DISC-INFOMAX: SELF-SUPERVISED LEARNING OF INVARIANT CAUSAL REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

InfoMax self-supervised learning (SSL) has driven significant advances in representation learning, but standard approaches often entangle true causal variables with spurious correlations, leading to catastrophic failures under distribution shifts. While multi-view causal representation learning offers theoretical guarantees for identifying true causal variables, practical implementations in SSL settings are bottlenecked by their reliance on explicit view labels and highly non-convex optimization. To bridge this gap, we address the broader challenge of causal representation learning without explicit view labels, which includes the InfoMax SSL setting. We make two primary contributions. First, we introduce a novel training objective, Disc-InfoMax, which dynamically infers a shared content mask from the empirical differences between paired representations. We prove that under asymptotic convergence, the optimal encoder is theoretically guaranteed to block-identify true underlying causal variables. Second, we propose BayesDisc as a practical algorithm that effectively mitigates non-convexity, only requiring standard deep learning optimizers. Finally, we demonstrate on several datasets that BayesDisc successfully improves causal variable identification over InfoMax SSL baselines.

1 INTRODUCTION

InfoMax self-supervised learning (SSL) builds robust representations by maximizing mutual information between the predicted latent vectors of paired observations of the data:

$$\max_{\theta} \mathcal{I}(g_\theta(\mathbf{X}^{(1)}), g_\theta(\mathbf{X}^{(2)})), \tag{1}$$

where $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ share some invariant traits. This objective encourages the encoder g_θ to extract information shared between the pair, effectively maximizing a lower bound on the mutual information between the observations and the predicted latent vectors (Tschannen et al., 2019; Oord et al., 2019).

However, there is a fundamental catch: standard SSL compresses all of this information into a single, dense latent vector. The information of the underlying true causal variables is entangled, making individual components of the vector uninterpretable. Consequently, this entangled monolith prevents the use of downstream causal algorithms, directly leading to a critical vulnerability in modern deep learning: a lack of out-of-distribution (OOD) generalization. For instance, a vision model trained to detect traffic lights in one country might secretly rely on local background scenery or specific pole designs. When deployed in another country, these spurious features shift, and prediction performance plummets. To achieve true OOD robustness, a model must isolate the underlying causal variables, as these represent the fundamental, invariant mechanisms that hold across different environments.

Causal Representation Learning (CRL) offers a principled path forward to both interpretability and OOD robustness by seeking to identify the true, underlying causal variables of a dataset (Schölkopf et al., 2021). Since fully unsupervised CRL is impossible without inductive biases (Locatello et al., 2019), multi-view setups have emerged as a highly promising framework and the closest theoretical counterpart to InfoMax SSL. Theoretical guarantees demonstrate that information shared across the views can be block-identified (Yao et al., 2023). Unfortunately, a fundamental barrier prevents the application of this theory to practical SSL. To understand this gap, we must clarify the terminology: a "view" (or "group" in sparsity-based CRL (Xu et al., 2024)) is defined by a partial observability pattern capturing only a subset of true causal variables for an observation. Importantly, multi-view CRL relies

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

on explicit *view labels*, which serve as the structural indices to identify exactly which view a sample originates from. Because manually grouping complex, augmented data into identical underlying observability patterns to acquire these labels is often challenging, this framework is difficult to deploy in scalable self-supervised learning pipelines. As a result, prior multi-view methods are fundamentally incompatible with the standard InfoMax SSL setting, which relies on a single shared encoder processing uncalibrated pairs with unlabeled, constantly shifting observability patterns.

To overcome this gap, we abandon the reliance on explicit view labels and instead infer the shared content mask. As a result, our approach seamlessly extends the theoretical guarantees of Causal Representation Learning (CRL) to naturally encompass the InfoMax SSL setting.

Specifically, our main contributions are as follows:

- **Theoretical Identifiability without View Labels:** We formulate a Disc-InfoMax objective and formally prove that, at the asymptotic limit of the proximity mask, optimizing the encoder guarantees the block-identifiability of true underlying causal variables without view labels.
- **The BayesDisc Algorithm:** Because practical optimization of Disc-InfoMax is highly non-convex, we propose BayesDisc. By dynamically inferring analytic Bayesian posteriors based on empirical latent residuals with bandwidth annealing, BayesDisc acts as a tractable surrogate that mitigates this non-convexity, enabling stable training with standard deep learning optimizers.
- **Empirical Validation:** We rigorously evaluate our approach on both structurally entangled synthetic numerical benchmarks and the complex, image-based Causal3DIdent dataset (von Kügelgen et al., 2021). Our results demonstrate that BayesDisc outperforms standard InfoMax SSL baselines (InfoNCE (Oord et al., 2019), LeJEPA (Balestriero & LeCun, 2025)) in identifying causal variables while retaining equal or greater information content.

2 Problem Formulation: The Paired Partially Observed Setting

In this section, we formalize the *Paired Partially Observed* setting, which covers the InfoMax SSL setting. This framework describes data-generating processes where paired observations are formed by capturing varying, instance-dependent subsets of true underlying causal variables.

Our generative model is defined by four main components: the true causal variables $\mathbf{C}$, the random mask vectors representing paired observability patterns $(\mathbf{M}^{(1)}, \mathbf{M}^{(2)})$, the resulting masked latent vectors $(\mathbf{Z}^{(1)}, \mathbf{Z}^{(2)})$, and the empirical paired observations $(\mathbf{X}^{(1)}, \mathbf{X}^{(2)})$. A comprehensive summary of mathematical notation, variables, and operators used throughout this work is provided in App. A.

True Causal Variables ($\mathbf{C}$) We define the underlying true causal variables as a continuous random vector $\mathbf{C} = (C_1, \dots, C_n)^\top$ taking values in an open, simply connected latent space $\mathcal{C} \subseteq \mathbb{R}^n$. We assume $\mathbf{C}$ follows a joint probability density $p_{\mathbf{C}}$ such that $p_{\mathbf{C}} > 0$ almost everywhere on $\mathcal{C}$. Rather than assuming a generic prior, we strictly assume $\mathbf{C}$ is governed by a Structural Causal Model (SCM) associated with a Directed Acyclic Graph (DAG), generated via nonlinear mechanisms.

Paired Observability Patterns ($\mathbf{M}^{(1)}, \mathbf{M}^{(2)}$) To model the partial observation process, we introduce a pair of random binary mask vectors $\mathbf{M}^{(i)} \in \mathcal{M} \subseteq \{0,1\}^n$ for $i \in \{1,2\}$. Each mask dictates the specific observability pattern (i.e., the "view") for that instance. The mask variables are independent of the causal variables ($\mathbf{M}^{(i)} \perp\!\!\!\perp \mathbf{C}$). Furthermore, the joint sampling distribution $p(\mathbf{M}^{(1)}, \mathbf{M}^{(2)})$ is constrained such that the two masks always possess a non-empty intersection, mathematically mirroring the content overlap guaranteed by semantic data augmentations. We denote the shared content mask by $\mathbf{S} \coloneqq \mathbf{M}^{(1)} \odot \mathbf{M}^{(2)} \in \{0,1\}^n$, with non-empty intersection implying $\|\mathbf{S}\|_0 \geq 1$. We strictly assume that the marginal distributions of the masks are identical (i.e., $p(\mathbf{M}^{(1)}) = p(\mathbf{M}^{(2)})$).

Masked Latent Vectors ($\mathbf{Z}^{(1)}, \mathbf{Z}^{(2)}$) The masked latent vectors $\mathbf{Z}^{(i)} \in \mathcal{Z} \subseteq \mathbb{R}^n$ represent the latent information preserved in each specific instance $i \in \{1,2\}$. They are formulated as a masked sum of the true causal variables and a random fill variable:

$$\mathbf{Z}^{(i)} = \mathbf{M}^{(i)} \odot \mathbf{C} + (\mathbf{1} - \mathbf{M}^{(i)}) \odot \mathbf{N}^{(i)}, \quad i \in \{1,2\}. \tag{2}$$

For any component $j \in \{1, \dots, n\}$, if $M_j^{(i)} = 1$, the true causal feature C_j is perfectly captured. If $M_j^{(i)} = 0$, the feature is occluded and replaced by $N_j^{(i)}$. Here, $\mathbf{N}^{(i)}$ represents uninformative nuisance noise drawn independently from a continuous distribution with positive density on an open support (e.g., continuous environmental noise or independent marginal resampling of $\mathbf{C}$), ensuring that the masked latent $\mathbf{Z}$ possesses a regular continuous density on $\mathcal{Z}$.[1]

Paired Observations ($\mathbf{X}^{(1)}, \mathbf{X}^{(2)}$) Finally, the empirical observations $\mathbf{X}^{(i)}$ taking values in the observation space $\mathcal{X} \subseteq \mathbb{R}^d$ are generated by applying a smooth, invertible mixing function (diffeomorphism onto its image) $f : \mathbb{R}^n \to \mathbb{R}^d$ to the masked latents:

$$\mathbf{X}^{(i)} = f(\mathbf{Z}^{(i)}), \quad i \in \{1, 2\}. \tag{3}$$

Since the marginal distributions of the masks are identical, both observations $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ share the identical marginal distribution $p(\mathbf{X}^{(1)}) = p(\mathbf{X}^{(2)})$.

Learning Setup and Block-Identifiability We assume the encoder g_θ belongs to a hypothesis class $\mathcal{G}$ of diffeomorphisms onto their range $\hat{\mathcal{Z}} := g_\theta(\mathcal{X}) \subseteq \mathbb{R}^m$, parameterized over a compact parameter space $\Theta \subset \mathbb{R}^p$ (standardly enforced via weight decay). For our theoretical identifiability guarantees, we analyze the well-specified setting where the representation dimension matches the true causal latent dimension ($m = n$). Throughout, a binary mask subscript (e.g., $\mathbf{C}_K \in \mathcal{C}_K$ or $\hat{\mathbf{Z}}_{K'} \in \hat{\mathcal{Z}}_{K'}$) denotes the subvector or coordinate projection restricted to its active coordinates, distinct from the dimension-preserving Hadamard product $\mathbf{m} \odot \mathbf{x}$ (excluding parameter subscripts like θ in g_θ). Our goal is to learn $g_\theta \in \mathcal{G}$ that identifies true causal variables $\mathbf{C}$ from paired observations $(\mathbf{X}^{(1)}, \mathbf{X}^{(2)})$. We formalize this by adapting block-identifiability (von Kügelgen et al., 2021) to our setting:

Definition 1 ***(Block-Identifiability).*** For any binary mask $K \in \{0, 1\}^n$, the causal subvector $\mathbf{C}_K$ is *block-identified* by g_θ if there exists a binary latent mask $K' \in \{0, 1\}^m$ with $\|K'\|_0 = \|K\|_0$ and a diffeomorphism $\psi : \mathcal{C}_K \to \hat{\mathcal{Z}}_{K'}$ such that $\hat{\mathbf{Z}}_{K'} = \psi(\mathbf{C}_K)$.

3 Discernment Mechanism for Identifiability

Standard self-supervised InfoMax frameworks train representations by balancing two opposing forces: an *alignment* objective that encourages representations across paired views to match, and an *entropy maximization* (or information maximization) objective that prevents representational collapse. In the paired partially observed setting, however, global alignment is fundamentally flawed: paired observations $(\mathbf{X}^{(1)}, \mathbf{X}^{(2)})$ share only an instance-dependent subset of causal variables governed by the latent shared content mask $\mathbf{S} := \mathbf{M}^{(1)} \odot \mathbf{M}^{(2)}$ (isolating $\mathbf{C}_S$). Standard InfoMax objectives blindly push *all* latent dimensions to be identical, forcing unshared coordinates to align and mixing the variables. To resolve this, our key insight is to introduce a latent mask into the alignment objective, allowing dimensions that are *not* supposed to be the same to remain different without incurring an optimization penalty. Because this selective alignment focuses exclusively on shared coordinates to extract invariant causal information while letting unshared variations fluctuate, we call it the *Discernment* term (replacing *Alignment* term), leading to our proposed *Disc-InfoMax* objective. This discernment mechanism directly unlocks our theoretical identifiability guarantees. In what follows, Sec. 3.1 provides the theoretical formulation and identifiability proof sketch for Disc-InfoMax, and Sec. 3.2 introduces our tractable optimization algorithm, BayesDisc.

3.1 Theoretical Foundation: The Disc-InfoMax Objective

Core Question Given paired observations $(\mathbf{X}^{(1)}, \mathbf{X}^{(2)})$, an encoder g_θ outputs representations $\hat{\mathbf{Z}}^{(i)} = g_\theta(\mathbf{X}^{(i)})$ with residual vector $\mathbf{\Delta} := \hat{\mathbf{Z}}^{(1)} - \hat{\mathbf{Z}}^{(2)}$. Since the true shared content mask $\mathbf{S}$ is strictly unobserved, the central theoretical question is: *how can an objective dynamically isolate which coordinates to align without access to oracle view labels?*

[1] Continuous noise $\mathbf{N}$ with positive density is assumed in our theoretical identifiability analysis to ensure almost sure non-zero residual fluctuations on unobserved coordinates; in empirical implementations, $\mathbf{N}$ may also be realized via standard discrete zero-masking or mean imputation without affecting practical optimization.

To answer this core question, we first introduce *proximity mask*: when coordinate differences are small, we treat them as shared content to be aligned; conversely, when differences are large, we treat them as unshared variations that should remain unaligned. We formally define this mechanism as follows:

Definition 2 *(Proximity Mask).* Let $\sigma > 0$ denote a bandwidth parameter. A proximity mask $\mathbf{\Phi}^{(\sigma)} : \mathbb{R}^m \to [0,1]^m$ operates element-wise on the residual $\mathbf{\Delta}$, where each coordinate function $\Phi_j^{(\sigma)} : \mathbb{R} \to [0,1]$ is continuous on $\mathbb{R}$ and satisfies:

1. **Peak Alignment:** $\Phi_j^{(\sigma)}(0) = 1$ for any $\sigma > 0$. The gate fully activates when representations perfectly align.
2. **Monotonic Decay:** $\Phi_j^{(\sigma)}(\delta)$ is strictly monotonically decreasing with respect to $|\delta|$ on $[0, \infty)$.
3. **Asymptotic Nullification:** $\lim_{\sigma \to 0^+} \Phi_j^{(\sigma)}(\delta) = 0$ for all $\delta \neq 0$. As the bandwidth becomes infinitely narrow, the gate vanishes on all unaligned coordinates.

