM with a confounder $Z \in \mathbb{R}^{10}$, a treatment $T \in \mathbb{R}^{10}$ decomposable into $d_c = 5$ causal and 5 non-causal dimensions, and an outcome Y that depends on T_C and Z but not T_{nC}. Full data-generating equations are given in Appendix E. Since Y is invariant to interventions on T_{nC}, a model that has correctly isolated the causal part of the treatment should also be invariant to those interventions. We measure the *intervention sensitivity* $|\hat{Y}(T) - \hat{Y}(T')|$ where T' has its non-causal coordinates resampled, and the *prediction MAE* on Y. In Figure 3 we sweep the outcome noise σ_Y from 0 to 1, and show how the baseline predictor is sensitive to interventions on the T_{nC} non-causal signal, whereas the CRL variant's intervention sensitivity sits essentially at zero across the entire range. The MAE is matched between the two methods.

4.2 Semi-Synthetic KuaiRand

In this section, we evaluate our method in a semi-synthetic setup using real user and video features from KuaiRand-Pure [Gao et al., 2022]. We generate synthetic clicks with a fixed nonlinear function of user features Z, video features T_C, and their interactions. We introduce a non-causal feature T_{nC} and append it to the video input. In the source distribution, this feature is correlated with Z, providing a shortcut which the model can exploit for prediction. In the target distribution, T_{nC} is assigned uniformly at random. The source and target evaluations use the same held-out users, videos, and synthetic clicks, with only the shortcut assignment changing.

We use the DeepFM [Guo et al., 2017] architecture with user and video features and learned ID embeddings. DeepFM is a standard baseline in the recommendation literature [see eg. Wang et al., 2021], and has been applied to the KuaiRand dataset [Wang et al., 2024, Huang et al., 2024]. Our CRL variant adds an InfoNCE critic between the treatment and user representations, with gradient reversal applied to the treatment representation. The predictor architecture and capacity are the same as the baseline. Details of the data-generation, architecture, training setup, and metrics are in Appendix F.

We compare our CRL approach with the baseline, DCCL [Zhao et al., 2023], and two oracle methods. DCCL is similar to our approach in that it adds contrastive losses during training without requiring target-domain data. However, it targets popularity and conformity through user-conditioned positive and negative interaction pairs. Our objective does not specify a particular bias mechanism and does not use user or outcome-specific positive/negative pair construction. We further note that DCCL's specific bias mechanism and pair construction are too restrictive for our production application. Oracle IPS uses the exact target-to-source T_{nC} probability ratio. Oracle T_{nC} randomization resamples each

Table 1: Semi-synthetic KuaiRand results over 100 seeds. AUC entries report means ± 95% confidence intervals. Ranking metrics report target means. Source and target use the same held-out rows, changing only T_{nC} assignment. Oracle methods receive information about the true assignment mechanism. Our CRL model improves all target metrics over the baseline and approaches oracle T_{nC} randomization performance.

Method	Source AUC	Target AUC	Recall@5	NDCG@5
Baseline	0.7249 ± 0.0004	0.6914 ± 0.0011	0.5697	0.4501
DCCL	0.7248 ± 0.0004	0.6912 ± 0.0011	0.5699	0.4501
Oracle IPS	0.7079 ± 0.0008	0.7078 ± 0.0008	0.5850	0.4678
Oracle T_{nC} randomization	0.7243 ± 0.0004	0.7243 ± 0.0004	0.5884	0.4717
CRL (ours)	0.7218 ± 0.0005	0.7212 ± 0.0004	0.5853	0.4676

Table 2: Online A/B test results on the personalised playlist generation surface. Percentage change of the CRL variant relative to the production baseline, with 95% confidence intervals.

Metric	% change	95% CI
Track Streams	$+0.75\%$	$[+0.54\%, +0.96\%]$
Track Skips	-0.61%	$[-0.74\%, -0.47\%]$
Minutes Played	$+0.50\%$	$[+0.27\%, +0.73\%]$

training example's T_{nC} uniformly. Both oracle approaches rely on knowledge of the synthetic assignment mechanism.

In Table 1, we show that the baseline is sensitive to distribution shifts in the non-causal shortcut T_{nC}, with AUC degrading between the source and target distributions. DCCL performs similarly to the baseline in this setup. Oracle IPS shows little or no degradation between source and target, but large weights and low effective sample size reduce overall performance in this setup. The baseline and CRL models are close on the source distribution, but CRL is robust to the distribution shift, substantially reducing the AUC degradation between source and target. CRL outperforms the baseline across all target metrics, and approaches the performance of the oracle shortcut randomization. This pattern of similar performance on the source distribution combined with gains on the target distribution is similar to what Taori et al. [2020] call *effective robustness*: improvement under shift not predicted by in-distribution performance alone. In the online A/B test (Section 4.3) we show how this improved generalisation ability translates to real gains in listener engagement.

4.3 Online A/B test on personalised playlist generation

We evaluate real-world impact through an online A/B test with millions of users on a production personalised playlist generation surface at Spotify. The test compares two versions of the production session-ranking model: the deployed baseline and a capacity-matched CRL variant trained with the disentanglement objective of §2. The CRL variant uses an adversarial loss to decorrelate the model's internal representation of the candidate track from its representations of session- and user-level context, mirroring the recipe used in the KuaiRand experiment. Both variants perform similarly on in-distribution offline metrics, but the CRL variant performs better online. Table 2 reports the headline engagement metrics, measured over two weeks during the test. The CRL variant increases track streams by $+0.75\%$ and minutes played by $+0.50\%$ while decreasing track skips by -0.61%, with all three changes significant at the 95% level. For context, these gains are of the same order of magnitude as the improvements that the current production ranker, itself a ground-up rebuild, delivered over its mature predecessor.

4.4 Discussion

Across all three settings, in-distribution performance remains unchanged while performance under distribution shift improves. The synthetic and semi-synthetic experiments demonstrate robustness to known non-causal treatment variation, and the production A/B test demonstrates online gains in listener engagement. Our theoretical analysis assumes that the relevant confounding information

is observed in Z. The penalty cannot remove dependence transmitted only through unobserved confounders. In production, the available user and session context almost surely does not include all relevant confounding factors. The observed gains nevertheless suggest that an informative but incomplete Z can still improve generalisation. Combining the criterion with proximal causal inference is a possible extension, requiring suitable proxies and additional identifying assumptions.

5 Conclusion

We proposed a CRL-motivated method for making predictive models more robust under the distribution shift characteristic of on-policy recommendation. We introduced an information-theoretic disentanglement criterion, characterised its properties, and derived a tractable training objective. The method uses existing observational logs and adds no inference-time cost. Synthetic and semi-synthetic experiments demonstrate improved robustness under controlled shifts, while a large-scale production A/B test showed significant gains in listener engagement.

Acknowledgments and Disclosure of Funding

This work was supported by Spotify. The majority of Yorgos Felekis's contribution was completed during an internship at Spotify. Michael O'Riordan, Oriol Corcoll, and Ciarán M. Gilligan-Lee were affiliated with Spotify during this work. The authors declare no additional competing interests.