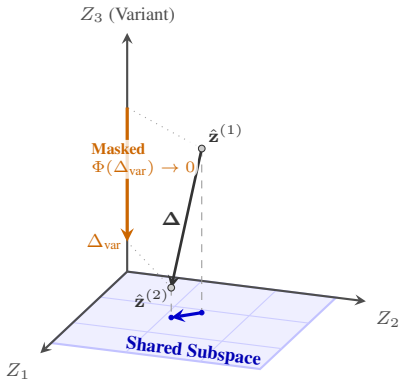


Figure 1: 3D geometric interpretation of the discernment term. The representation residual $\mathbf{\Delta}$ is decomposed into a shared subspace and a variant subspace. In the **Shared Subspace**, $\Delta \approx 0$ drives the proximity mask toward 1, whereas in the **Variant Subspace**, large residuals from instance-specific noise softly force the proximity mask to 0, resulting in selective alignment in the shared subspace.

The Disc-InfoMax Objective This proximity mask enables selective alignment through a simple mechanism: shared components should align, yielding small residuals ($\mathbf{\Delta} \approx \mathbf{0}$), whereas unshared components do the opposite. The proximity mask thus acts as a gate that opens when representations come close and shuts when they are far apart (Fig. 1), resulting in the *discernment term*: $\mathbb{E}_{\mathbf{X}}\left[\left\|\mathbf{\Phi}^{(\sigma)}(\mathbf{\Delta}) \odot \mathbf{\Delta}\right\|_2^2\right]$. Based on this mechanism, we formulate Disc-InfoMax as follows:

$$\mathcal{L}(g_\theta, t; \mathbf{\Phi}^{(\sigma)}) = \underbrace{\mathbb{E}_{\mathbf{X}}\left[\left\|\mathbf{\Phi}^{(\sigma)}(\mathbf{\Delta}) \odot \mathbf{\Delta}\right\|_2^2\right]}_{\text{Discernment}} - \underbrace{\sum\nolimits_{i \in \{1,2\}} H\left(t(\hat{\mathbf{Z}}^{(i)})\right)}_{\text{Entropy Regularization}} \tag{4}$$

$$\text{Reg}(\mathbf{\Phi}^{(\sigma)}) = -\mathbb{E}_{\mathbf{X}}\left[\left\|\mathbf{\Phi}^{(\sigma)}(\mathbf{\Delta})\right\|_1\right], \tag{5}$$

with $H(\cdot)$ denoting differential entropy after an auxiliary projection $t \in \mathcal{T}$ (Def. 4 in App. B) to prevent representational collapse while allowing dependencies between learned variables. By forming a bi-level optimization with the regularization term:

$$\text{Inner (InfoMax):} \quad \mathcal{G}_\sigma^\circ = \operatorname*{argmin}_{g_\theta \in \mathcal{G}} \inf_{t \in \mathcal{T}} \mathcal{L}(g_\theta, t; \mathbf{\Phi}^{(\sigma)}), \tag{6}$$

$$\text{Outer (Regularization):} \quad \mathcal{G}_\sigma^* = \operatorname*{argmin}_{g_{\theta^\circ} \in \mathcal{G}_\sigma^\circ} \text{Reg}(\mathbf{\Phi}^{(\sigma)}), \tag{7}$$

we theoretically guarantee block-identifiability:

Theorem 1 *(Block-Identifiability of Disc-InfoMax).* In the asymptotic limit $\sigma \to 0^+$, any optimal encoder $g^* \in \mathcal{G}_{0^+}^* := \limsup_{\sigma \to 0^+} \mathcal{G}_\sigma^*$ achieving the theoretical bi-level Disc-InfoMax objective (Eqs. (6) and (7)) block-identifies the shared causal variables $\mathbf{C}_S$ dynamically defined for any given observability pattern.

Proof Sketch. Because unshared coordinates contain independent noise, exact pairwise invariance ($\Delta_j = 0$) requires coordinate j to depend exclusively on $\mathbf{C}_S$. Under maximal differential entropy, at most $\|\mathbf{S}\|_0$ coordinates can achieve zero residual; aligning any further coordinate forces representation collapse onto the lower-dimensional manifold of $\mathbf{C}_S$, driving entropy to $-\infty$. Thus, maximizing the regularizer $\mathbb{E}[\|\mathbf{\Phi}^{(\sigma)}(\mathbf{\Delta})\|_1]$ drives the encoder to align precisely $\|\mathbf{S}\|_0$ coordinates to capture $\mathbf{C}_S$. In the limit $\sigma \to 0^+$, $\mathbf{\Phi}^{(\sigma)}$ converges to the discrete indicator $\mathbb{I}(\mathbf{\Delta} = \mathbf{0})$, nullifying unshared fluctuations and guaranteeing diffeomorphic block-identifiability of $\mathbf{C}_S$ (full proof in App. B.1).

While Thm. 1 analyzes the asymptotic bi-level formulation with static bandwidth, practical neural networks face vanishing gradients at the limit $\sigma = 0^+$ and suffer from the computational load of bi-level optimization. Instead, practical training typically optimizes the inner and outer problem simultaneously while annealing σ. In App. B.2, we analyze this parameter-space optimization and prove that this joint optimization along an annealing schedule strictly converges to the asymptotic bi-level optimum $\Theta^*_{0^+}$, preserving the identical block-identifiability guarantee (Thm. 2).

Dynamic Identifiability Algebra While Disc-InfoMax identifies the shared causal variables $\mathbf{C}_S$, varying observability patterns across pairs induce overlaps among them. In App. B.3, we establish a dynamic identifiability algebra for Disc-InfoMax, proving that intersection and union operations natively preserve block-identifiability. Consequently, identified blocks can be algebraically refined to achieve finer granularity than raw observation masks: any subset formed via intersection and union closures (formalized as the identifiable index set Q in App. C.3) is also guaranteed to be identified.

Theoretical details on distributional symmetry ($p(\mathbf{X}^{(1)}) = p(\mathbf{X}^{(2)})$), the expanded information capacity of Disc-InfoMax over vanilla InfoMax, and connections to practical SSL are provided in App. C.

3.2 Tractable Optimization via BayesDisc

While the asymptotic optimum guarantees block-identifiability (Thm. 1), finding it via empirical training poses severe bottlenecks. First, orthodox bi-level optimization is computationally intractable. Second, the inner problem suffers from a chicken-and-egg co-adaptation trap: at initialization, uninformative residuals produce an arbitrary mask $\mathbf{\Phi}^{(\sigma)}$, which misguides alignment and reinforces an erroneous dimension allocation. The encoder easily gets trapped in this deep local minimum, failing to identify the true causal variables. To bypass these issues, we propose *BayesDisc*, a practical algorithm for Disc-InfoMax that couples a Bayesian proximity mask with bandwidth annealing.

3.2.1 Bayesian Posterior as Proximity Mask

To resolve the chicken-and-egg problem, we take the Bayesian approach. Here, the core question is: *given residual Δ_j, what is the probability that coordinate j truly encodes shared content (i.e., that the proximity mask should be active)?* We formalize this as a probabilistic mixture model:

$$p(\Delta_j) = p(\Delta_j|a)p(a) + p(\Delta_j|\neg a)p(\neg a), \tag{8}$$

where event a denotes that coordinate j encodes shared content with prior $w \coloneqq p(a)$ (and $\neg a$ denotes unshared event). By Bayes' theorem, the posterior of sharing given observed residual Δ_j yields:

$$P(a|\Delta_j) = \frac{p(\Delta_j|a) \cdot w}{p(\Delta_j)}. \tag{9}$$

We directly deploy this posterior as our proximity mask, setting $\Phi_j^{(\sigma)}(\Delta_j) = P(a|\Delta_j)$, and detach it from the computation graph during backpropagation to break the co-adaptation cycle.

Prior Determination via Peak Alignment To use this posterior directly as a proximity mask, it must satisfy the three conditions of Def. 2. First, enforcing *Peak Alignment* ($\Phi_j^{(\sigma)}(0) = P(a|0) = 1$ at $\Delta_j = 0$) anchors the prior probability by the density ratio at the origin: $w = \frac{p(0)}{p(0|a)}$.

Gaussian Priors via SIGReg To satisfy the remaining two conditions and render the posterior analytically computable, we employ Sketched Isotropic Gaussian Regularization (SIGReg) (Balestriero & LeCun, 2025) for entropy maximization. SIGReg regularizes representations $\hat{\mathbf{Z}}$ toward an isotropic standard normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$, inducing a marginal residual distribution $p(\Delta_j) \sim \mathcal{N}(0, 2.0)$. Because shared dimensions should naturally be closer to each other than the unconstrained marginal, we model their conditional distribution as a narrower zero-mean Gaussian $p(\Delta_j|a) \sim \mathcal{N}(0, \sigma^2)$ with bandwidth parameter $\sigma^2 < 2.0$. Under these Gaussian densities, the Peak Alignment constraint determines the prior as $w = \sigma/\sqrt{2.0}$. Evaluating Eq. (9) analytically, normalization constants cancel to yield the closed-form proximity mask $\Phi_j^{(\sigma)}(\Delta_j) = P(a|\Delta_j) = \exp\left(-\frac{2-\sigma^2}{4\sigma^2}\Delta_j^2\right)$. Because

$\sigma^2 < 2.0$, this closed-form posterior strictly satisfies *Monotonic Decay* (decreasing monotonically in $|\Delta_j|$) and *Asymptotic Nullification* ($\lim_{\sigma\to 0^+} \Phi_j^{(\sigma)}(\Delta_j) = 0$ for all $\Delta_j \neq 0$). Thus, all three conditions of Def. 2 are satisfied and analytically computable via a single scalar parameter σ.[2]

3.2.2 Bandwidth Annealing as an Outer Optimization Proxy

Combining Bayesian mask inference with SIGReg resolves the non-convexity of the inner problem. To address the outer problem without expensive nested loops, we leverage bandwidth annealing as an automatic optimization proxy. Specifically, the bandwidth σ directly couples both optimization levels: for the *inner problem*, decreasing $\sigma \to 0$ narrows the acceptance window toward the ideal indicator $\lim_{\sigma\to 0^+} \Phi_j^{(\sigma)}(\Delta_j) = \mathbb{I}(\Delta_j = 0)$, driving representations toward sharp discernment; for the *outer problem*, decreasing σ shrinks the gating kernel, which increases the regularization penalty $\mathrm{Reg}(\mathbf{\Phi}^{(\sigma)}) = -\mathbb{E}[\|\mathbf{\Phi}^{(\sigma)}(\mathbf{\Delta})\|_1]$ unless representations adaptively align to keep residuals within the narrowing window. Consequently, annealing σ at a calibrated rate acts as an effective proxy that optimizes both levels simultaneously—preventing premature gate closure while steadily driving representations to the optimum.[3] Moreover, as outlined in Alg. 1 (App. C.5), BayesDisc integrates into standard InfoMax SSL workflows with few extra lines and incurs negligible computational overhead, as the posterior mask is evaluated analytically.

This proxy, however, raises a question: *does BayesDisc truly achieve the core goal of Disc-InfoMax, identifying the causal variables?* We answer this through extensive empirical validations in Sec. 4.

4 Experiments

In this section, we empirically evaluate BayesDisc on both synthetic numerical benchmarks and a complex 3D visual dataset, Causal3DIdent (von Kügelgen et al., 2021).

Baselines and Reference Methods Because no existing CRL method handles varying shared contents without view labels, we compare BayesDisc against standard self-supervised baselines: InfoNCE (Chen et al., 2020) and LeJEPA (Balestriero & LeCun, 2025) (the underlying SSL framework of our method). We also include a *True Mask* baseline trained using Multi-View CRL (Yao et al., 2023). Because ground-truth mask labels $\mathbf{m}_k$ are provided in this setup, Multi-View CRL is directly applicable and serves to calibrate empirical performance against supervised view-label methods.

Evaluation Protocol Following standard self-supervised evaluation, all evaluations are conducted on single-view, fully observed observations ($\mathbf{M} = \mathbf{1}$). We evaluate via Mean Correlation Coefficient (MCC) for element-wise identifiability and R^2 score for non-linear decodability (detailed in App. F.3). Across all tables, the best performance among self-supervised methods (excluding the supervised True Mask) are bolded.

Element-Level Identifiability and Training Setup To guarantee element-wise identifiability, the pool of observability patterns across both benchmarks are chosen to ensure that each causal variable can be identified. All results are aggregated across 16 independent training runs with different random seeds controlling data generation and network initialization. For the details on training, architecture, bandwidth annealing, and computation load, see App. F.

4.1 Numerical Dataset

Data Generation and Setup We generate ground-truth causal variables $\mathbf{C} \in \mathbb{R}^n$ from an SCM over a DAG (see App. E.1 for generation details), apply view-specific masks $\mathbf{M}^{(i)} \in \{0,1\}^n$, and map the masked latents $\mathbf{Z}^{(i)}$ to observations $\mathbf{X}^{(i)} \in \mathbb{R}^{64}$ via an invertible MLP. We evaluate in the well-specified regime where the latent dimension matches the ground truth ($m = n$). We evaluate systematic sweeps over graph density (k), depth (L), and dimension (n), with default parameters $k = 1, L = 4, n = 8$ (detailed in App. E).

[2] We demonstrate in App. D.1 that the inferred posterior remains asymptotically stable and correctly bounded even in the theoretical limit of perfect discernment ($\sigma \to 0^+$), despite the prior w approaching zero.

[3] Practical annealing heuristics are provided in App. D.2.

Table 1: **Numerical results on fully observed test distribution** ($\mathbf{M = 1}$) across different parameters (MCC ↑, R^2 ↑). Default setting is $k = 1, L = 4, n = 8$.

	InfoNCE		LeJEPA		BayesDisc (Ours)		True Mask	
Setting	MCC	R^2	MCC	R^2	MCC	R^2	MCC	R^2
Default	$0.456_{\pm 0.039}$	$0.821_{\pm 0.055}$	$0.457_{\pm 0.045}$	$0.834_{\pm 0.049}$	$\mathbf{0.963_{\pm 0.024}}$	$\mathbf{0.986_{\pm 0.006}}$	$0.973_{\pm 0.031}$	$0.993_{\pm 0.002}$
$k = 0$	$0.451_{\pm 0.060}$	$0.801_{\pm 0.048}$	$0.508_{\pm 0.061}$	$0.857_{\pm 0.059}$	$\mathbf{0.987_{\pm 0.001}}$	$\mathbf{0.989_{\pm 0.001}}$	$0.987_{\pm 0.000}$	$0.995_{\pm 0.000}$
$k = 2$	$0.474_{\pm 0.040}$	$0.842_{\pm 0.040}$	$0.437_{\pm 0.033}$	$0.850_{\pm 0.037}$	$\mathbf{0.955_{\pm 0.025}}$	$\mathbf{0.986_{\pm 0.004}}$	$0.977_{\pm 0.004}$	$0.994_{\pm 0.002}$
$k = 3$	$0.486_{\pm 0.036}$	$0.855_{\pm 0.041}$	$0.460_{\pm 0.037}$	$0.861_{\pm 0.043}$	$\mathbf{0.951_{\pm 0.026}}$	$\mathbf{0.985_{\pm 0.007}}$	$0.976_{\pm 0.005}$	$0.994_{\pm 0.002}$
$L = 2$	$0.449_{\pm 0.048}$	$0.823_{\pm 0.060}$	$0.454_{\pm 0.035}$	$0.834_{\pm 0.050}$	$\mathbf{0.969_{\pm 0.012}}$	$\mathbf{0.986_{\pm 0.003}}$	$0.981_{\pm 0.004}$	$0.994_{\pm 0.001}$
$L = 6$	$0.446_{\pm 0.033}$	$0.820_{\pm 0.054}$	$0.446_{\pm 0.044}$	$0.833_{\pm 0.049}$	$\mathbf{0.959_{\pm 0.037}}$	$\mathbf{0.985_{\pm 0.005}}$	$0.981_{\pm 0.004}$	$0.994_{\pm 0.001}$
$L = 8$	$0.463_{\pm 0.045}$	$0.824_{\pm 0.056}$	$0.458_{\pm 0.049}$	$0.836_{\pm 0.052}$	$\mathbf{0.968_{\pm 0.013}}$	$\mathbf{0.987_{\pm 0.002}}$	$0.980_{\pm 0.004}$	$0.993_{\pm 0.001}$
$n = 4$	$0.516_{\pm 0.090}$	$0.870_{\pm 0.058}$	$0.489_{\pm 0.055}$	$0.873_{\pm 0.056}$	$\mathbf{0.965_{\pm 0.025}}$	$\mathbf{0.994_{\pm 0.004}}$	$0.974_{\pm 0.032}$	$0.997_{\pm 0.001}$
$n = 12$	$0.465_{\pm 0.027}$	$0.840_{\pm 0.035}$	$0.445_{\pm 0.025}$	$0.844_{\pm 0.033}$	$\mathbf{0.961_{\pm 0.021}}$	$\mathbf{0.983_{\pm 0.003}}$	$0.981_{\pm 0.002}$	$0.992_{\pm 0.001}$
$n = 16$	$0.470_{\pm 0.037}$	$0.831_{\pm 0.028}$	$0.447_{\pm 0.021}$	$0.856_{\pm 0.031}$	$\mathbf{0.799_{\pm 0.150}}$	$\mathbf{0.945_{\pm 0.051}}$	$0.982_{\pm 0.003}$	$0.988_{\pm 0.002}$

Identifiability on In-Distribution Test Data As shown in Tab. 1, BayesDisc substantially outperforms both vanilla SSL baselines across all configurations, closely matching the performance of the True Mask, successfully identifying the underlying causal variables. Notably, SSL baselines achieve lower R^2 since they hold smaller theoretical information capacity compared to BayesDisc (App. C.4). Evaluation on the train (masked) distribution displays an identical trend (App. G.1). While moderate latent dimensionality ($n = 16$) increases seed variance due to optimization non-convexity (App. G.2.1), violation of mixing function invertibility causes major failure (App. G.2.2).

Out-of-Distribution Generalization To evaluate representation robustness, we benchmark 5 systematic OOD environments on the numerical dataset: (1) *Cause Intervention*, (2) *Effect Intervention*, (3) *Mechanism Shift*, (4) *Edge Addition*, and (5) *Support Extrapolation*, with fixed mixing (App. E.5). As shown in Tab. 2, BayesDisc maintains high identifiability (MCC $\geq$ 0.90) across all shifts. Under atomic interventions, BayesDisc exhibits superior resilience (0.922 and 0.900) compared even to the supervised True Mask baseline (0.857 and 0.852), confirming that learned coordinates reflect modular, autonomous mechanisms. In contrast, vanilla SSL remains uniformly entangled (MCC $\approx$ 0.42–0.47), demonstrating that InfoMax without discernment term fails to identify causal variables.

Table 2: **OOD Generalization** (MCC ↑, mean ± std). Full sweeps with InfoNCE in App. G.2.3.

Shift	LeJEPA	BayesDisc (Ours)	True Mask
Standard	$0.457_{\pm 0.045}$	$\mathbf{0.963_{\pm 0.024}}$	$0.973_{\pm 0.031}$
Cause Interv.	$0.429_{\pm 0.035}$	$\mathbf{0.922_{\pm 0.019}}$	$0.857_{\pm 0.008}$
Effect Interv.	$0.428_{\pm 0.041}$	$\mathbf{0.900_{\pm 0.033}}$	$0.852_{\pm 0.030}$
Mech. Shift	$0.436_{\pm 0.046}$	$\mathbf{0.962_{\pm 0.024}}$	$0.973_{\pm 0.031}$
Edge Add.	$0.467_{\pm 0.050}$	$\mathbf{0.958_{\pm 0.023}}$	$0.972_{\pm 0.031}$
Extrapolation	$0.421_{\pm 0.034}$	$\mathbf{0.947_{\pm 0.027}}$	$0.962_{\pm 0.026}$

Causal Discovery We evaluate whether learned representations facilitate downstream causal graph recovery by applying LiNGAM, PC, and GES directly to the latents (Tab. 3). BayesDisc consistently outperforms the vanilla SSL baseline, achieving lower Structural Hamming Distance (SHD) across all algorithms and closely approaching the True Mask baseline. While ground-truth causal variables yield lower SHD due to the absence of empirical estimation noise, BayesDisc reliably preserves causal dependencies.

Table 3: **Causal Discovery** (SHD ↓, mean ± std across 16 seeds on default setting). Full results in App. G.2.4.

Representation	LiNGAM (↓)	PC (↓)	GES (↓)
SSL (LeJEPA)	$20.25_{\pm 8.56}$	$19.94_{\pm 9.31}$	$21.62_{\pm 15.61}$
BayesDisc (Ours)	$\mathbf{15.75_{\pm 5.44}}$	$\mathbf{14.50_{\pm 4.75}}$	$\mathbf{16.38_{\pm 5.98}}$
True Mask	$13.62_{\pm 6.11}$	$13.56_{\pm 12.00}$	$12.69_{\pm 13.46}$
True Variables	$8.50_{\pm 9.25}$	$7.31_{\pm 4.34}$	$7.19_{\pm 7.78}$

4.2 CAUSAL3DIDENT DATASET

Dataset Setup and Overparameterization We evaluate BayesDisc on Causal3DIdent (von Kügelgen et al., 2021), which renders complex 3D scenes ($224 \times 224 \times 3$ images) governed by an 11-variable SCM encompassing object poses, spotlight positions, and hues (sample pairs visualized in Fig. 2, App. F.2). While our theoretical identifiability guarantees analyze the well-specified regime ($m = n$), the true causal dimensionality n is typically unknown for real-world visual observations. To evaluate robustness under an unknown causal dimension, we intentionally overparameterize the latent representation space to $m = 32$ for the $n = 11$ underlying generative factors.

Pairing Protocol and Symmetrization Since Causal3DIdent is pre-rendered, pairs are constructed by retrieving nearest neighbors in latent space for partially shared factors (App. F.2). The order-symmetry in discernment term effectively makes the marginal distributions identical, $p(\mathbf{X}^{(1)}) = p(\mathbf{X}^{(2)})$ (App. C.1), fulfilling the theoretical symmetry requirement of Disc-InfoMax.

Results and Analysis As shown in Tab. 4, BayesDisc achieves an MCC of 0.663, outperforming InfoNCE (0.560) and LeJEPA (0.513). Notably, BayesDisc matches the supervised True Mask baseline (0.648), demonstrating that BayesDisc incurs no empirical penalty over true mask labels. The MCC ceiling of True Mask shows that it stems from the complexities of the dataset, such as approximate nearest-neighbor pairing and unordered discrete variable (object shape), rather than BayesDisc itself. Furthermore, under latent overparameterization ($m = 32$ for $n = 11$), the causal variables are identified without dimensional collapse in the excess dimensions. We analyze element-wise identifiability in App. G.3.

Table 4: Summary results on Causal3DIdent.

Method	MCC (↑)	R^2 (↑)
InfoNCE	$0.560_{\pm 0.017}$	$0.962_{\pm 0.001}$
LeJEPA	$0.513_{\pm 0.018}$	$\mathbf{0.971_{\pm 0.002}}$
BayesDisc (Ours)	$\mathbf{0.663_{\pm 0.032}}$	$0.969_{\pm 0.001}$
True Mask	$0.648_{\pm 0.035}$	$0.921_{\pm 0.001}$

5 RELATED WORK

5.1 SELF-SUPERVISED LEARNING

Self-supervised learning (SSL) has driven a paradigm shift in representation learning. In Natural Language Processing, SSL has achieved massive success through discrete token prediction frameworks. This includes bidirectional masked language modeling (Devlin et al., 2019; Liu et al., 2019), as well as autoregressive next-token prediction, which serves as the foundational SSL objective for large-scale generative models (Radford et al., 2018; Brown et al., 2020). However, as highlighted by broader SSL perspectives (LeCun & Misra, 2021), extending predictive reconstruction to continuous, high-dimensional, and noisy domains like vision is notoriously difficult. Modeling every pixel often forces the network to focus on unpredictable, low-level details rather than semantic content.

To overcome this, modern computer vision relies on Joint-Embedding Architectures (JEAs) that perform prediction directly in the representation space, often leveraging the InfoMax principle to maximize information content between paired observations. However, foundational critiques of early InfoMax frameworks demonstrated that maximizing mutual information (MI) is not sufficient on its own to guarantee useful representations (Tschannen et al., 2019). Because MI is invariant to invertible transformations, naive maximization can yield highly entangled spaces unless constrained by strong inductive biases. Since then, the field has advanced significantly to prevent representation collapse through asymmetric architectural constraints (e.g., BYOL (Grill et al., 2020), DINO (Caron et al., 2021; Oquab et al., 2023; Siméoni et al., 2025)) and advanced contrastive regularizers, including *sample-contrastive* (e.g., SimCLR (Chen et al., 2020), MoCo (He et al., 2020)) and *dimension-contrastive* approaches (e.g., Barlow Twins (Zbontar et al., 2021), VICReg (Bardes et al., 2021), LeJEPA (Balestriero & LeCun, 2025)). Despite these advances, current methods still rely on implicit architectural biases and blindly decorrelate all dimensions alike. Disc-InfoMax and BayesDisc is built on this modern InfoMax foundation but replaces these implicit biases with the discernment mechanism. By selectively aligning only the inferred shared content dimensions, our approach identifies causal variables directly in the InfoMax SSL pipeline.

5.2 Causal Representation Learning

Causal Representation Learning (CRL) seeks to identify ground-truth latent causal variables from high-dimensional observations (Schölkopf et al., 2021). To achieve identifiability, existing methods carefully craft assumptions regarding the data-generating process, often relying on underlying invariance principles. For example, Interventional CRL exploits data collected under different interventional environments (von Kügelgen et al., 2021; Ahuja et al., 2023; Squires et al., 2023; Liang et al., 2023; Zhang et al., 2023; Varici et al., 2023; 2024). Similarly, Temporal CRL leverages the natural structure of time-series data (Yao et al., 2021; 2022; Lippe et al., 2022b;a; 2023; Lachapelle et al., 2022; Li et al., 2024), and Sparsity CRL (Xu et al., 2024) uses the sparsity principle along group identities under an unpaired partially observed setting, while Multi-task CRL utilizes a supervised setting with multiple task labels (Lachapelle et al., 2023; Fumero et al., 2023). Operating under a similar paradigm is Multi-View (or Counterfactual) CRL, which identifies shared latent blocks from simultaneously observed views generated from partially overlapping subsets of causal variables (von Kügelgen et al., 2021; Yao et al., 2023; Locatello et al., 2020; Daunhawer et al., 2022; Ahuja et al., 2022). Recently, a unified identifiability framework was proposed encompassing these CRL setups under an invariance principle (Yao et al., 2024); however, it still fundamentally presumes access to auxiliary view or environment labels, and thus does not apply to our view-label-free setting.

Our work aligns closest with the Multi-View CRL paradigm, which exploits the sample-level invariance of the latents shared across views. However, existing multi-view frameworks (Yao et al., 2023; 2024) rely on explicit view labels, and optimizing their information-sharing regularizers (e.g., combinatorial content selectors) is highly non-convex. While recent work investigates identifiability without view labels (Xu et al., 2026), it only covers specific data distributions (potentially degenerate Gaussian mixture models under piecewise affine mixing) and requires specialized, complicated training procedures. In contrast, BayesDisc addresses general causal DAG distributions under smooth nonlinear diffeomorphisms. By dynamically inferring the shared content via an analytic Bayesian posterior, BayesDisc eliminates the need for view labels and trains directly with standard deep learning optimizers, providing a tractable path toward CRL under SSL pipeline.

6 Conclusion

We introduced Disc-InfoMax and *BayesDisc*, proving that optimizing paired representations under a dynamic proximity mask guarantees block-identifiability without view labels, while providing an analytic Bayesian surrogate for stable end-to-end training. As the first work to make CRL viable for InfoMax SSL, our framework marks an essential step toward overcoming the dataset scaling bottleneck in CRL and learning causal representations across InfoMax SSL applications. Immediate next steps must address these boundaries by automating the σ curriculum and extending our probabilistic formulation to discrete variables. Ultimately, adapting dynamic shared-content inference from continuous SSL frameworks to modern autoregressive paradigms and world models offers a highly promising path for scaling causal representation learning to foundational models.

Limitations Examining our framework's practical boundaries contextualizes not only its current scope, but also why its core contributions remain foundational and practically viable. First, while our identifiability guarantees hold at the population level without an end-to-end finite-sample consistency proof (Sec. H.1), Disc-InfoMax shows the theoretical plausibility of block-identifiability in an InfoMax self-supervised learning setup, which is a prerequisite for empirical exploration. Moreover, BayesDisc provides a tractable surrogate that empirically enables standard optimizers to reach the theoretical optimum across diverse regimes. Second, because proximity masks evaluate observation pairs, single test samples cannot compute the mask; in practice, downstream predictors can operate on the complete, block-identified representation vector, while pair-based tasks (e.g., counterfactual reasoning) directly evaluate the posterior mask across test pairs (Sec. H.2). Third, flexible projection heads t for entropy regularization risk dimensional collapse; we resolve this by omitting t and using isotropic Gaussian regularization (SIGReg). In practice, since training loss never reaches strictly zero, the discernment term and the entropy maximization term balance and recover the dependencies between the identified variables (App. H.3, H.4, and Fig. 3 in App. G). Finally, without the ground truth causal variables, no metrics can be computed for monitoring σ annealing. We address unsupervised tuning via an annealing heuristic that monitors representation variance to stop safely above the collapse threshold (Sec. D.2). We detail these theoretical boundaries in Sec. H.