References

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

David Barber and Felix V. Agakov. Information maximization in noisy channels: A variational approach. In *Advances in Neural Information Processing Systems 16 (NIPS)*. MIT Press, 2003.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *International Conference on Machine Learning (ICML)*, 2018.

Simon Bing, Jonas Wahl, and Jakob Runge. Structural causal bottleneck models, 2026. URL `https://arxiv.org/abs/2603.08682`.

Stephen Bonner and Flavian Vasile. Causal embeddings for recommendation. In *Proceedings of the 12th ACM Conference on Recommender Systems (RecSys)*, 2018.

Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. Bias and debias in recommender system: A survey and future directions. *ACM Transactions on Information Systems*, 41(3):1–39, 2023.

Pengyu Cheng, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan, and Lawrence Carin. CLUB: A contrastive log-ratio upper bound of mutual information. In *International Conference on Machine Learning (ICML)*, 2020.

Oriol Corcoll, Athanasios Vlontzos, Michael O'Riordan, and Ciarán M. Gilligan-Lee. Contrastive representations of structured treatments. *npj Artificial Intelligence*, 2:49, 2026. doi: 10.1038/s44387-026-00105-2. URL `https://www.nature.com/articles/s44387-026-00105-2`.

Jianfeng Deng, Qingfeng Chen, Debo Cheng, Jiuyong Li, Lin Liu, and Xiaojing Du. Mitigating dual latent confounding biases in recommender systems. *arXiv preprint arXiv:2410.12451*, 2024.

Jake Fawkes, Michael O'Riordan, Athanasios Vlontzos, Oriol Corcoll, and Ciarán Mark Gilligan-Lee. The hardness of validating observational studies with experimental data. *arXiv preprint arXiv:2503.14795*, 2025.

Yorgos Felekis, Fabio Massimo Zennaro, Nicola Branchini, and Theodoros Damoulas. Causal optimal transport of abstractions. In *Causal Learning and Reasoning*, pages 462–498. PMLR, 2024.

Yorgos Felekis, Theodoros Damoulas, and Paris Giampouras. Distributionally robust causal abstractions, 2026. URL https://arxiv.org/abs/2510.04842.

Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning (ICML)*, 2015.

Chongming Gao, Shijun Li, Yuan Zhang, Jiawei Chen, Biao Li, Wenqiang Lei, Peng Jiang, and Xiangnan He. KuaiRand: An unbiased sequential recommendation dataset with randomly exposed videos. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management (CIKM)*, 2022.

Jingyue Gao, Shuguang Han, Han Zhu, Siran Yang, Yuning Jiang, Jian Xu, and Bo Zheng. Rec4Ad: A free lunch to mitigate sample selection bias for ads CTR prediction in Taobao. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM)*, 2023. doi: 10.1145/3583780.3615496.

Logan Graham, Ciarán M Lee, and Yura Perov. Copy, paste, infer: a robust analysis of twin networks for counterfactual inference. In *NeurIPS19 CausalML workshop*, 2019.

Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, and Xiuqiang He. DeepFM: A factorization-machine based neural network for CTR prediction. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2017.

Shonosuke Harada and Hisashi Kashima. GraphITE: Estimating individual effects of graph-structured treatments. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management (CIKM)*, pages 659–668, 2021. doi: 10.1145/3459637.3482349.

R. Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations (ICLR)*, 2019.

Jiahui Huang, Lan Zhang, Junhao Wang, Shanyang Jiang, Dongbo Huang, Cheng Ding, and Lan Xu. Utilizing non-click samples via semi-supervised learning for conversion rate prediction. In *Proceedings of the 18th ACM Conference on Recommender Systems (RecSys)*, pages 350–359, 2024. doi: 10.1145/3640457.3688151.

Olivier Jeunen, Ciarán Gilligan-Lee, Rishabh Mehrotra, and Mounia Lalmas. Disentangling causal effects from sets of interventions in the presence of unobserved confounders. *Advances in Neural Information Processing Systems*, 35:27850–27861, 2022.

Thorsten Joachims, Adith Swaminathan, and Tobias Schnabel. Unbiased learning-to-rank with biased feedback. In *Proceedings of the 10th ACM International Conference on Web Search and Data Mining (WSDM)*, 2017.

Jean Kaddour, Yuchen Zhu, Qi Liu, Matt J. Kusner, and Ricardo Silva. Causal effect inference for structured treatments. In *Advances in Neural Information Processing Systems 34 (NeurIPS)*, 2021.

Armin Kekić, Bernhard Schölkopf, and Michel Besserve. Targeted reduction of causal models. In *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, UAI '24. JMLR.org, 2024.

Ilyes Khemakhem, Diederik P. Kingma, Ricardo Pio Monti, and Aapo Hyvärinen. Variational autoencoders and nonlinear ICA: a unifying framework. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020.

Paul Pu Liang, Zihao Deng, Martin Q. Ma, James Zou, Louis-Philippe Morency, and Ruslan Salakhutdinov. Factorized contrastive learning: going beyond multi-view redundancy. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning (ICML)*, 2019.

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

Judea Pearl. Causal inference in statistics: An overview. *Statistics Surveys*, 3:96–146, 2009.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B*, 78(5):947–1012, 2016.

Reid Pryzant, Dallas Card, Dan Jurafsky, Victor Veitch, and Dhanya Sridhar. Causal effects of linguistic properties. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 4095–4109. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.naacl-main.323.

Hadrien Reynaud, Athanasios Vlontzos, Mischa Dombrowski, Ciarán Gilligan Lee, Arian Beqiri, Paul Leeson, and Bernhard Kainz. D'artagnan: Counterfactual video generation. In *International conference on medical image computing and computer-assisted intervention*, pages 599–609. Springer, 2022.

Jonathan G Richens, Ciarán M Lee, and Saurabh Johri. Improving the accuracy of medical diagnosis with causal machine learning. *Nature communications*, 11(1):3923, 2020.

Paul K Rubenstein, Sebastian Weichwald, Stephan Bongers, Joris M Mooij, Dominik Janzing, Moritz Grosse-Wentrup, and Bernhard Schölkopf. Causal consistency of structural equation models. In *33rd Conference on Uncertainty in Artificial Intelligence (UAI 2017)*, pages 808–817. Curran Associates, Inc., 2017.

Yuta Saito. Asymmetric tri-training for debiasing missing-not-at-random explicit feedback. In *Proceedings of the 43rd ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020.

Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. Recommendations as treatments: Debiasing learning and evaluation. In *International Conference on Machine Learning (ICML)*, 2016.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

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

Claudia Shi, David M. Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, 2019.

Francisco N. F. Q. Simoes, Mehdi Dastani, and Thijs van Ommen. The causal information bottleneck and optimal causal variable abstractions. In Silvia Chiappa and Sara Magliacane, editors, *Proceedings of the Forty-first Conference on Uncertainty in Artificial Intelligence*, volume 286 of *Proceedings of Machine Learning Research*, pages 3878–3897. PMLR, 21–25 Jul 2025. URL `https://proceedings.mlr.press/v286/simoes25a.html`.

Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Graham Van Goffrier, Lucas Maystre, and Ciarán Mark Gilligan-Lee. Estimating long-term causal effects from short-term experiments and long-term observational data with unobserved confounding. In *Conference on Causal Learning and Reasoning*, pages 791–813. PMLR, 2023.

Athanasios Vlontzos, Bernhard Kainz, and Ciarán M Gilligan-Lee. Estimating categorical counterfactuals via deep twin networks. *Nature Machine Intelligence*, 5(2):159–168, 2023.

Athanasios Vlontzos, David Gustafsson, Michael O'Riordan, and Ciarán M Gilligan-Lee. Incremental recommendation via causal models. *arXiv preprint arXiv:2608.26804*, 2026.

Ruoxi Wang, Rakesh Shivanna, Derek Cheng, Sagar Jain, Dong Lin, Lichan Hong, and Ed H. Chi. DCN V2: Improved deep & cross network and practical lessons for web-scale learning to rank systems. In *Proceedings of the Web Conference 2021 (WWW)*, pages 1785–1797, 2021. doi: 10.1145/3442381.3450078.

Wenjie Wang, Xinyu Lin, Fuli Feng, Xiangnan He, Min Lin, and Tat-Seng Chua. Causal representation learning for out-of-distribution recommendation. In *Proceedings of the ACM Web Conference 2022 (WWW)*, pages 3562–3571, 2022. doi: 10.1145/3485447.3512251.

Yejing Wang, Dong Xu, Xiangyu Zhao, Zhiren Mao, Peng Xiang, Ling Yan, Yao Hu, Zijian Zhang, Xuetao Wei, and Qidong Liu. Bi-level user modeling for deep recommenders. In *2024 IEEE International Conference on Data Mining (ICDM)*, pages 510–519, 2024. doi: 10.1109/ICDM59182.2024.00058.

Yixin Wang, Dawen Liang, Laurent Charlin, and David M. Blei. Causal inference for recommender systems. In *Proceedings of the 14th ACM Conference on Recommender Systems (RecSys)*, pages 426–431, 2020. doi: 10.1145/3383313.3412225.

Kevin Xia and Elias Bareinboim. Neural causal abstractions. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(18):20585–20595, Mar. 2024. doi: 10.1609/aaai.v38i18.30044. URL `https://ojs.aaai.org/index.php/AAAI/article/view/30044`.

Fabio Massimo Zennaro, Máté Drávucz, Geanina Apachitei, W. Dhammika Widanage, and Theodoros Damoulas. Jointly learning consistent causal abstractions over multiple interventional distributions. In Mihaela van der Schaar, Cheng Zhang, and Dominik Janzing, editors, *Proceedings of the Second Conference on Causal Learning and Reasoning*, volume 213 of *Proceedings of Machine Learning Research*, pages 88–121. PMLR, 11–14 Apr 2023. URL `https://proceedings.mlr.press/v213/zennaro23a.html`.

Qing Zhang, Xiaoying Zhang, Yang Liu, Hongning Wang, Min Gao, Jiheng Zhang, and Ruocheng Guo. Debiasing recommendation by learning identifiable latent confounders. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2023.

Yang Zhang, Fuli Feng, Xiangnan He, Tianxin Wei, Chonggang Song, Guohui Ling, and Yongdong Zhang. Causal intervention for leveraging popularity bias in recommendation. In *Proceedings of the 44th ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021.

Weiqi Zhao, Dian Tang, Xin Chen, Dawei Lv, Daoli Ou, Biao Li, Peng Jiang, and Kun Gai. Disentangled causal embedding with contrastive learning for recommender system. In *Companion Proceedings of the ACM Web Conference 2023 (WWW '23 Companion)*, 2023.

Yu Zheng, Chen Gao, Xiang Li, Xiangnan He, Yong Li, and Depeng Jin. Disentangling user interest and conformity for recommendation with causal embedding. In *Proceedings of the web conference 2021*, pages 2980–2991, 2021.

A Proofs for §2

Throughout, we use the structural assumptions introduced in §2. The observed treatment decomposes as

$$T = m(T_C, T_{nC}),$$

the outcome is conditionally independent of the non-causal latents,

$$\textbf{(A1)} \qquad Y \perp T_{nC} \mid (T_C, Z),$$

and the non-causal latents contain residual confounder information beyond the causal latents,

$$\textbf{(A2)} \qquad I(T_{nC}; Z \mid T_C) > 0.$$

Proof of Proposition 1

We compare the ideal representation $g^\star(T) = T_C$ with the naive representation $\tilde{g}(T) = T = (T_C, T_{nC})$.

Utility. By the chain rule for conditional mutual information,

$$\begin{aligned} I(\tilde{g}(T); Y \mid Z) &= I(T_C, T_{nC}; Y \mid Z) \\ &= I(T_C; Y \mid Z) + I(T_{nC}; Y \mid T_C, Z). \end{aligned}$$

The second term is zero by **A1**. Hence

$$I(\tilde{g}(T); Y \mid Z) = I(T_C; Y \mid Z) = I(g^\star(T); Y \mid Z).$$

Penalty. Again, by the mutual-information chain rule,

$$\begin{aligned} I(\tilde{g}(T); Z) &= I(T_C, T_{nC}; Z) \\ &= I(T_C; Z) + I(T_{nC}; Z \mid T_C). \end{aligned}$$

By **A2**, $I(T_{nC}; Z \mid T_C) > 0$. Therefore

$$I(\tilde{g}(T); Z) > I(T_C; Z) = I(g^\star(T); Z).$$

Equivalently, using the entropy chain rule,

$$\begin{aligned} I(T_C, T_{nC}; Z) &= H(T_C, T_{nC}) - H(T_C, T_{nC} \mid Z) \\ &= \big[H(T_C) + H(T_{nC} \mid T_C)\big] - \big[H(T_C \mid Z) + H(T_{nC} \mid T_C, Z)\big] \\ &= I(T_C; Z) + I(T_{nC}; Z \mid T_C). \end{aligned}$$

Thus, the excess penalty of the naive representation is exactly the unique spurious leakage term.