AI USE STATEMENT

In this work, generative AI tools (LLMs) were utilized under strict human oversight: **(1) AI-assisted tasks:** LLMs assisted in drafting initial code from author-specified architectures, drafting mathematical derivations from author-conceived proof sketches, refining informal prose into formal academic text, formatting LaTeX tables, and serving as an exploratory brainstorming aid. **(2) Tasks without AI assistance:** Generative AI was not used for the literature survey and citations, the core conceptualization of the discernment mechanism, experimental design, or the interpretation of empirical results. **(3) Verification and oversight:** The authors rigorously verified all AI-assisted outputs: code was validated through benchmark execution, proof steps were verified line-by-line for deductive correctness, and all text was critically reviewed and approved. The authors take full responsibility for all content, claims, and artifacts in this paper.

REPRODUCIBILITY STATEMENT

Complete theoretical claims and mathematical proofs are provided in App. B. Detailed algorithmic steps and pseudocode are outlined in App. C.5. Full experimental configurations, neural network architectures, hyperparameter choices, evaluation metrics, and dataset generation pipelines are detailed in App. E and App. F. Anonymized code, configuration files, and instructions to reproduce all empirical benchmarks are available at `https://anonymous.4open.science/r/disc-infomax`.

REFERENCES

Kartik Ahuja, Jason S. Hartford, and Yoshua Bengio. Weakly Supervised Representation Learning with Sparse Perturbations. In *Advances in Neural Information Processing Systems*, volume 35, pp. 15516–15528, December 2022.

Kartik Ahuja, Divyat Mahajan, Yixin Wang, and Yoshua Bengio. Interventional Causal Representation Learning. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 372–407. PMLR, July 2023.

Randall Balestriero and Yann LeCun. LeJEPA: Provable and Scalable Self-Supervised Learning Without the Heuristics, November 2025. arXiv:2511.08544 [cs].

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning. In *International Conference on Learning Representations (ICLR)*, October 2021.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jegou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging Properties in Self-Supervised Vision Transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9630–9640, October 2021. doi: 10.1109/ICCV48922.2021.00951. ISSN: 2380-7504.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 1597–1607. PMLR, November 2020.

Imant Daunhawer, Alice Bizeul, Emanuele Palumbo, Alexander Marx, and Julia E. Vogt. Identifiability Results for Multimodal Contrastive Learning. In *International Conference on Learning Representations (ICLR)*, September 2022.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Jill Burstein, Christy Doran, and

Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423.

Marco Fumero, Florian Wenzel, Luca Zancato, Alessandro Achille, Emanuele Rodolà, Stefano Soatto, Bernhard Schölkopf, and Francesco Locatello. Leveraging sparse and shared feature activations for disentangled representation learning. *Advances in Neural Information Processing Systems*, 36: 27682–27698, 2023.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap Your Own Latent - A New Approach to Self-Supervised Learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 21271–21284. Curran Associates, Inc., 2020.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum Contrast for Unsupervised Visual Representation Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9729–9738, 2020.

Sebastien Lachapelle, Pau Rodriguez, Yash Sharma, Katie E. Everett, Rémi Le Priol, Alexandre Lacoste, and Simon Lacoste-Julien. Disentanglement via Mechanism Sparsity Regularization: A New Principle for Nonlinear ICA. In *Proceedings of the First Conference on Causal Learning and Reasoning*, pp. 428–484. PMLR, June 2022.

Sebastien Lachapelle, Tristan Deleu, Divyat Mahajan, Ioannis Mitliagkas, Yoshua Bengio, Simon Lacoste-Julien, and Quentin Bertrand. Synergies between Disentanglement and Sparsity: Generalization and Identifiability in Multi-Task Learning. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 18171–18206. PMLR, July 2023.

Yann LeCun and Ishan Misra. Self-supervised learning: The dark matter of intelligence, March 2021. Published: Meta AI Blog.

Zijian Li, Yifan Shen, Kaitao Zheng, Ruichu Cai, Xiangchen Song, Mingming Gong, Guangyi Chen, and Kun Zhang. On the Identification of Temporal Causal Representation with Instantaneous Dependence. In *International Conference on Learning Representations (ICLR)*, October 2024.

Wendong Liang, Armin Kekić, Julius von Kügelgen, Simon Buchholz, Michel Besserve, Luigi Gresele, and Bernhard Schölkopf. Causal Component Analysis. In *Advances in Neural Information Processing Systems*, volume 36, pp. 32481–32520. Curran Associates, Inc., 2023. doi: 10.52202/075280-1411.

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M. Asano, Taco Cohen, and Efstratios Gavves. Causal Representation Learning for Instantaneous and Temporal Effects in Interactive Systems. In *International Conference on Learning Representations (ICLR)*, September 2022a.

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M. Asano, Taco Cohen, and Stratis Gavves. CITRIS: Causal Identifiability from Temporal Intervened Sequences. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 13557–13603. PMLR, June 2022b.

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M. Asano, Taco Cohen, and Efstratios Gavves. BISCUIT: Causal Representation Learning from Binary Interactions. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, pp. 1263–1273. PMLR, July 2023.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach, July 2019. arXiv:1907.11692 [cs].

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 4114–4124. PMLR, May 2019.

Francesco Locatello, Ben Poole, Gunnar Raetsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. Weakly-Supervised Disentanglement Without Compromises. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 6348–6359. PMLR, November 2020.

Qi Lyu, Xiao Fu, Weiran Wang, and Songtao Lu. Understanding Latent Correlation-Based Multiview Learning and Self-Supervision: An Identifiability Perspective. In *International Conference on Learning Representations (ICLR)*, October 2021.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation Learning with Contrastive Predictive Coding, January 2019. arXiv:1807.03748 [cs].

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research*, July 2023.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. *OpenAI*, 2018.

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, May 2021. ISSN 1558-2256. doi: 10.1109/JPROC.2021.3058954.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, August 2025. arXiv:2508.10104 [cs].

Chandler Squires, Anna Seigal, Salil S. Bhate, and Caroline Uhler. Linear Causal Disentanglement via Interventions. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 32540–32560. PMLR, July 2023.

Michael Tschannen, Josip Djolonga, Paul K. Rubenstein, Sylvain Gelly, and Mario Lucic. On Mutual Information Maximization for Representation Learning. In *International Conference on Learning Representations (ICLR)*, September 2019.

Burak Varici, Emre Acarturk, Karthikeyan Shanmugam, Abhishek Kumar, and Ali Tajer. Score-based Causal Representation Learning with Interventions, May 2023. arXiv:2301.08230 [stat].

Burak Varici, Emre Acartürk, Karthikeyan Shanmugam, and Ali Tajer. General Identifiability and Achievability for Causal Representation Learning. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, pp. 2314–2322. PMLR, April 2024.

Julius von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-Supervised Learning with Data Augmentations Provably Isolates Content from Style. In *Advances in Neural Information Processing Systems*, volume 34, pp. 16451–16467. Curran Associates, Inc., 2021.

Danru Xu, Dingling Yao, Sebastien Lachapelle, Perouz Taslakian, Julius von Kügelgen, Francesco Locatello, and Sara Magliacane. A Sparsity Principle for Partially Observable Causal Representation Learning. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 55389–55433. PMLR, July 2024.

Danru Xu, Sebastien Lachapelle, and Sara Magliacane. Identifiability of Potentially Degenerate Gaussian Mixture Models With Piecewise Affine Mixing. In *Proceedings of The 29th International Conference on Artificial Intelligence and Statistics*, pp. 1252–1260. PMLR, August 2026.

Dingling Yao, Danru Xu, Sebastien Lachapelle, Sara Magliacane, Perouz Taslakian, Georg Martius, Julius von Kügelgen, and Francesco Locatello. Multi-View Causal Representation Learning with Partial Observability. In *International Conference on Learning Representations (ICLR)*, October 2023.

Dingling Yao, Dario Rancati, Riccardo Cadei, Marco Fumero, and Francesco Locatello. Unifying Causal Representation Learning with the Invariance Principle. In *International Conference on Learning Representations (ICLR)*, October 2024.

Weiran Yao, Yuewen Sun, Alex Ho, Changyin Sun, and Kun Zhang. Learning Temporally Causal Latent Processes from General Temporal Data. In *International Conference on Learning Representations (ICLR)*, October 2021.

Weiran Yao, Guangyi Chen, and Kun Zhang. Temporally disentangled representation learning. *Advances in Neural Information Processing Systems*, 35:26492–26503, 2022.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stephane Deny. Barlow Twins: Self-Supervised Learning via Redundancy Reduction. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 12310–12320. PMLR, July 2021.

Jiaqi Zhang, Kristjan Greenewald, Chandler Squires, Akash Srivastava, Karthikeyan Shanmugam, and Caroline Uhler. Identifiability guarantees for causal disentanglement from soft interventions. *Advances in Neural Information Processing Systems*, 36:50254–50292, 2023.

A NOTATION AND TERMINOLOGY

We adopt the following mathematical and typographical conventions throughout the paper:

- **Vectors and Scalars:** Bold uppercase and lowercase letters denote random and deterministic vectors ($\mathbf{C}, \mathbf{Z}, \mathbf{X}, \mathbf{M}, \mathbf{S}, \mathbf{\Delta}, \mathbf{N}$), whereas regular italic letters denote scalars, coordinate indices, or dimensions ($n, m, d, C_j, \Delta_j, Z_j$).
- **Spaces and Manifolds:** Calligraphic letters denote functional hypothesis classes, probability domains, parameter manifolds, and graphs ($\mathcal{C}, \mathcal{Z}, \mathcal{X}, \hat{\mathcal{Z}}, \mathcal{M}, \mathcal{G}, \mathcal{T}, \Theta, \mathcal{G}_{\text{DAG}}$).
- **Subvectors and Projections:** For any binary mask $K \in \{0,1\}^n$, the subscript notation $\mathbf{C}_K \in \mathcal{C}_K$ denotes the coordinate subvector restricted to the active coordinate indices $\{j \mid K_j = 1\}$, distinguishing it from the dimension-preserving Hadamard product $\mathbf{K} \odot \mathbf{C}$.
- **Estimators and Limits:** Hats denote learned or estimated quantities (e.g., $\hat{\mathbf{Z}}$ for encoder representations, $\hat{\boldsymbol{\phi}}$ for the limiting inferred mask).

Disambiguation of Latent and Mask Notations. To ensure clarity across theoretical proofs and empirical setups, we highlight key distinctions:

- **Causal Latents vs. Masked States vs. Learned Representations:** $\mathbf{C} = (C_1, \ldots, C_n)^\top \in \mathcal{C}$ denotes the true, underlying causal variables generated by the Structural Causal Model (SCM). $\mathbf{Z}^{(i)} \in \mathcal{Z}$ denotes the partially occluded latent states before mixing, where inactive coordinates are filled with nuisance noise $\mathbf{N}^{(i)}$. In contrast, $\hat{\mathbf{Z}} = g_\theta(\mathbf{X}) \in \hat{\mathcal{Z}} \subseteq \mathbb{R}^m$ denotes the encoder representations.
- **Mask Progression:** $\mathbf{M}^{(1)}, \mathbf{M}^{(2)} \in \{0,1\}^n$ denote the paired instance observability masks, inducing the true shared content mask $\mathbf{S} \coloneqq \mathbf{M}^{(1)} \odot \mathbf{M}^{(2)}$. During training, the continuous proximity mask $\mathbf{\Phi}^{(\sigma)}(\mathbf{\Delta}) \in [0,1]^m$ dynamically gates alignment, asymptotically converging as $\sigma \to 0^+$ to an oracle discrete mask $\hat{\boldsymbol{\phi}} \in \varphi(\mathbf{S})$.

Table 5: Comprehensive summary of mathematical notation, variables, and operators used throughout the paper.

Notation	Type / Domain	Description
Data Generating Process (Sec. 2)		
n	$\mathbb{N}$	Dimension of the true causal latent space
d	$\mathbb{N}$	Dimension of the empirical observation space
$\mathcal{C}, \mathcal{Z}$	$\subseteq \mathbb{R}^n$	Open domains of true causal variables and partially masked latents
$\mathcal{X}$	$\subseteq \mathbb{R}^d$	Empirical observation space
$\mathbf{C}$	Random vector	True underlying causal variables $(C_1, \ldots, C_n)^\top \in \mathcal{C}$ governed by an SCM
$\mathcal{G}_{\text{DAG}}$	Graph	Structural Directed Acyclic Graph $(\mathcal{V}, \mathcal{E})$ representing causal dependencies
$pa(j)$	Subset of $\mathcal{V}$	Set of direct parent node indices of causal variable j in $\mathcal{G}_{\text{DAG}}$
$\mathbf{M}^{(1)}, \mathbf{M}^{(2)}$	$\mathcal{M} \subseteq \{0,1\}^n$	Random binary mask vectors dictating paired instance observability patterns
$\mathbf{S}$	$\{0,1\}^n$	Shared content mask, $\mathbf{S} \coloneqq \mathbf{M}^{(1)} \odot \mathbf{M}^{(2)}$ with non-empty overlap ($\lVert\mathbf{S}\rVert_0 \geq 1$)
$\mathbf{N}^{(1)}, \mathbf{N}^{(2)}$	$\mathbb{R}^n$	Uninformative continuous nuisance noise vectors drawn independently
$\mathbf{Z}^{(1)}, \mathbf{Z}^{(2)}$	$\mathcal{Z} \subseteq \mathbb{R}^n$	Partially masked latent states, $\mathbf{M}^{(i)} \odot \mathbf{C} + (\mathbf{1} - \mathbf{M}^{(i)}) \odot \mathbf{N}^{(i)}$
f	$\mathbb{R}^n \to \mathbb{R}^d$	Smooth, invertible generative mixing function (diffeomorphism onto its image)
$\mathbf{X}^{(1)}, \mathbf{X}^{(2)}$	$\mathcal{X} \subseteq \mathbb{R}^d$	Empirical paired observations, $\mathbf{X}^{(i)} = f(\mathbf{Z}^{(i)})$
$\odot$	Operator	Element-wise (Hadamard) vector product