Conclusion. The two representations have equal utility, but the naive representation has a strictly larger penalty. Hence

$$\begin{aligned} J(g^\star) - J(\tilde{g}) &= \big[I(g^\star; Y \mid Z) - I(\tilde{g}; Y \mid Z)\big] + \lambda\big[I(\tilde{g}; Z) - I(g^\star; Z)\big] \\ &= \lambda\, I(T_{nC}; Z \mid T_C) > 0. \end{aligned}$$

Therefore $J(g^\star) > J(\tilde{g})$. □

Proof of Proposition 2

Consider a representation satisfying the recoverability assumption:

$$g(T) = h\big(\phi(T_C), \ell(T_{nC})\big),$$

where the causal component $\phi(T_C)$ is deterministically recoverable from $g(T)$. That is, there exists a map r such that

$$r(g(T)) = \phi(T_C).$$

Define the purified representation by

$$\bar{g}(T) := r(g(T)) = \phi(T_C).$$

Thus $\bar{g}$ is a deterministic function of g. At the representation level, we assume that the residual information in $g(T)$ beyond the recovered causal component contains no additional outcome information but does contain residual confounder information:

$$\textbf{(R1)} \qquad I(g(T); Y \mid \phi(T_C), Z) = 0,$$

and

$$\textbf{(R2)} \qquad I(g(T); Z \mid \phi(T_C)) > 0.$$

Utility preservation. Since $\bar{g} = \phi(T_C)$ is a deterministic function of g, the pair $(g, \bar{g})$ contains the same information as g. Hence

$$I(g; Y \mid Z) = I(g, \bar{g}; Y \mid Z).$$

By the chain rule for conditional mutual information,

$$I(g, \bar{g}; Y \mid Z) = I(\bar{g}; Y \mid Z) + I(g; Y \mid \bar{g}, Z).$$

Since $\bar{g} = \phi(T_C)$, the second term is

$$I(g; Y \mid \bar{g}, Z) = I(g; Y \mid \phi(T_C), Z),$$

which is zero by **R1**. Therefore

$$I(g; Y \mid Z) = I(\bar{g}; Y \mid Z).$$

Thus purification preserves the utility term.

Penalty reduction. Again, since $\bar{g}$ is a deterministic function of g,

$$I(g; Z) = I(g, \bar{g}; Z).$$

Applying the mutual-information chain rule,

$$I(g, \bar{g}; Z) = I(\bar{g}; Z) + I(g; Z \mid \bar{g}).$$

Using $\bar{g} = \phi(T_C)$, we obtain

$$I(g; Z) - I(\bar{g}; Z) = I(g; Z \mid \bar{g}) = I(g; Z \mid \phi(T_C)).$$

By **R2**, this quantity is strictly positive. Hence

$$I(\bar{g}; Z) < I(g; Z).$$

Objective comparison. Combining utility preservation and penalty reduction,

$$\begin{aligned} J(\bar{g}) - J(g) &= \big[I(\bar{g}; Y \mid Z) - I(g; Y \mid Z)\big] + \lambda\big[I(g; Z) - I(\bar{g}; Z)\big] \\ &= \lambda\, I(g; Z \mid \phi(T_C)) > 0. \end{aligned}$$

Therefore $J(\bar{g}) > J(g)$ for every $\lambda > 0$. □

General Predictive Purification Result

The purification result in the main text assumes that the causal component $\phi(T_C)$ is recoverable from the representation. We now show that a more general purification principle holds without this assumption.

Let

$$G := g(T)$$

be an arbitrary representation. We call $\bar{G} = s(G)$ a *predictive purification* of G if it is a deterministic function of G satisfying

$$Y \perp G \mid (\bar{G}, Z).$$

In words, once $\bar{G}$ and Z are known, the remaining information in G does not provide additional predictive information about Y.

Existence. Under standard regularity conditions, such a statistic exists. One canonical construction is the conditional response profile

$$s(G) := \big\{p(Y \mid G, Z = z) : z \in \mathcal{Z}\big\}.$$

This object records, for each value of G, the conditional law of Y across values of the confounder. By construction,

$$p(Y \mid G, Z) = p(Y \mid s(G), Z),$$

which implies

$$Y \perp G \mid (s(G), Z).$$

This statistic may be infinite-dimensional and is used only as a population-level object.

Proposition 4 (General predictive purification)**.** *Let $G = g(T)$ be any representation, and let $\bar{G} = s(G)$ be a predictive purification. Then, for any $\lambda > 0$,*

$$J(\bar{G}) \geq J(G),$$

with exact improvement

$$J(\bar{G}) - J(G) = \lambda I(G; Z \mid \bar{G}).$$

The inequality is strict whenever $I(G; Z \mid \bar{G}) > 0$.

Proof. Since $\bar{G} = s(G)$ is a deterministic function of G,

$$I(G; Y \mid Z) = I(G, \bar{G}; Y \mid Z).$$

By the chain rule,

$$I(G, \bar{G}; Y \mid Z) = I(\bar{G}; Y \mid Z) + I(G; Y \mid \bar{G}, Z).$$

The second term is zero by predictive sufficiency, so

$$I(G; Y \mid Z) = I(\bar{G}; Y \mid Z).$$

For the penalty,

$$I(G; Z) = I(G, \bar{G}; Z) = I(\bar{G}; Z) + I(G; Z \mid \bar{G}).$$

Thus

$$I(G; Z) - I(\bar{G}; Z) = I(G; Z \mid \bar{G}) \geq 0.$$

Combining,

$$J(\bar{G}) - J(G) = \lambda I(G; Z \mid \bar{G}) \geq 0.$$

□

Connection to the causal setting. The result above is purely information-theoretic and applies to any representation. Its causal interpretation follows from assumption **A1**. For the full entangled latent representation

$$G = (T_C, T_{nC}),$$

we have

$$p(Y \mid T_C, T_{nC}, Z) = p(Y \mid T_C, Z),$$

so the predictive purification depends only on the causal component T_C, up to outcome-equivalence. Thus, in the SCM considered in this work, the population target of purification can be written as

$$g(T) = \phi(T_C).$$

Relation to the main result. Proposition 4 shows that the disentanglement criterion removes information that is irrelevant for predicting $Y \mid Z$ but still informative about the confounder, without requiring any structural assumptions on g. The main-text result strengthens this conclusion by imposing a recoverability assumption, which allows the purified representation to be identified explicitly as a function of the causal component T_C.

B Reparameterisation of the Disentanglement Criterion

The objective in (1),

$$J(g) = I(g(T); Y \mid Z) - \lambda I(g(T); Z),$$

uses the unbounded penalty weight $\lambda \in [0, \infty)$. It is often more convenient to work with an equivalent bounded parameter $\gamma \in [0, 1]$.

Reparameterisation. For $\lambda > 0$, maximising $J(g)$ is equivalent to minimising

$$-J(g) = \lambda I(g(T); Z) - I(g(T); Y \mid Z).$$

Dividing by the positive constant $1 + \lambda$ does not change the minimiser, so the same solutions are obtained by minimising

$$\frac{\lambda}{1+\lambda} I(g(T); Z) - \frac{1}{1+\lambda} I(g(T); Y \mid Z).$$

Define

$$\gamma := \frac{1}{1+\lambda} \in (0, 1),$$

so that $1 - \gamma = \lambda/(1 + \lambda)$. The equivalent bounded objective is

$$L_\gamma(g) := (1-\gamma) I(g(T); Z) - \gamma I(g(T); Y \mid Z), \tag{13}$$

which we minimise over g. The limiting case $\lambda = 0$ corresponds to $\gamma = 1$, i.e. pure utility maximisation. Conversely, $\lambda \to \infty$ corresponds to $\gamma \to 0$, i.e. pure penalty minimisation.

Optimality condition under γ. Using the same notation as Proposition 3, compare the uncompressed representation T_C with a lossy alternative $\phi(T_C)$. The uncompressed representation is preferred under L_γ when

$$L_\gamma(T_C) < L_\gamma(\phi(T_C)).$$

Expanding this inequality gives

$$(1-\gamma)\Delta P(\phi) - \gamma \Delta U(\phi) < 0,$$

or equivalently

$$\gamma > \frac{\Delta P(\phi)}{\Delta U(\phi) + \Delta P(\phi)} =: \Gamma(\phi).$$

Thus $\Gamma(\phi)$ is the critical utility weight above which the full causal component is preferred to the lossy compression ϕ.

Proposition 5 (High-utility regime under γ)**.** *Let $\mathcal{H}$, $\Delta U(\phi)$, and $\Delta P(\phi)$ be as in Proposition 3, and assume the no-free-lunch condition $\Delta U(\phi) > 0$ and $\Delta P(\phi) > 0$ for every $\phi \in \mathcal{H}$. Define*

$$\gamma^\star := \sup_{\phi \in \mathcal{H}} \Gamma(\phi) = \sup_{\phi \in \mathcal{H}} \frac{\Delta P(\phi)}{\Delta U(\phi) + \Delta P(\phi)}.$$

Then for any $\gamma \in (\gamma^\star, 1)$, the uncompressed representation T_C strictly dominates every lossy alternative $\phi(T_C)$, $\phi \in \mathcal{H}$, under the objective L_γ.

Proof. Fix any $\phi \in \mathcal{H}$. The condition $L_\gamma(T_C) < L_\gamma(\phi(T_C))$ holds if and only if $\gamma > \Gamma(\phi)$. By definition of the supremum,

$$\Gamma(\phi) \le \gamma^\star \qquad \forall\, \phi \in \mathcal{H}.$$

Hence, for any $\gamma > \gamma^\star$,

$$\gamma > \Gamma(\phi) \qquad \forall\, \phi \in \mathcal{H}.$$

Therefore $L_\gamma(T_C) < L_\gamma(\phi(T_C))$ uniformly over $\mathcal{H}$, so the uncompressed causal representation strictly dominates all lossy alternatives. □ □

Correspondence with Proposition 3. The λ- and γ-parameterisations describe the same trade-off. Since

$$\gamma = \frac{1}{1+\lambda},$$

the condition

$$\lambda < \Lambda(\phi) = \frac{\Delta U(\phi)}{\Delta P(\phi)}$$

is equivalent to

$$\gamma > \frac{\Delta P(\phi)}{\Delta U(\phi) + \Delta P(\phi)} = \Gamma(\phi).$$

Thus the low-λ lossless regime in Proposition 3 is the same as the high-γ regime in Proposition 5.

C Toy Example

We give a closed-form sanity check for the Disentanglement Criterion:

$$J(g) = I(g(T); Y \mid Z) - \lambda I(g(T); Z)$$

in a linear additive Gaussian model. The treatment consists of causal and non-causal components, but only the causal ones affects the outcome. Consider the following SCM:

$$Z = \epsilon_Z, \tag{14}$$
$$T_C = \alpha Z + \epsilon_C, \tag{15}$$
$$T_{nC} = \beta Z + \epsilon_N, \tag{16}$$
$$Y = \rho T_C + \delta Z + \epsilon_Y, \tag{17}$$

where all noise variables are mutually independent and Gaussian:

$$\epsilon_Z \sim \mathcal{N}(0, \sigma_Z^2), \quad \epsilon_C \sim \mathcal{N}(0, \sigma_C^2), \quad \epsilon_N \sim \mathcal{N}(0, \sigma_N^2), \quad \epsilon_Y \sim \mathcal{N}(0, \sigma_Y^2).$$

The observed treatment is

$$T = (T_C, T_{nC}).$$

By construction, the outcome depends on the treatment only through the causal component:

$$Y \perp T_{nC} \mid (T_C, Z).$$

Moreover, under non-degenerate parameters, e.g. $\beta \neq 0$, $\sigma_Z^2 > 0$, $\sigma_C^2 > 0$, and $\sigma_N^2 > 0$, the non-causal component contains residual information about the confounder beyond the causal component:

$$I(T_{nC}; Z \mid T_C) > 0.$$

Indeed, since the variables are jointly Gaussian,

$$I(T_{nC}; Z \mid T_C) = \frac{1}{2} \log \frac{\mathrm{Var}(T_{nC} \mid T_C)}{\mathrm{Var}(T_{nC} \mid Z, T_C)}.$$

Now

$$\mathrm{Var}(T_{nC} \mid Z, T_C) = \sigma_N^2,$$

whereas

$$\mathrm{Var}(T_{nC} \mid T_C) = \beta^2 \mathrm{Var}(Z \mid T_C) + \sigma_N^2,$$

with

$$\mathrm{Var}(Z \mid T_C) = \frac{\sigma_Z^2 \sigma_C^2}{\alpha^2 \sigma_Z^2 + \sigma_C^2}.$$

Therefore,

$$I(T_{nC}; Z \mid T_C) = \frac{1}{2} \log \left(1 + \frac{\beta^2 \sigma_Z^2 \sigma_C^2}{\sigma_N^2 (\alpha^2 \sigma_Z^2 + \sigma_C^2)} \right) > 0.$$

Thus the Gaussian SCM satisfies assumptions **A1** and **A2** from the main text. Consequently, we now compare the ideal causal representation with the naive entangled representation:

$$g_1(T) = T_C, \quad g_2(T) = T = (T_C, T_{nC}).$$

For scalar jointly Gaussian variables,

$$I(U; V \mid W) = \frac{1}{2} \log \frac{\mathrm{Var}(V \mid W)}{\mathrm{Var}(V \mid U, W)}.$$

Conditioned on Z,

$$Y = (\rho\alpha + \delta) Z + \rho \epsilon_C + \epsilon_Y,$$

so

$$\mathrm{Var}(Y \mid Z) = \rho^2 \sigma_C^2 + \sigma_Y^2.$$

Conditioned on (T_C, Z),

$$Y = \rho T_C + \delta Z + \epsilon_Y,$$

and therefore

$$\mathrm{Var}(Y \mid T_C, Z) = \sigma_Y^2.$$

Hence

$$I(T_C; Y \mid Z) = \frac{1}{2} \log \left(\frac{\rho^2 \sigma_C^2 + \sigma_Y^2}{\sigma_Y^2} \right) \tag{18}$$

$$= \frac{1}{2} \log \left(1 + \frac{\rho^2 \sigma_C^2}{\sigma_Y^2} \right). \tag{19}$$

Since $Y \perp T_{nC} \mid (T_C, Z) \implies I(T_C, T_{nC}; Y \mid Z) = I(T_C; Y \mid Z)$. So the ideal and naive representations have identical utility.