Continued on next page

Table 5 – *Continued from previous page*

Notation	Type / Domain	Description
Representation Learning & Identifiability (Sec. 2, Sec. B)		
m	$\mathbb{N}$	Dimension of learned representation space ($m = n$ in well-specified theory)
$g_\theta, \mathcal{G}$	Function, Class	Neural encoder with parameters $\theta \in \Theta$, belonging to hypothesis class $\mathcal{G}$
Θ	$\subset \mathbb{R}^p$	Compact parameter space of encoder weights (enforced via weight decay)
$\hat{\mathcal{Z}}$	$\subseteq \mathbb{R}^m$	Range of encoder representations, $\hat{\mathcal{Z}} := g_\theta(\mathcal{X})$
$\hat{\mathbf{Z}}^{(1)}, \hat{\mathbf{Z}}^{(2)}$	$\hat{\mathcal{Z}} \subseteq \mathbb{R}^m$	Encoded representation vectors, $\hat{\mathbf{Z}}^{(i)} = g_\theta(\mathbf{X}^{(i)})$
$\mathbf{E}$	Vector	Vector of unshared style and nuisance variables, $(\mathbf{C}_{\setminus S}, \mathbf{N})$
K, K'	Binary masks	Binary causal subvector mask $K \in \{0,1\}^n$ and latent mask $K' \in \{0,1\}^m$
Q	$\subseteq \{1, \ldots, n\}$	Identifiable index set formed by algebraic intersections and unions of basis blocks
$\mathbf{C}_K, \mathbf{C}_Q, \mathbf{C}_S$	Vectors	Causal subvectors restricted to active indices of masks $K, Q, \mathbf{S}$
ψ	Diffeomorphism	Diffeomorphism mapping causal subvectors to latent representation blocks
h	$\mathbb{R}^n \to \mathbb{R}^m$	Composite latent-to-representation map, $h := g_\theta \circ f$
$\mathbf{P}$	$\mathcal{P}_m$	Permutation matrix over representation coordinate indices
$\cap, \setminus, \cup$	Operators	Boolean intersection, relative complement, and union operations on masks
$\|\cdot\|_0$	Pseudo-norm	L_0 pseudo-norm (Hamming weight, count of non-zero entries)
Objective Formulation & BayesDisc (Sec. 3, Sec. B)		
$\boldsymbol{\Delta}$	$\mathbb{R}^m$	Pairwise representation residual vector, $\boldsymbol{\Delta} := \hat{\mathbf{Z}}^{(1)} - \hat{\mathbf{Z}}^{(2)}$
Δ_j	$\mathbb{R}$	Representation residual on the j-th latent coordinate, $\hat{Z}_j^{(1)} - \hat{Z}_j^{(2)}$
σ	$\mathbb{R}_{>0}$	Proximity mask alignment bandwidth parameter
$\boldsymbol{\Phi}^{(\sigma)}$	Function	Continuous proximity mask mapping residuals $\boldsymbol{\Delta} \in \mathbb{R}^m \to [0,1]^m$
$\Phi_j^{(\sigma)}$	Function	Coordinate gating function satisfying peak alignment and asymptotic nullification
$\varphi(\mathbf{S})$	Set	Set of oracle shared content masks, $\{\boldsymbol{\phi} \in \{0,1\}^m \mid \|\boldsymbol{\phi}\|_0 = \|\mathbf{S}\|_0\}$
$\boldsymbol{\phi}, \hat{\boldsymbol{\phi}}$	Binary	Oracle mask $\boldsymbol{\phi} \in \varphi(\mathbf{S})$ and limiting inferred mask $\lim_{\sigma \to 0^+} \boldsymbol{\Phi}^{(\sigma)}(\boldsymbol{\Delta})$
$\mathcal{L}$	Functional	Disc-InfoMax objective combining discernment and differential entropy terms
Reg	Functional	Continuous mask regularizer, $-\mathbb{E}_{\mathbf{X}}[\|\boldsymbol{\Phi}^{(\sigma)}(\boldsymbol{\Delta})\|_1]$
$\mathcal{L}_{\text{joint}}$	Functional	Joint scalarized objective function, $\mathcal{L} + \lambda \text{Reg}$
λ	$\mathbb{R}_{>0}$	Trade-off weight hyperparameter penalizing the regularizer in $\mathcal{L}_{\text{joint}}$
$\mathcal{G}_\sigma^\circ, \mathcal{G}_\sigma^*$	Sets	Inner candidate minimizer set and outer regularizer minimizer set
$\mathcal{G}_{0^+}^*, \Theta_{0^+}^*$	Sets	Asymptotic optimal encoder set and parameter set as $\sigma \to 0^+$
$H(\cdot)$	Functional	Differential entropy functional evaluated over continuous densities
$\mathcal{I}(\cdot,\cdot)$	Functional	Mutual information between random variables
$t, \mathcal{T}$	Diffeomorphism	Auxiliary bounded-Jacobian projection $t : \hat{\mathcal{Z}} \to (0,1)^m$ and class $\mathcal{T}$
$v_\sigma(\cdot)$	Function	Scalar potential function, $v_\sigma(\delta) := \delta^2 \Phi^{(\sigma)}(\delta) - \lambda \Phi^{(\sigma)}(\delta)$
$a, \neg a$	Events	Discrete events indicating coordinate j encodes shared vs. unshared content
w	$[0,1]$	Prior probability of sharing, $w := p(a) = \sigma/\sqrt{2.0}$
$p(\Delta_j)$	Density	Marginal residual density, regularized via SIGReg toward $\mathcal{N}(0, 2.0)$
$p(\Delta_j \mid a)$	Density	Conditional residual density given event a, modeled as $\mathcal{N}(0, \sigma^2)$
$\tilde{\mathbf{z}}^{(1)}, \tilde{\mathbf{z}}^{(2)}$	$\mathbb{R}^m$	Filtered representation vectors, $\tilde{\mathbf{z}}^{(i)} := \boldsymbol{\Phi} \odot \hat{\mathbf{z}}^{(i)}$ in Alg. 1
Synthetic Dataset & Empirical Setup (Sec. 4, Sec. E)		
k	$\mathbb{R}_{\geq 0}$	Expected causal DAG edge density parameter ($n \times k$ edges)
L	$\mathbb{N}$	Number of dense linear layers (depth) in generative mixing MLP
C_j	$\mathbb{R}$	Scalar value of the j-th causal variable generated via non-linear structural equation
$do(\cdot)$	Operator	Pearl's interventional do-operator for atomic interventions ($do(C_r = 2.0)$)
MCC, R^2	Metrics	Mean Correlation Coefficient (identifiability) and R^2 score (decodability)
SHD	Metric	Structural Hamming Distance measuring edge errors in recovered DAGs

B PROOFS FOR IDENTIFIABILITY OF DISC-INFOMAX

In this section, we provide the theoretical guarantees for the Disc-InfoMax objective with Regularization introduced in Sec. 3. While our problem formulation utilizes a single shared encoder and dynamically shifting, unlabeled views rather than fixed views with multiple encoders, our proof structure closely mirrors the foundational identifiability proofs established in multi-view CRL (von Kügelgen et al., 2021; Yao et al., 2023). We build upon their algebraic insights to accommodate our proximity mask formulation.

The proof is structured as follows: First, in Sec. B.1, we establish the theoretical guarantees of our Disc-InfoMax objective with regularization, proving that its asymptotic global optimum as $\sigma \to 0^+$ recovers an oracle shared content mask, achieves dynamic content-style isolation, and block-identifies the shared causal variables. Second, in Sec. B.2, we bridge the theoretical objective with practical optimization. We prove that minimizing the continuous joint objective along an annealing schedule $\sigma \to 0^+$ mathematically guarantees convergence to the asymptotic bi-level optimal set $\Theta^*_{0^+}$ and recovers the block-identified causal variables. Finally, in Sec. B.3, we establish the dynamic identifiability algebra for our setting, proving that intersection and union operations natively achieve strict block-identifiability, and analyzing why relative complements are not automatically block-identified in general.

To aid the reader, we collect the theoretical definitions and training objective required for our analysis. We restate the proximity mask and Disc-InfoMax objective from Sec. 3, and provide formal definitions for the oracle shared content mask and the auxiliary projection:

Definition 2 *(Proximity Mask).* Let $\sigma > 0$ denote a bandwidth parameter. A proximity mask $\mathbf{\Phi}^{(\sigma)} : \mathbb{R}^m \to [0,1]^m$ operates element-wise on the residual $\mathbf{\Delta}$, where each coordinate function $\Phi_j^{(\sigma)} : \mathbb{R} \to [0,1]$ is continuous on $\mathbb{R}$ and satisfies:

1. **Peak Alignment:** $\Phi_j^{(\sigma)}(0) = 1$ for any $\sigma > 0$. The gate fully activates when representations perfectly align.
2. **Monotonic Decay:** $\Phi_j^{(\sigma)}(\delta)$ is strictly monotonically decreasing with respect to $|\delta|$ on $[0, \infty)$.
3. **Asymptotic Nullification:** $\lim_{\sigma \to 0^+} \Phi_j^{(\sigma)}(\delta) = 0$ for all $\delta \neq 0$. As the bandwidth becomes infinitely narrow, the gate vanishes on all unaligned coordinates.

Definition 3 *(Oracle Shared Content Mask).* Since learned representations lack an inherent dimensional ordering, the oracle mask is not unique. For any true shared causal mask $\mathbf{S} \in \{0,1\}^n$, the *set of oracle shared content masks* $\varphi(\mathbf{S})$ is defined as:

$$\varphi(\mathbf{S}) \coloneqq \{\boldsymbol{\phi} \in \{0,1\}^m \mid \|\boldsymbol{\phi}\|_0 = \|\mathbf{S}\|_0\}. \tag{10}$$

Any element $\boldsymbol{\phi} \in \varphi(\mathbf{S})$ is referred to as an *oracle shared content mask*. Note that any such mask can be equivalently expressed as $\boldsymbol{\phi} = \mathbf{PS}$ for some permutation matrix $\mathbf{P} \in \mathcal{P}_m$.

Definition 4 *(Auxiliary Projection t).* Following multi-view CRL (Yao et al., 2023), we define $t : \hat{\mathcal{Z}} \to (0,1)^m$ as a smooth diffeomorphism from the encoder range $\hat{\mathcal{Z}}$ to the unit hypercube of the same dimension, belonging to a class $\mathcal{T}$ of smooth diffeomorphisms with bounded Jacobian determinant: $\frac{1}{L} \leq |\det J_t(\hat{\mathbf{z}})| \leq L$ for a constant $L \geq 1$. This projection accommodates arbitrary causal dependencies while mapping the latent distribution to a uniform distribution to rigorously evaluate differential entropy without dimensional collapse or arbitrary scaling artifacts.

The Disc-InfoMax objective $\mathcal{L}$ is defined as:

$$\mathcal{L}(g_\theta, t; \mathbf{\Phi}^{(\sigma)}) = \underbrace{\mathbb{E}_{\mathbf{X}}\left[\left\|\mathbf{\Phi}^{(\sigma)}(\mathbf{\Delta}) \odot \mathbf{\Delta}\right\|_2^2\right]}_{\text{Discernment}} - \underbrace{\sum\nolimits_{i \in \{1,2\}} H\left(t(\hat{\mathbf{Z}}^{(i)})\right)}_{\text{Entropy Regularization}} \tag{4}$$

$$\mathrm{Reg}(\mathbf{\Phi}^{(\sigma)}) = -\mathbb{E}_{\mathbf{X}}\left[\left\|\mathbf{\Phi}^{(\sigma)}(\mathbf{\Delta})\right\|_1\right], \tag{5}$$

with differential entropy evaluated via an auxiliary projection $t \in \mathcal{T}$ (Def. 4). By forming a bi-level optimization with the continuous regularizer $\mathrm{Reg}(\mathbf{\Phi}^{(\sigma)}) \coloneqq -\mathbb{E}_{\mathbf{X}}\left[\left\|\mathbf{\Phi}^{(\sigma)}(\mathbf{\Delta})\right\|_1\right]$:

$$\text{Inner (InfoMax):} \quad \mathcal{G}_\sigma^\circ = \operatorname*{argmin}_{g_\theta \in \mathcal{G}} \inf_{t \in \mathcal{T}} \mathcal{L}(g_\theta, t; \mathbf{\Phi}^{(\sigma)}), \tag{6}$$

$$\text{Outer (Regularization):} \quad \mathcal{G}_\sigma^* = \operatorname*{argmin}_{g_{\theta^\circ} \in \mathcal{G}_\sigma^\circ} \mathrm{Reg}(\mathbf{\Phi}^{(\sigma)}), \tag{7}$$

B.1 Properties of the Theoretical Disc-InfoMax Objective

Throughout this theoretical subsection, we analyze the well-specified setting where the encoder latent dimension matches the true causal latent dimension ($m = n$). At any finite bandwidth $\sigma > 0$, the bi-level optimization problem is:

$$\mathcal{G}_\sigma^\circ \coloneqq \operatorname*{argmin}_{g_\theta \in \mathcal{G}} \inf_{t \in \mathcal{T}} \mathcal{L}(g_\theta, t; \mathbf{\Phi}^{(\sigma)}), \tag{11}$$

$$\mathcal{G}_\sigma^* \coloneqq \operatorname*{argmin}_{g_{\theta^\circ} \in \mathcal{G}_\sigma^\circ} \mathrm{Reg}(\mathbf{\Phi}^{(\sigma)}). \tag{12}$$

The theoretical target of Disc-InfoMax is the asymptotic optimal set $\mathcal{G}_{0^+}^* \coloneqq \limsup_{\sigma \to 0^+} \mathcal{G}_\sigma^*$, defined as the set of accumulation points of optimal encoders as $\sigma \to 0^+$.

Remark *(The Encoder as the Sole Deployable Representation Model).* We emphasize that while the inner InfoMax objective utilizes an auxiliary projection $t \in \mathcal{T}$ to evaluate differential entropy and accommodate causal dependencies, the representation learning objective outputs strictly the encoder $g_{\theta^*} \in \mathcal{G}_{0^+}^*$. The outer regularizer $\mathrm{Reg}(\mathbf{\Phi}^{(\sigma)})$ and downstream evaluations depend exclusively on the encoder g_{θ^*} through representations $\hat{\mathbf{Z}} = g_{\theta^*}(\mathbf{X})$ and pairwise residuals $\mathbf{\Delta} = g_{\theta^*}(\mathbf{X}^{(1)}) - g_{\theta^*}(\mathbf{X}^{(2)})$. The auxiliary projection t is not an output of the optimization and is discarded after pre-training.