For the penalty term,

$$I(T_C; Z) = \frac{1}{2} \log \left(\frac{\alpha^2 \sigma_Z^2 + \sigma_C^2}{\sigma_C^2} \right) = \frac{1}{2} \log \left(1 + \frac{\alpha^2 \sigma_Z^2}{\sigma_C^2} \right).$$

By the mutual-information chain rule,

$$I(T_C, T_{nC}; Z) = I(T_C; Z) + I(T_{nC}; Z \mid T_C).$$

The second term is strictly positive by the calculation above, so

$$I(g_2(T); Z) > I(g_1(T); Z).$$

Therefore, for any $\lambda > 0$,

$$J(g_1) - J(g_2) = \lambda \big[I(g_2(T); Z) - I(g_1(T); Z) \big] \tag{20}$$

$$= \lambda I(T_{nC}; Z \mid T_C) > 0. \tag{21}$$

Thus, in this Gaussian SCM, the criterion strictly prefers the causal representation T_C over the naive entangled representation (T_C, T_{nC}). Figure 4 shows the same effect empirically in a multivariate simulation where the representation progressively removes coordinates of T_{nC}. Retaining more non-causal information leaves the utility unchanged but increases the confounding penalty, so the corresponding objective values decrease more rapidly as λ grows.

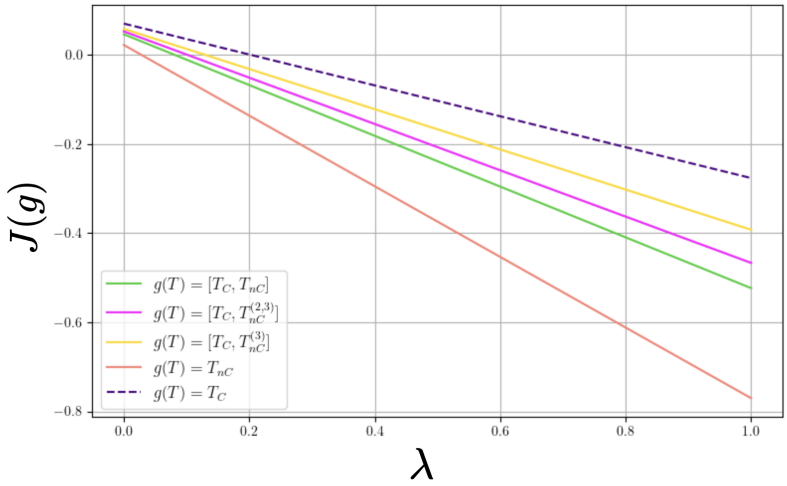

Figure 4: **Progressive removal of non-causal leakage.** The objective $J(g)$ is plotted as a function of the penalty weight λ for representations retaining different amounts of the non-causal component T_{nC}. Since T_{nC} contributes confounder information without adding outcome information beyond T_C, representations with more non-causal leakage incur a larger penalty. The purely causal representation $g(T) = T_C$ achieves the highest score among the tested representations.

After the non-causal component has been removed, the remaining question is whether the optimal representation should preserve all of T_C or compress it. To model lossy compression analytically, consider the stochastic representation channel

$$G = T_C + \epsilon_G, \qquad \epsilon_G \sim \mathcal{N}(0, \sigma_G^2),$$

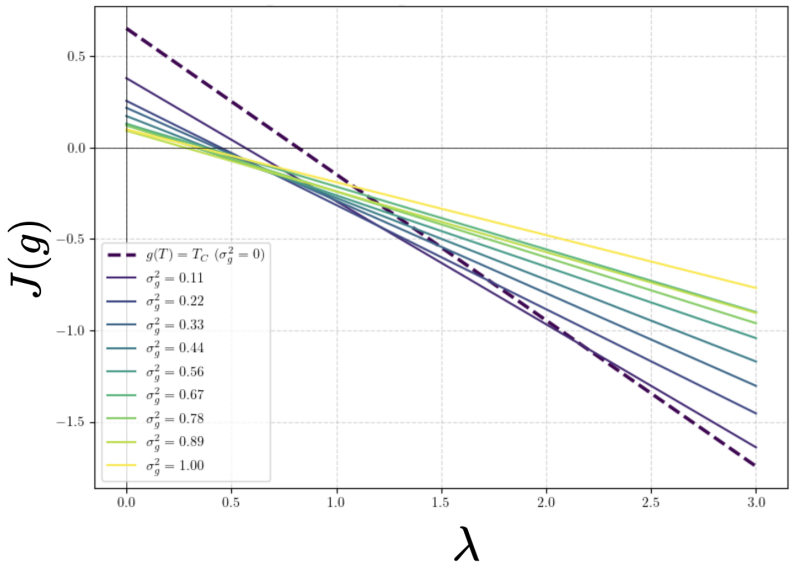

Figure 5: **Lossless regime and critical penalty weight.** The objective $J(g)$ is plotted as a function of λ for stochastic compressed representations $G = T_C + \epsilon_G$, where $\epsilon_G \sim \mathcal{N}(0, \sigma_G^2)$. The dashed curve corresponds to the uncompressed causal representation $\sigma_G^2 = 0$. For small λ, preserving the full causal component is optimal; for larger λ, compression can become preferable because the reduction in confounder information outweighs the utility loss.

with ϵ_G independent of all other variables. This can be viewed as a Gaussian Markov kernel from T_C to G, where σ_G^2 controls the amount of compression. The penalty is

$$P(\sigma_G^2) := I(G; Z) = \frac{1}{2} \log \left(1 + \frac{\alpha^2 \sigma_Z^2}{\sigma_C^2 + \sigma_G^2} \right),$$

and the utility is

$$U(\sigma_G^2) := I(G; Y \mid Z) = \frac{1}{2} \log \left(\frac{\rho^2 \sigma_C^2 + \sigma_Y^2}{(\rho^2 \sigma_C^2 + \sigma_Y^2) - \frac{\rho^2 \sigma_C^4}{\sigma_C^2 + \sigma_G^2}} \right).$$

Both terms decrease as σ_G^2 increases: compression removes information about the confounder, but it also removes predictive information about the outcome. The objective is

$$J(\sigma_G^2) = U(\sigma_G^2) - \lambda P(\sigma_G^2).$$

At the uncompressed boundary $\sigma_G^2 = 0$,

$$\left. \frac{dP}{d\sigma_G^2} \right|_{\sigma_G^2 = 0} = -\frac{\alpha^2 \sigma_Z^2}{2\sigma_C^2 (\sigma_C^2 + \alpha^2 \sigma_Z^2)},$$

and

$$\left. \frac{dU}{d\sigma_G^2} \right|_{\sigma_G^2 = 0} = -\frac{\rho^2}{2\sigma_Y^2}.$$

Thus the uncompressed representation is locally optimal whenever

$$\left. \frac{dJ}{d\sigma_G^2} \right|_{\sigma_G^2 = 0} = \left. \frac{dU}{d\sigma_G^2} \right|_{\sigma_G^2 = 0} - \lambda \left. \frac{dP}{d\sigma_G^2} \right|_{\sigma_G^2 = 0} \leq 0.$$

Equivalently,

$$\lambda \leq \lambda_{\text{crit}} := \frac{\left| \frac{dU}{d\sigma_G^2} \right|_{\sigma_G^2 = 0}}{\left| \frac{dP}{d\sigma_G^2} \right|_{\sigma_G^2 = 0}}.$$

This threshold is the ratio between the marginal utility loss and the marginal penalty reduction at zero compression. Figure 5 visualises this trade-off: for small λ, preserving the full causal component is optimal; for larger λ, the penalty becomes sufficiently expensive that a noisy compressed representation can achieve a higher objective value. This is the Gaussian analogue of the lossless regime in Proposition 3.