Lemma 1 *(Attained Global Minimum and Maximal Differential Entropy of Inner Minimizers).* For any bandwidth $\sigma > 0$, the inner Disc-InfoMax objective satisfies:

$$\min_{g_\theta \in \mathcal{G}} \inf_{t \in \mathcal{T}} \mathcal{L}(g_\theta, t; \mathbf{\Phi}^{(\sigma)}) \geq 0, \tag{13}$$

and in the asymptotic limit $\sigma \to 0^+$, the global minimum value across all encoders converges to zero. Consequently, for any asymptotic inner minimizer $g_{\theta^\circ} \in \limsup_{\sigma \to 0^+} \mathcal{G}_\sigma^\circ$, there exists an auxiliary projection $t \in \mathcal{T}$ under which representations achieve maximal differential entropy:

$$t(g_{\theta^\circ}(\mathbf{X})) \sim \mathrm{Uniform}((0,1)^m) \quad \text{a.s.}, \tag{14}$$

strictly preventing dimensional collapse across all m latent dimensions.

Proof. We prove the claim in three steps:

1. Universal Lower Bound of the Inner Objective (≥ 0): The inner Disc-InfoMax objective consists of the discernment loss and the negative differential entropy:

$$\mathcal{L}(g_\theta, t; \mathbf{\Phi}^{(\sigma)}) = \mathbb{E}_{\mathbf{X}}\left[\left\|\mathbf{\Phi}^{(\sigma)}(\mathbf{\Delta}) \odot \mathbf{\Delta}\right\|_2^2\right] + \sum_{i \in \{1,2\}} \Big(- H\left(t(g_\theta(\mathbf{X}^{(i)}))\right)\Big). \tag{15}$$

The discernment term is non-negative: $\left\|\mathbf{\Phi}^{(\sigma)}(\mathbf{\Delta}) \odot \mathbf{\Delta}\right\|_2^2 \geq 0$. Furthermore, because the auxiliary projection $t : \hat{\mathcal{Z}} \to (0,1)^m$ maps onto the unit hypercube of volume 1, the negative differential entropy is bounded below by zero via its equivalence to the Kullback-Leibler divergence:

$$-H(t(g_\theta(\mathbf{X}))) = \int_{(0,1)^m} p(u) \log p(u)\, du = D_{\mathrm{KL}}(p_{t(g_\theta(\mathbf{X}))} \parallel \mathrm{Uniform}((0,1)^m)) \geq 0, \tag{16}$$

with equality if and only if $t(g_\theta(\mathbf{X})) \sim \mathrm{Uniform}((0,1)^m)$ almost surely. Being the sum of two non-negative terms, $\mathcal{L}(g_\theta, t; \mathbf{\Phi}^{(\sigma)}) \geq 0$ for any encoder $g_\theta \in \mathcal{G}$, auxiliary projection $t \in \mathcal{T}$, and bandwidth $\sigma > 0$, implying $\lim_{\sigma \to 0^+} \mathcal{L}(g_\theta, t; \mathbf{\Phi}^{(\sigma)}) \geq 0$.

2. Pointwise Vanishing of Discernment and Attainability of the Minimum (0): For any fixed encoder $g_\theta \in \mathcal{G}$, consider each coordinate residual $\Delta_j = g_\theta(\mathbf{X}^{(1)})_j - g_\theta(\mathbf{X}^{(2)})_j$. On the event

$\Delta_j = 0$, $\Phi_j^{(\sigma)}(\Delta_j)\Delta_j = \Phi_j^{(\sigma)}(0) \cdot 0 = 0$ for all $\sigma > 0$. On the event $\Delta_j \neq 0$, by Property 3 of Def. 2 (Asymptotic Nullification), $\lim_{\sigma\to 0^+} \Phi_j^{(\sigma)}(\Delta_j) = 0$, yielding $\lim_{\sigma\to 0^+} \Phi_j^{(\sigma)}(\Delta_j)\Delta_j = 0$. Since $\Phi_j^{(\sigma)}(\Delta_j) \leq 1$ and $\mathbf{\Delta}$ has finite moments, the Dominated Convergence Theorem yields:

$$\lim_{\sigma\to 0^+} \mathbb{E}_{\mathbf{X}} \left[\left\| \mathbf{\Phi}^{(\sigma)}(\mathbf{\Delta}) \odot \mathbf{\Delta} \right\|_2^2 \right] = 0 \quad \forall g_\theta \in \mathcal{G}. \tag{17}$$

Next, consider the oracle encoder $g^* = f^{-1} \in \mathcal{G}$ such that $g^*(\mathbf{X}) = \mathbf{Z} \in \mathcal{Z}$. Because $\mathbf{C}$ and $\mathbf{N}$ have continuous, positive densities on open supports (Sec. 2) and f is a diffeomorphism onto its image, $\mathbf{Z}$ has an open domain $\mathcal{Z} \subseteq \mathbb{R}^m$ with continuous positive density $0 < p_{\mathbf{Z}} < \infty$. By Moser's theorem / Brenier's theorem, there exists a smooth diffeomorphism $t^* \in \mathcal{T}$ pushing forward $p_{\mathbf{Z}}$ to the uniform density on $(0,1)^m$, achieving $D_{\mathrm{KL}}(p_{t^*(g^*(\mathbf{X}))} \parallel \mathrm{Uniform}) = 0$. Therefore, the fixed candidate (g^*, t^*) achieves $\lim_{\sigma\to 0^+} \mathcal{L}(g^*, t^*; \mathbf{\Phi}^{(\sigma)}) = 0$. For any bandwidth $\sigma > 0$, the minimum value over all encoders is bounded:

$$0 \leq \min_{g_\theta \in \mathcal{G}} \inf_{t\in\mathcal{T}} \mathcal{L}(g_\theta, t; \mathbf{\Phi}^{(\sigma)}) \leq \mathcal{L}(g^*, t^*; \mathbf{\Phi}^{(\sigma)}). \tag{18}$$

Taking the limit $\sigma \to 0^+$ on both sides and applying the Squeeze Theorem proves that the global minimum value converges to exactly zero:

$$\lim_{\sigma\to 0^+} \min_{g_\theta \in \mathcal{G}} \inf_{t\in\mathcal{T}} \mathcal{L}(g_\theta, t; \mathbf{\Phi}^{(\sigma)}) = 0. \tag{19}$$

We emphasize that the limit $\lim_{\sigma\to 0^+}$ is placed outside the minimum purely via the Squeeze Theorem applied to the fixed test candidate (g^*, t^*) (whose residuals $\mathbf{\Delta}^*$ are fixed random variables independent of σ). We do not interchange the limit and the minimum, nor do we evaluate Property 3 on an encoder that adapts with σ.

3. Inner Minimization Forces Maximal Differential Entropy: For any sequence of inner minimizers $g_{\theta^\circ}^{(\sigma)} \in \mathcal{G}_\sigma^\circ$ over σ, attaining the minimum value of zero as $\sigma \to 0^+$ strictly requires minimizing the negative entropy term:

$$\lim_{\sigma\to 0^+} \inf_{t\in\mathcal{T}} D_{\mathrm{KL}}(p_{t(g_{\theta^\circ}^{(\sigma)}(\mathbf{X}))} \parallel \mathrm{Uniform}((0,1)^m)) = 0. \tag{20}$$

If an encoder collapses any latent coordinate (causing representation rank deficiency), $t(g_\theta(\mathbf{X}))$ is supported on a lower-dimensional submanifold of $(0,1)^m$, causing its differential entropy to diverge to $-\infty$ ($D_{\mathrm{KL}} = +\infty$). Any such encoder incurs an infinite loss penalty and is discarded by the argmin operation. Conversely, for any non-collapsing encoder achieving the infimum of zero, there exists a smooth diffeomorphism $t \in \mathcal{T}$ that pushes forward $p_{g_{\theta^\circ}(\mathbf{X})}$ to $\mathrm{Uniform}((0,1)^m)$, achieving $D_{\mathrm{KL}} = 0$ exactly. Thus, asymptotic inner minimizers are strictly forced to satisfy:

$$t(g_{\theta^\circ}(\mathbf{X})) \sim \mathrm{Uniform}((0,1)^m) \quad \text{a.s.}, \tag{21}$$

which achieves maximal differential entropy $H(t(g_{\theta^\circ}(\mathbf{X}))) = 0$ and strictly prevents marginal variance collapse across all m latent dimensions. □

Lemma 2 *(Representation Invariance on Shared Dimensions).* Let $g_{\theta^*} \in \mathcal{G}_{0^+}^*$ be a bi-level optimal encoder minimizing the asymptotic regularizer (Eq. (12)). For any paired observation $(\mathbf{X}^{(1)}, \mathbf{X}^{(2)})$ with true shared causal mask $\mathbf{S}$, the optimal encoder g_{θ^*} achieves pairwise representation invariance on exactly $\|\mathbf{S}\|_0$ latent coordinates corresponding to the true shared causal variables $\mathbf{C}_S$:

$$g_{\theta^*}(\mathbf{X}^{(1)})_j = g_{\theta^*}(\mathbf{X}^{(2)})_j \quad (\Delta_j = 0) \quad \text{a.s.} \tag{22}$$

for all coordinates j where the dynamically inferred mask activates ($\lim_{\sigma\to 0^+} \Phi_j^{(\sigma)}(\Delta_j) = 1$). Conversely, on all remaining $m - \|\mathbf{S}\|_0$ coordinates encoding unshared variation, representations maintain non-zero residual fluctuations:

$$|\Delta_j| > 0 \quad \text{a.s.} \tag{23}$$

causing the dynamically inferred mask to asymptotically vanish: $\lim_{\sigma\to 0^+} \Phi_j^{(\sigma)}(\Delta_j) = 0$.

Proof. We analyze coordinates encoding unshared and shared variation separately:

1. Nuisance Noise Injection and Dimension Bound on Aligned Coordinates: By the data-generating formulation (Sec. 2, Eq. (2)), for any unshared coordinate $l \notin S$, the shared content mask satisfies $S_l = M_l^{(1)} M_l^{(2)} = 0$. Thus, at least one view $i \in \{1, 2\}$ has $M_l^{(i)} = 0$, where the latent feature is replaced by continuous nuisance noise $Z_l^{(i)} = N_l^{(i)}$, drawn independently of all causal variables and of the paired view: $\mathbf{N}^{(i)} \perp\!\!\!\perp (\mathbf{C}, \mathbf{N}^{(\setminus i)})$.

Because the representations are generated by a single shared encoder g_θ via smooth composite mapping $h := g_\theta \circ f : \mathcal{Z} \to \mathbb{R}^m$, suppose an encoder coordinate h_j maintains non-trivial functional dependence on an unshared coordinate $l \notin S$. In the view i where $M_l^{(i)} = 0$, the representation $\hat{Z}_j^{(i)} = h_j(\mathbf{Z}^{(i)})$ fluctuates with the continuous independent noise $N_l^{(i)}$. Because $N_l^{(i)}$ is statistically independent of the paired observation $\mathbf{Z}^{(\setminus i)}$, the conditional distribution of $\hat{Z}_j^{(i)}$ given $\mathbf{Z}^{(\setminus i)}$ possesses a non-degenerate continuous density, yielding zero probability of exact equality: $\mathbb{P}(\Delta_j = 0 \mid \mathbf{Z}^{(\setminus i)}) = 0$, and hence $\mathbb{P}(\Delta_j = 0) = 0$.

Therefore, achieving pairwise invariance $\Delta_j = 0$ almost surely strictly requires h_j to have zero partial derivatives with respect to all unshared coordinates $l \notin S$. Consequently, any aligned coordinate $g_\theta(\mathbf{X})_j$ must depend exclusively on the shared causal variables $\mathbf{C}_S$.

Now suppose an encoder attempts to align k distinct latent coordinates, so that the subvector $(g_\theta(\mathbf{X}))_K$ depends exclusively on $\mathbf{C}_S$, where $|K| = k$. Since $\mathbf{C}_S$ has Hausdorff dimension $\|\mathbf{S}\|_0$, the subvector $(g_\theta(\mathbf{X}))_K$ is constrained to a submanifold of dimension at most $\|\mathbf{S}\|_0$. If $k > \|\mathbf{S}\|_0$, the entire latent vector $\hat{\mathbf{Z}} \in \mathbb{R}^m$ is supported on a manifold of dimension at most $\|\mathbf{S}\|_0 + (m - k) < m$. The pushforward measure $p_{t(\hat{\mathbf{Z}})}$ under any diffeomorphism $t \in \mathcal{T}$ is then singular with respect to the Lebesgue measure on $(0, 1)^m$, driving the joint differential entropy $H(t(\hat{\mathbf{Z}})) = -\infty$ and strictly violating membership in $\mathcal{G}_{0^+}^\circ$. Therefore, at most $\|\mathbf{S}\|_0$ coordinates can satisfy $\Delta_j = 0$ almost surely.

Conversely, on any of the remaining $m - \|\mathbf{S}\|_0$ coordinates, $g_\theta(\mathbf{X})_j$ must maintain non-trivial functional dependence on at least one unshared coordinate $l \notin S$ to preserve full m-dimensional representation rank and maximal differential entropy. By the noise fluctuation argument above, in the view where coordinate l is unobserved ($M_l^{(i)} = 0$), $g_\theta(\mathbf{X}^{(i)})_j$ fluctuates with the independent nuisance noise $N_l^{(i)}$, which guarantees $\mathbb{P}(\Delta_j = 0) = 0$. Hence, representations maintain non-zero fluctuations $|\Delta_j| > 0$ almost surely. Consequently, by Property 3 of Def. 2 (Asymptotic Nullification), the dynamically inferred gate satisfies:

$$\lim_{\sigma \to 0^+} \Phi_j^{(\sigma)}(\Delta_j) = 0 \quad \text{a.s.} \tag{24}$$

2. Regularizer Maximization and Exact Allocation to the Shared Block: By the Dominated Convergence Theorem and the tower property of expectation, the asymptotic regularizer is:

$$\lim_{\sigma \to 0^+} \mathrm{Reg}(\mathbf{\Phi}^{(\sigma)}) = -\lim_{\sigma \to 0^+} \mathbb{E}_{\mathbf{S}} \left[\sum_{j=1}^m \mathbb{E}_{\mathbf{X}} \left[\Phi_j^{(\sigma)}(\Delta_j) \,\middle|\, \mathbf{S} \right] \right] = -\mathbb{E}_{\mathbf{S}} \left[\sum_{j=1}^m \mathbb{P}(\Delta_j = 0 \mid \mathbf{S}) \right]. \tag{25}$$

By Step 1, for any fixed observability pattern $\mathbf{S}$, no encoder in $\mathcal{G}_{0^+}^\circ$ can align more than $\|\mathbf{S}\|_0$ coordinates, giving $\sum_{j=1}^m \mathbb{P}(\Delta_j = 0 \mid \mathbf{S}) \leq \|\mathbf{S}\|_0$. Consequently, the asymptotic regularizer is bounded from below:

$$\lim_{\sigma \to 0^+} \mathrm{Reg}(\mathbf{\Phi}^{(\sigma)}) \geq -\mathbb{E}_{\mathbf{S}} \left[\|\mathbf{S}\|_0 \right]. \tag{26}$$

Minimizing this regularizer over $\mathcal{G}_{0^+}^\circ$ is strictly equivalent to achieving this lower bound, which holds if and only if the encoder achieves the maximal capacity $\sum_{j=1}^m \mathbb{P}(\Delta_j = 0 \mid \mathbf{S}) = \|\mathbf{S}\|_0$ for every $\mathbf{S}$ in the support of the observability distribution. Because $\mathbf{C}_S$ is identical in both views ($\mathbf{C}_S^{(1)} = \mathbf{C}_S^{(2)}$), allocating a dedicated subvector of exactly $\|\mathbf{S}\|_0$ coordinates to encode a diffeomorphism of the shared causal block $\mathbf{C}_S$ achieves $\Delta_j = 0$ on all $\|\mathbf{S}\|_0$ dimensions without violating the maximal entropy condition. Within this active block, diffeomorphic coupling between variables is fully permitted by block-identifiability.

Thus, any optimal encoder $g_{\theta^*} \in \mathcal{G}_{0^+}^*$ achieves $\Delta_j = 0$ almost surely on exactly $\|\mathbf{S}\|_0$ coordinates (where $\lim_{\sigma\to 0^+} \Phi_j^{(\sigma)}(\Delta_j) = 1$), while the remaining $m - \|\mathbf{S}\|_0$ coordinates satisfy $|\Delta_j| > 0$ almost surely (where $\lim_{\sigma\to 0^+} \Phi_j^{(\sigma)}(\Delta_j) = 0$). □

Lemma 3 *(Asymptotic Recovery of an Oracle Mask).* Let $g_{\theta^*} \in \mathcal{G}_{0^+}^*$ be a bi-level optimal encoder satisfying (Eqs. (11) and (12)). In the asymptotic limit $\sigma \to 0^+$, the dynamically inferred proximity mask $\mathbf{\Phi}^{(\sigma)}(\mathbf{\Delta})$ converges almost surely to an oracle shared content mask:

$$\lim_{\sigma\to 0^+} \mathbf{\Phi}^{(\sigma)}(\mathbf{\Delta}) \in \varphi(\mathbf{S}) \quad \text{a.s.} \tag{27}$$

for any paired observation $(\mathbf{X}^{(1)}, \mathbf{X}^{(2)})$ with true shared mask $\mathbf{S}$. Moreover, the active latent coordinate allocation is globally consistent across all sample pairs sharing that pattern.

Proof. By Lem. 2, for any paired observation $(\mathbf{X}^{(1)}, \mathbf{X}^{(2)})$ with shared mask $\mathbf{S}$, the latent coordinates under the optimal encoder g_{θ^*} are partitioned into exactly $\|\mathbf{S}\|_0$ shared coordinates where $\Delta_j = 0$ almost surely, and $m - \|\mathbf{S}\|_0$ unshared coordinates where $|\Delta_j| > 0$ almost surely.

On the unshared coordinates, Property 3 of Def. 2 (Asymptotic Nullification) gives:

$$\lim_{\sigma\to 0^+} \Phi_j^{(\sigma)}(\Delta_j) = 0 \quad \text{a.s.} \tag{28}$$

On the shared coordinates, because $\Delta_j = 0$ almost surely, Property 1 of Def. 2 (Peak Alignment) gives $\Phi_j^{(\sigma)}(0) = 1$ for all $\sigma > 0$:

$$\lim_{\sigma\to 0^+} \Phi_j^{(\sigma)}(\Delta_j) = 1 \quad \text{a.s.} \tag{29}$$

Thus, the limiting dynamically inferred mask $\hat{\boldsymbol{\phi}} \coloneqq \lim_{\sigma\to 0^+} \mathbf{\Phi}^{(\sigma)}(\mathbf{\Delta})$ is a binary vector with exactly $\|\mathbf{S}\|_0$ entries equal to 1 and $m - \|\mathbf{S}\|_0$ entries equal to 0.

Furthermore, this convergence to an oracle mask holds regardless of whether the observability distribution provides fine element-wise masking or coarse block-wise masking. By Def. 3, $\varphi(\mathbf{S})$ consists of all binary vectors of Hamming weight $\|\mathbf{S}\|_0$. For any shared causal block $\mathbf{C}_S$ presented in the dataset, regularizer minimization forces the encoder to allocate a dedicated coordinate subvector of matching dimension $\|\mathbf{S}\|_0$ that aligns when and only when that block is shared, ensuring $\hat{\boldsymbol{\phi}} \in \varphi(\mathbf{S})$ across all observability structures. □

Lemma 4 *(Dynamic Content-Style Isolation under Theoretical Optimum).* Let $g_{\theta^*} \in \mathcal{G}_{0^+}^*$ be an asymptotic bi-level optimal encoder, and let $\hat{\boldsymbol{\phi}} = \lim_{\sigma\to 0^+} \mathbf{\Phi}^{(\sigma)}(\mathbf{\Delta}) \in \varphi(\mathbf{S})$ denote the recovered oracle mask. Then the selected latent subvector $\hat{\mathbf{Z}}_{\hat{\boldsymbol{\phi}}} \coloneqq (g_{\theta^*}(\mathbf{X}))_{\hat{\boldsymbol{\phi}}}$ depends exclusively on the true shared causal variables $\mathbf{C}_S$ and is invariant to the unshared style variables, defining a smooth mapping $\psi : \mathcal{C}_S \to \hat{\mathcal{Z}}_{\hat{\boldsymbol{\phi}}}$ such that $\hat{\mathbf{Z}}_{\hat{\boldsymbol{\phi}}} = \psi(\mathbf{C}_S)$.

Proof. By Lems. 2 and 3, on the active coordinates where $\hat{\phi}_j = 1$, the learned representations satisfy the invariance condition:

$$(g_{\theta^*}(\mathbf{X}^{(1)}))_{\hat{\boldsymbol{\phi}}} = (g_{\theta^*}(\mathbf{X}^{(2)}))_{\hat{\boldsymbol{\phi}}} \quad \text{a.s.} \tag{30}$$

By problem formulation (Sec. 2), the observations $\mathbf{X}$ are generated via a smooth mixing function f (a diffeomorphism onto its image) from the masked latents $\mathbf{Z}$, and the encoder g_{θ^*} is a smooth mapping on the observation space. Hence, the composite representation mapping $h \coloneqq g_{\theta^*} \circ f : \mathcal{Z} \to \hat{\mathcal{Z}}$ is smooth everywhere.

The selected latent subvector $\hat{\mathbf{Z}}_{\hat{\boldsymbol{\phi}}} = h_{\hat{\boldsymbol{\phi}}}(\mathbf{Z})$ can thus be written as a smooth function of the shared causal variables $\mathbf{C}_S$ and the unshared style variables $\mathbf{E} = (\mathbf{C}_{\setminus S}, \mathbf{N})$. By Lem. 2, any functional dependence on unshared coordinates $l \notin S$ injects independent nuisance noise variations in at least one view, strictly violating pairwise invariance. Therefore, all partial derivatives with respect to style variables vanish identically everywhere: $\frac{\partial \hat{\mathbf{Z}}_{\hat{\boldsymbol{\phi}}}}{\partial \mathbf{E}} = \mathbf{0}$.

Furthermore, $\hat{\mathbf{Z}}_{\hat{\boldsymbol{\phi}}}$ cannot be a trivial constant, because g_{θ^*} satisfies the maximal entropy condition (Lem. 1), ensuring full rank and non-zero differential entropy across all coordinates. Restricting the smooth composite mapping $h_{\hat{\boldsymbol{\phi}}}$ to $\mathbf{C}_S$ defines a smooth function $\psi : \mathcal{C}_S \to \hat{\mathcal{Z}}_{\hat{\boldsymbol{\phi}}}$ such that $\hat{\mathbf{Z}}_{\hat{\boldsymbol{\phi}}} = \psi(\mathbf{C}_S)$, which depends exclusively on $\mathbf{C}_S$ and is invariant to all unshared style variables. □

Theorem 1 *(Block-Identifiability of Disc-InfoMax).* In the asymptotic limit $\sigma \to 0^+$, any optimal encoder $g^* \in \mathcal{G}^*_{0^+} := \limsup_{\sigma \to 0^+} \mathcal{G}^*_\sigma$ achieving the theoretical bi-level Disc-InfoMax objective (Eqs. (6) and (7)) block-identifies the shared causal variables $\mathbf{C}_S$ dynamically defined for any given observability pattern.

Proof. By Lem. 4, $\hat{\mathbf{Z}}_{\hat{\boldsymbol{\phi}}} = \psi(\mathbf{C}_S)$ holds for a smooth mapping $\psi : \mathcal{C}_S \to \hat{\mathcal{Z}}_{\hat{\boldsymbol{\phi}}}$ with $\frac{\partial \hat{\mathbf{Z}}_{\hat{\boldsymbol{\phi}}}}{\partial \mathbf{E}} = \mathbf{0}$. To establish block-identifiability in the sense of Def. 1, it remains to show that ψ is a diffeomorphism.

By problem formulation (Sec. 2), the latent domain $\mathcal{Z} \subseteq \mathbb{R}^m$ is open and simply connected, the mixing function $f : \mathcal{Z} \to \mathcal{X}$ is a smooth diffeomorphism onto its image, and the encoder hypothesis class $\mathcal{G}$ consists of diffeomorphisms onto their image $\hat{\mathcal{Z}} = g_{\theta^*}(\mathcal{X})$ (von Kügelgen et al., 2021; Yao et al., 2023). Consequently, the composite representation mapping on the full latent space:

$$h \coloneqq g_{\theta^*} \circ f : \mathcal{Z} \to \hat{\mathcal{Z}} \tag{31}$$

is a global diffeomorphism from $\mathcal{Z} \subseteq \mathbb{R}^m$ onto $\hat{\mathcal{Z}} \subseteq \mathbb{R}^m$.

We now examine the structure of h partitioned into active coordinates $\hat{\boldsymbol{\phi}}$ and unshared coordinates $\setminus\hat{\boldsymbol{\phi}}$ with respect to the input variables $(\mathbf{C}_S, \mathbf{E})$. By Lem. 4, the active subvector satisfies $h_{\hat{\boldsymbol{\phi}}}(\mathbf{C}_S, \mathbf{E}) = \psi(\mathbf{C}_S)$, with $\frac{\partial h_{\hat{\boldsymbol{\phi}}}}{\partial \mathbf{E}} = \mathbf{0}$. Therefore, the Jacobian matrix of the full mapping h is block lower-triangular:

$$J_h(\mathbf{C}_S, \mathbf{E}) = \begin{pmatrix} \dfrac{\partial \psi(\mathbf{C}_S)}{\partial \mathbf{C}_S} & \mathbf{0} \\ \dfrac{\partial h_{\setminus\hat{\boldsymbol{\phi}}}(\mathbf{C}_S, \mathbf{E})}{\partial \mathbf{C}_S} & \dfrac{\partial h_{\setminus\hat{\boldsymbol{\phi}}}(\mathbf{C}_S, \mathbf{E})}{\partial \mathbf{E}} \end{pmatrix}. \tag{32}$$

Because h is a global diffeomorphism, its Jacobian determinant is non-zero everywhere on $\mathcal{Z}$:

$$\det J_h(\mathbf{C}_S, \mathbf{E}) = \det\left(\frac{\partial \psi(\mathbf{C}_S)}{\partial \mathbf{C}_S}\right) \cdot \det\left(\frac{\partial h_{\setminus\hat{\boldsymbol{\phi}}}(\mathbf{C}_S, \mathbf{E})}{\partial \mathbf{E}}\right) \neq 0. \tag{33}$$

This strictly requires the sub-Jacobian determinant to be non-zero everywhere:

$$\det\left(\frac{\partial \psi(\mathbf{C}_S)}{\partial \mathbf{C}_S}\right) \neq 0 \quad \forall \mathbf{C}_S \in \mathcal{C}_S. \tag{34}$$

By the Inverse Function Theorem, a non-singular Jacobian implies that ψ is locally invertible with a smooth local inverse around every point, making ψ a local diffeomorphism.

Finally, we elevate this local property to a global diffeomorphism. Because $h : \mathcal{Z} \to \hat{\mathcal{Z}}$ is a global diffeomorphism, it admits a smooth global inverse $h^{-1} : \hat{\mathcal{Z}} \to \mathcal{Z}$. By the Inverse Function Theorem, the Jacobian of the inverse mapping is given by $J_{h^{-1}}(\hat{\mathbf{z}}) = [J_h(h^{-1}(\hat{\mathbf{z}}))]^{-1}$. Since the inverse of an invertible block lower-triangular matrix is also block lower-triangular:

$$\begin{pmatrix} A & \mathbf{0} \\ B & C \end{pmatrix}^{-1} = \begin{pmatrix} A^{-1} & \mathbf{0} \\ -C^{-1}BA^{-1} & C^{-1} \end{pmatrix}, \tag{35}$$

the Jacobian of the inverse mapping satisfies:

$$J_{h^{-1}}(\hat{\mathbf{z}}) = \begin{pmatrix} \left(\dfrac{\partial \psi(\mathbf{C}_S)}{\partial \mathbf{C}_S}\right)^{-1} & \mathbf{0} \\ * & \left(\dfrac{\partial h_{\setminus\hat{\boldsymbol{\phi}}}(\mathbf{C}_S, \mathbf{E})}{\partial \mathbf{E}}\right)^{-1} \end{pmatrix}. \tag{36}$$

Because the upper-right block vanishes identically everywhere, the first coordinate block of the inverse mapping, $(h^{-1})_1 : \hat{\mathcal{Z}}_{\hat{\boldsymbol{\phi}}} \to \mathcal{C}_S$, depends exclusively on $\hat{\mathbf{Z}}_{\hat{\boldsymbol{\phi}}}$ with $\frac{\partial (h^{-1})_1}{\partial \hat{\mathbf{Z}}_{\setminus \hat{\boldsymbol{\phi}}}} = \mathbf{0}$. Consequently, $(h^{-1})_1$ defines a smooth two-sided global inverse for ψ: for any $\mathbf{c} \in \mathcal{C}_S$, $(h^{-1})_1(\psi(\mathbf{c})) = \mathbf{c}$, and for any $\hat{\mathbf{z}} \in \hat{\mathcal{Z}}_{\hat{\boldsymbol{\phi}}}$, $\psi((h^{-1})_1(\hat{\mathbf{z}})) = \hat{\mathbf{z}}$. Thus, $\psi : \mathcal{C}_S \to \hat{\mathcal{Z}}_{\hat{\boldsymbol{\phi}}}$ is a global diffeomorphism, establishing block-identifiability of the shared causal variables $\mathbf{C}_S$ in the sense of Def. 1. □

Remark *(Identification of All Observability Blocks).* While the content $\mathbf{C}_S$ and style $\mathbf{E}$ are dynamically defined for each specific shared mask $\mathbf{S}$, the optimal encoder g_{θ^*} is a single shared model optimized across the entire dataset. Because the identifiability guarantees in Sec. B.1 hold for any $\mathbf{S}$ in the support of the observability distribution $p(\mathbf{M}^{(1)}, \mathbf{M}^{(2)})$, the single encoder g_{θ^*} simultaneously block-identifies all causal blocks $\mathbf{C}_S$ corresponding to every observability pattern presented in the dataset.