D NCE-CLUB Upper Bound and Critic Interpretation

To construct a tractable surrogate for the disentanglement objective

$$J(g) = I(g(T); Y \mid Z) - \lambda I(g(T); Z),$$

we require an upper bound on the penalty term $I(g(T); Z)$.

General NCE-CLUB bound. We use the NCE-CLUB estimator Liang et al. [2023]. For a pair of random variables (X_1, X_2), NCE-CLUB provides the upper bound

$$I(X_1; X_2) \leq I_{\text{NCE-CLUB}}(X_1; X_2),$$

where, for the optimal critic $f^\star$,

$$I_{\text{NCE-CLUB}}(X_1; X_2) := \mathbb{E}_{(x_1, x_2^+) \sim p(x_1, x_2)}\big[f^\star(x_1, x_2^+)\big] - \mathbb{E}_{x_1 \sim p(x_1),\, x_2^- \sim p(x_2)}\big[f^\star(x_1, x_2^-)\big].$$

The first expectation is over joint, or positive, pairs; the second is over independently sampled, or negative, pairs.

Application to the penalty term. Applying this construction to the penalty with

$$X_1 = g(T), \qquad X_2 = Z,$$

gives

$$I(g(T); Z) \leq I_{\text{NCE-CLUB}}(g(T); Z).$$

In practice, the optimal critic $f^\star$ is unknown, so we use a learned critic d_ϕ and obtain the plug-in estimator

$$\begin{aligned} I_{\text{NCE-CLUB}}(g(T); Z; d_\phi) := \mathbb{E}_{p(g(T), z)}\big[d_\phi(g(T), z)\big] \\ - \mathbb{E}_{p(g(T))p(z)}\big[d_\phi(g(T), z)\big]. \end{aligned}$$

This is the quantity used as the penalty estimator in the main text.

Interpretation of the critic d_ϕ. The optimal critic corresponds to the log-density ratio

$$d^\star(g, z) = \log \frac{p(z \mid g)}{p(z)} = \log p(z \mid g) - \log p(z),$$

which is intractable because the true conditional $p(z \mid g)$ is unknown. We therefore approximate it with a learnable critic $d_\phi(g, z)$, which can be interpreted as a variational estimate

$$d_\phi(g_\theta(T), z) \approx \log q_\phi(z \mid g_\theta(T)) - \log p(z),$$

where $q_\phi(z \mid g)$ is a learned conditional model.

The critic is trained to distinguish positive pairs

$$(g, z) \sim p(g, z)$$

from negative pairs

$$(g, z) \sim p(g)p(z).$$

This training aims to move d_ϕ toward the optimal log-ratio and thereby tighten the plug-in estimate.

Adversarial training dynamics. The optimisation alternates between critic fitting and encoder updates. The critic parameters ϕ are trained to maximise the InfoNCE objective, approximating the optimal critic for the current representation. Then, holding the critic fixed, the encoder parameters θ are trained to minimise the NCE-CLUB penalty together with the utility loss. This encourages representations that remain predictive of Y while discouraging dependence on Z. The gradient-reversed InfoNCE implementation used in our experiments is described in Section 2.4.

E Synthetic SCM: experimental details

We generate 1,000 training samples, and 1,000 validation samples from the following linear additive SCM with $d = 10$, $d_c = 5$:

$$\begin{aligned} Z &\sim \mathcal{N}(0, I_d), \\ T_C &= Z_{0:d_c} + \epsilon_{T_C}, & \epsilon_{T_C} &\sim \mathcal{N}(0, I_{d_c}), \\ T_{nC} &= Z_{d_c:d} + \epsilon_{T_{nC}}, & \epsilon_{T_{nC}} &\sim \mathcal{N}(0, I_{d-d_c}), \\ T &= [T_C, T_{nC}], \\ Y &= \mathbf{1}^\top T_C + \mathbf{1}^\top Z + \epsilon_Y, & \epsilon_{Y,i} &\overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_Y^2). \end{aligned}$$

The predictor is a linear model $\hat{Y} = w_Z^\top Z + w_T^\top T + b$. We take the treatment representation to be the per-coordinate contribution of T to the prediction, $g_T = T \odot w_T \in \mathbb{R}^{10}$, where w_T is the treatment slice of the predictor's own weight vector. The MI critic therefore decorrelates Z from the components of T that the predictor actually uses. Under gradient reversal this pushes w_T toward the causal coordinates of T. The critic itself is bilinear: two bias-free linear projections $W_g, W_z \in \mathbb{R}^{10\times 10}$ to a shared embedding space, with InfoNCE scores given by the full $B \times B$ matrix $S = (W_g g_T)(W_z Z)^\top$ over a minibatch of size B. We train with lr 3×10^{-4}, batch size 64, 4,000 epochs, and GRL schedule with $\lambda = 0$ for epochs 0–200, cubic ramp $0 \to \lambda_{\max} = 0.5$ over epochs 200–1000, held constant thereafter.

F Semi-Synthetic KuaiRand: experimental details

We use the KuaiRand-Pure Gao et al. [2022] (Zenodo 10439422) dataset with 27,285 users, 7,583 items, and ∼1.186M distinct user-video pairs from the randomized-exposure log. KuaiRand provides 30 user features, of which we keep 24 including four counts, two binary indicators, and 18 anonymized category codes. We drop the four bucketed versions of the counts, the categorical activity label, and one binary indicator that is constant in KuaiRand-Pure. For simplicity when generating synthetic outcomes we treat each anonymized category code as numerical. KuaiRand also provides 62 video features, of which we keep 54 including duration, width, height, and 51 video statistics. We drop identifiers, categorical features, upload date, and constant features. We log transform counts, duration, video statistics, and then standardize them. We standardize the numerical user features.

Let Z denote the 24 user features and T_C the 54 video features. We generate a synthetic click outcome from a user term $h_u(Z)$, video term $h_v(T_C)$, and user-video interaction term $h_{uv}(Z, T_C)$ as

$$Y \sim \text{Bernoulli}\left(\text{sigmoid}\left(\text{std}\left(4\, h_u(Z) + h_v(T_C) + h_{uv}(Z, T_C)\right) + b\right)\right)$$

where std corresponds to centering and scaling. The intercept b calibrates the mean click probability to the actual click rate in the randomized log. We define the user, video, and user-video interaction terms as follows. Let $q_1, q_2 \in \mathbb{R}^{24}$, $q_3 \in \mathbb{R}^{54}$, $Q_1 \in \mathbb{R}^{24\times 16}$, and $Q_2 \in \mathbb{R}^{54\times 16}$ be fixed random projections. We define

$$\begin{aligned} h_u(Z) &= \text{std}\left(\tanh\big(\text{std}(Z^\top q_1)\big)\tanh\big(\text{std}(Z^\top q_2)\big)\right), \\ h_v(T_C) &= \text{std}(T_C^\top q_3), \\ h_{uv}(Z, T_C) &= \text{std}\left(\frac{1}{16}\sum_{j=1}^{16}\tanh\big((Z^\top Q_1)_j\big)\tanh\big((T_C^\top Q_2)_j\big)\right). \end{aligned}$$

The entries of each projection are independent zero-mean Gaussian samples with variance equal to the inverse of its input dimension. To be clear, the data generation process above is not intended to be a realistic model of user behaviour. It provides a nonlinear outcome mechanism with known causal structure built from real KuaiRand features.

To test our method, we introduce a non-causal treatment component $T_{nC} \in \{1, \ldots, 16\}$, which does not impact the outcome, but is related to Z and can be exploited by the model when making predictions. We append this non-causal shortcut feature to the observed treatment input $T = (T_C, T_{nC})$. In the

source distribution, T_{nC} is strongly related to the user response term. Let $Q(Z)$ be the quantile bin of $h_u(Z)$, then

$$p(T_{nC} = t \mid Z) = \frac{\exp\left(\alpha \mathbf{1}\left[t = Q(Z)\right]\right)}{\exp(\alpha) + 15}.$$

The parameter α controls the shortcut strength, and we choose α so that the mutual information is $I(T_{nC}; Q \mid V) = 2$ nats, where V is the video ID. For the target distribution, we instead set $\alpha = 0$, sampling the non-causal feature uniformly and removing the Z shortcut. Source and target test sets contain the same users, videos, features, and synthetic clicks, only T_{nC} changes. We encode the code as 16 one-hot values scaled by 3, so the treatment encoder receives $54 + 16 = 70$ inputs.

We train on 70% of the data, use 15% as our validation set for early stopping, and the remaining 15% as the test set for calculating source and target AUC and ranking metrics. Our model is a DeepFM Guo et al. [2017] backbone with 16-dimensional learned user-ID and video-ID embeddings. The treatment encoder maps $70 \to 32 \to 16$ with a ReLU hidden layer. Appending the video-ID embedding gives a 32-dimensional treatment representation. Appending the user-ID embedding to the 24 user features gives a 40-dimensional user representation. The DeepFM network has hidden sizes $[32, 16]$, dropout 0.2, and FM embedding dimension 8. The CRL variant uses the same predictor and adds an InfoNCE critic. The predictor architecture and capacity are identical across configurations. Each critic input is mapped through a two-layer projection with widths 64 and 32, followed by a dot-product score. Gradient reversal acts on the treatment representation. The user representation is detached from the critic loss. The reversal weight λ is zero for 25 steps, rises linearly to 0.3 over the next 50 steps, and then remains fixed. We train for four epochs with Adam, learning rate 2×10^{-3}, weight decay 10^{-6}, batch size $4{,}096$, and gradient-norm clipping at 1. We evaluate source-validation AUC every 50 steps and retain the best checkpoint.

When applying DCCL Zhao et al. [2023] to our KuaiRand setup, we split each 16-dimensional user and video ID embedding equally into interest and conformity components, preserving the predictor's total capacity. Two popularity-weighted contrastive losses supplement the prediction loss, using clicked examples as anchors and up to 128 in-batch negatives. Both auxiliary losses use weight 0.01.

As well as source and target AUC, in Table 1 we report target Recall@5 and target NDCG@5 [see eg. Zhang et al., 2023, Deng et al., 2024, who report similar metrics on KuaiRand]. For each user in the test split we rank the items they were exposed to by predicted click probability. Recall@K is the fraction of that user's positives that appear in the top-K, and NDCG@K uses binary relevance with discount $1/\log_2(i+1)$ at rank i and IDCG computed over the user's true positives. Both are averaged over users with at least one positive and at least five exposed items.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and §1 state three contributions—an information-theoretic disentanglement criterion with a population-level analysis (§2, Appendix A), a tractable variational surrogate (§2), and an empirical chain consisting of a synthetic SCM, semi-synthetic KuaiRand, and an online A/B test (§4)—each of which corresponds to a section in the paper.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: §4.4 discusses the assumption that relevant confounding information is observed in Z, the inability of the penalty term to remove dependence transmitted only via unobserved confounders, and the implications of incomplete confounder data in production. §2.4 distinguishes the theoretical NCE-CLUB guarantee from the practical InfoNCE implementation and discusses the risk of removing potentially useful predictive information.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

   Justification: The structural assumptions (**A1**, **A2**) are stated in the Setting subsection of §2. Each numbered theoretical result in §2 (Propositions on the population-level optimum, the lossless regime, and the variational surrogate) is restated and proved in full in Appendix A, with a separate worked sanity check on a linear-Gaussian SCM in Appendix C. Proof sketches for the main propositions appear inline in §2.

   Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: For the synthetic SCM (Appendix E) we give the full details. For KuaiRand (Appendix F) we give the dataset version (Zenodo `10439422`), and details. The online A/B test uses proprietary production data and a deployed ranker. The underlying data and model cannot be shared (Section 4.3).

   Guidelines:

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

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

   Answer: [No]

   Justification: We describe the complete data generation process for the synthetic example in Appendix E. KuaiRand is a public dataset (Zenodo record `10439422`), cited in Appendix F. The online A/B test uses proprietary production data and a deployed ranker, neither of which can be released.

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: §4.1 and §4.2 describe the experimental setup at the level needed to interpret the results, with details in Appendices E and F.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

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

Answer: [Yes]

Justification: The synthetic experiment (§4.1) reports medians with interquartile-range bands across 2,000 random seeds; variability is over training seed and data resampling. The semi-synthetic KuaiRand experiment (§4.2) reports AUC means with 95% confidence intervals across 100 data-generation seeds. Ranking metrics report target means. The online A/B test (§4.3) reports point estimates with 95% confidence intervals computed by the platform's standard online-experimentation infrastructure (variance is over user-level outcomes within the randomised treatment/control assignment). Error-bar conventions are stated where they appear.

Guidelines:

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

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: The synthetic and KuaiRand experiments were run on a laptop with 10 CPU cores and 64GB memory. The model for the online A/B test was trained on production infrastructure.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification:

Guidelines:

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

10. **Broader impacts**

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

    Answer: [N/A]

    Justification: The authors present a theoretical result in the field of causal representation learning, and apply this methodology to a commercial music streaming platform.

    Guidelines:

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

11. **Safeguards**

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

    Answer: [N/A]

    Justification: The paper releases neither pre-trained generative models nor scraped datasets.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: KuaiRand-Pure Gao et al. [2022] is cited with its dataset version (Zenodo 10439422) in Appendix F and is used in accordance with the terms attached to that record.

    Guidelines:

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

13. **New assets**

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

    Answer: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

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

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

    Answer: [N/A]

    Justification:

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification:

    Guidelines:

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

16. **Declaration of LLM usage**

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

    Answer: [N/A]

    Justification: LLMs are not part of the core methodology, the disentanglement criterion, the variational surrogate, or any of the experimental pipelines.

    Guidelines:

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