B.2 Annealing the Continuous Joint Objective

To bridge the gap between the theoretical analysis and practical implementation, we first clarify the conceptual distinction between Sec. B.1 and the present subsection:

- **Theoretical Setting (Sec. B.1):** For any *given, fixed* bandwidth value $\sigma > 0$, the optimization is solved under a strict bi-level hierarchy: we first solve the inner InfoMax problem to obtain the candidate set $\mathcal{G}_\sigma^\circ$, and then among these candidates we solve the outer problem by minimizing the regularizer to obtain the optimal set $\mathcal{G}_\sigma^*$. If we solve this problem for each $\sigma > 0$, the asymptotic solution set as $\sigma \to 0^+$, defined by $\mathcal{G}_{0^+}^* := \limsup_{\sigma \to 0^+} \mathcal{G}_\sigma^*$, is proven in Sec. B.1 to achieve block-identifiability. We emphasize that this does *not* imply that the solution was found while annealing σ. In the theoretical formulation, σ is always a static, given value for each optimization problem, and the limit $\sigma \to 0^+$ merely shows where all of those distinct solution sets asymptotically converge. The corresponding parameter set is:
$$\Theta_{0^+}^* := \{\theta \in \Theta \mid g_\theta \in \mathcal{G}_{0^+}^*\}. \tag{37}$$
- **Practical Neural Network Training (This Subsection):** In real deep learning, optimization cannot follow the decoupled theoretical procedure above due to two practical obstacles:
  (a) *Zero Gradients at the Limit:* Directly setting $\sigma = 0$ is impossible because the gated discernment product $\boldsymbol{\Phi}^{(0^+)}(\boldsymbol{\Delta}) \odot \boldsymbol{\Delta}$ vanishes identically everywhere, yielding zero gradient signals to train the neural network.
  (b) *Intractable Nested Hierarchy:* Fully solving the inner InfoMax problem to exact convergence and subsequently optimizing over the resulting non-convex candidate manifold is computationally intractable.

  Consequently, practical neural networks optimize the inner InfoMax objective and outer regularizer *simultaneously* using a scalarized joint objective with penalty weight $\lambda > 0$:
$$\mathcal{L}_{\text{joint}}(g_\theta, t; \sigma) := \mathcal{L}(g_\theta, t; \boldsymbol{\Phi}^{(\sigma)}) + \lambda \text{Reg}(\boldsymbol{\Phi}^{(\sigma)}), \tag{38}$$
  while *simultaneously annealing* the bandwidth $\sigma_k \to 0$ dynamically across training iterations.

This raises a fundamental question: *when the inner problem, outer problem, and bandwidth annealing $\sigma_k \to 0$ all happen simultaneously in a single training run, does the resulting optimization trajectory still converge to the asymptotic bi-level optimum $\Theta_{0^+}^*$ and preserve the block-identifiability guarantee established in Sec. B.1?*

The following theorem resolves this question in the affirmative: it proves that minimizing the continuous joint objective $\mathcal{L}_{\text{joint}}$ along any annealing schedule $\sigma_k \to 0$ strictly converges to the asymptotic bi-level optimal set $\Theta_{0^+}^*$, ensuring that the practical simultaneous training trajectory achieves exact block-identifiability of the shared causal variables:

Theorem 2 *(Block Identifiability via Annealed Joint Optimization).* Let $\{\sigma_k\}_{k=1}^\infty \subset (0, \sigma_{\text{init}}]$ be any annealing schedule strictly decreasing to zero ($\lim_{k\to\infty} \sigma_k = 0$), and let $\{\theta^*(\sigma_k)\}_{k=1}^\infty$ with $\theta^*(\sigma_k) \in \Theta^*(\sigma_k)$ be a sequence of encoder parameters minimizing the continuous joint objective for each σ_k. Then, every accumulation point θ^* of this sequence achieves the asymptotic bi-level optimum ($\theta^* \in \Theta_{0^+}^*$), and the induced limiting encoder g_{θ^*} block-identifies the shared causal variables $\mathbf{C}_S$ dynamically defined for any given observability pattern.

Proof. **Formulation in the Parameter Space:** Recall from Sec. 2 that encoders $g_\theta \in \mathcal{G}$ are parameterized over a compact parameter space $\Theta \subset \mathbb{R}^p$. For any bandwidth $\sigma > 0$, let $\Theta^*(\sigma) \coloneqq \operatorname{argmin}_{\theta\in\Theta} \mathcal{L}_{\text{joint}}(g_\theta, t; \sigma)$ denote the set of global parameter minimizers. Furthermore, let $\Theta^*_{0^+} \subset \Theta$ denote the set of parameters whose induced encoders solve the asymptotic bi-level formulation (Eq. (37)). The proof proceeds in three steps:

1. Compactness and Existence of Accumulation Points: Because the encoder parameter space $\Theta \subset \mathbb{R}^p$ is compact and $\mathcal{L}_{\text{joint}}(\cdot; \sigma)$ is continuous on Θ for every finite $\sigma > 0$, the Weierstrass extreme value theorem guarantees that the minimizer set $\Theta^*(\sigma)$ is non-empty for all $\sigma > 0$. Furthermore, by compactness of Θ, any sequence of minimizers $\{\theta^*(\sigma_k)\}_{k=1}^\infty$ with $\theta^*(\sigma_k) \in \Theta^*(\sigma_k)$ possesses at least one accumulation point $\theta^* \in \Theta$ by the Bolzano-Weierstrass theorem. By continuity of the neural network parameterization $\theta \mapsto g_\theta$, every accumulation point θ^* induces a corresponding limiting encoder g_{θ^*}.

2. Collapsing Bounds and Limiting Consistency: For any finite bandwidth $\sigma > 0$ and each latent coordinate $j \in \{1, \dots, m\}$, consider the coordinate objective function:

$$v_\sigma(\Delta_j) \coloneqq \left(\Phi_j^{(\sigma)}(\Delta_j)\Delta_j\right)^2 - \lambda\Phi_j^{(\sigma)}(\Delta_j). \tag{39}$$

Because $\Phi_j^{(\sigma)}(\Delta_j) \in (0, 1]$ and $(\Phi_j^{(\sigma)}(\Delta_j)\Delta_j)^2 \geq 0$, we have $v_\sigma(\Delta_j) \geq -\lambda\Phi_j^{(\sigma)}(\Delta_j) \geq -\lambda$ for all $\sigma > 0$. The absolute global minimum $v_\sigma(\Delta_j) = -\lambda$ is attained if and only if $\Delta_j = 0$, where Peak Alignment gives $\Phi_j^{(\sigma)}(0) = 1$. Furthermore, by Lem. 1, the negative differential entropy satisfies $-\sum_{i\in\{1,2\}} H(t(g_\theta(\mathbf{X}^{(i)}))) \geq 0$ for all $\theta \in \Theta$, with equality if and only if $t(g_\theta(\mathbf{X})) \sim \text{Uniform}((0, 1)^m)$.

We now establish that for any $\sigma > 0$, the infimum value $\inf_{\theta\in\Theta} \mathcal{L}_{\text{joint}}(g_\theta, t; \sigma)$ is trapped in a bracket that collapses to $-\lambda\|\mathbf{S}\|_0$ as $\sigma \to 0^+$:

- **Universal Lower Bound for Any $\sigma > 0$:** By Lem. 2, achieving pairwise invariance $\Delta_j = 0$ a.s. requires coordinate j to depend exclusively on the shared causal variables $\mathbf{C}_S$, since dependence on unshared variables injects independent nuisance noise. Consequently, aligning more than $\|\mathbf{S}\|_0$ coordinates forces them to depend entirely on the $\|\mathbf{S}\|_0$-dimensional vector $\mathbf{C}_S$. This collapses the representation manifold dimension below m, driving the joint differential entropy to $-\infty$ and incurring an infinite loss penalty. Thus, any finite-loss encoder can achieve $\Delta_j = 0$ a.s. on at most $\|\mathbf{S}\|_0$ coordinates. On the remaining coordinates where $|\Delta_j| > 0$ a.s., $v_\sigma(\Delta_j) \geq -\lambda\Phi_j^{(\sigma)}(\Delta_j)$. Summing across all m dimensions yields:

$$\mathcal{L}_{\text{joint}}(g_\theta, t; \sigma) \geq -\lambda\|\mathbf{S}\|_0 - \lambda \sum_{j:|\Delta_j|>0} \mathbb{E}\left[\Phi_j^{(\sigma)}(\Delta_j)\right]. \tag{40}$$

  In addition, for any sequence of minimizers $\theta^*(\sigma_k) \in \Theta^*(\sigma_k)$ along $\sigma_k \to 0^+$, an encoder might theoretically attempt to adapt to σ_k by artificially shrinking the unaligned fluctuations $\Delta_j^{(k)}$ toward zero to falsely maximize the regularizer penalty $-\lambda\Phi_j^{(\sigma_k)}$. However, under the bounded Jacobian determinant condition on $\mathcal{T}$ ($|\det J_t| \leq L$ from Def. 4), the differential entropy satisfies $-H(t(\hat{\mathbf{Z}})) = -H(\hat{\mathbf{Z}}) - \mathbb{E}[\log|\det J_t|] \geq -H(\hat{\mathbf{Z}}) - \log L$. If the encoder shrinks the variance of an unshared coordinate to $O(\sigma_k^2)$, the representation entropy $-H(\hat{\mathbf{Z}})$ diverges to $+\infty$, inflicting an infinite $+\infty$ penalty via the $-H$ term in $\mathcal{L}_{\text{joint}}$. Because the regularizer gain is bounded by $-\lambda$, the infinite entropy penalty strictly prevents this pathological collapse. The variance of unshared coordinates thus remains macroscopic. With non-collapsing residual fluctuations $|\Delta_j^{(k)}| > 0$ almost surely, Property 3 of Def. 2 (Asymptotic Nullification) guarantees $\Phi_j^{(\sigma_k)}(\Delta_j^{(k)}) \to 0$ almost surely. Since $\Phi_j^{(\sigma_k)} \leq 1$ is uniformly bounded by the integrable constant 1, the Dominated Convergence Theorem yields:

$$\lim_{k\to\infty} \mathbb{E}\left[\Phi_j^{(\sigma_k)}(\Delta_j^{(k)})\right] = \mathbb{E}\left[\lim_{k\to\infty} \Phi_j^{(\sigma_k)}(\Delta_j^{(k)})\right] = 0. \tag{41}$$

  Consequently, the remainder term

$$\varepsilon(\sigma_k) \coloneqq \lambda \sum_{j:|\Delta_j^{(k)}|>0} \mathbb{E}\left[\Phi_j^{(\sigma_k)}(\Delta_j^{(k)})\right] \tag{42}$$

satisfies $\lim_{k\to\infty} \varepsilon(\sigma_k) = 0$, bounding the limit infimum:

$$\liminf_{k\to\infty} \mathcal{L}_{\text{joint}}(g_{\theta^*(\sigma_k)}, t; \sigma_k) \geq -\lambda \|\mathbf{S}\|_0. \tag{43}$$

- **Attainable Upper Bound for Any $\sigma > 0$:** By Lems. 1 and 2, there exists an ideal bi-level optimal encoder parameterized by $\theta^{\text{ideal}} \in \Theta^*_{0^+}$. Under θ^{ideal}, representations achieve pairwise invariance ($\Delta_j = 0$ a.s.) on exactly $\|\mathbf{S}\|_0$ coordinates for all sample pairs, while preserving maximal entropy ($-\sum H = 0$). On the remaining unshared coordinates, residuals fluctuate naturally ($|\Delta_j| > 0$ a.s.). Evaluating $\mathcal{L}_{\text{joint}}$ on θ^{ideal} at any $\sigma > 0$:

$$\mathcal{L}_{\text{joint}}(g_{\theta^{\text{ideal}}}, t; \sigma) = -\lambda \|\mathbf{S}\|_0 + \sum_{j:|\Delta_j|>0} \mathbb{E}\left[v_\sigma(\Delta_j^{\text{ideal}})\right]. \tag{44}$$

  Because $v_\sigma(\Delta_j^{\text{ideal}})$ is bounded and converges pointwise to 0 as $\sigma \to 0^+$, swapping the limit and the integral (expectation) via the Dominated Convergence Theorem yields $\lim_{\sigma\to 0^+} \mathbb{E}[v_\sigma(\Delta_j^{\text{ideal}})] = \mathbb{E}[\lim_{\sigma\to 0^+} v_\sigma(\Delta_j^{\text{ideal}})] = 0$, giving $\lim_{\sigma\to 0^+} \mathcal{L}_{\text{joint}}(g_{\theta^{\text{ideal}}}, t; \sigma) = -\lambda\|\mathbf{S}\|_0$. Since $\theta^*(\sigma_k)$ minimizes $\mathcal{L}_{\text{joint}}(\cdot; \sigma_k)$, $\mathcal{L}_{\text{joint}}(g_{\theta^*(\sigma_k)}, t; \sigma_k) \leq \mathcal{L}_{\text{joint}}(g_{\theta^{\text{ideal}}}, t; \sigma_k)$, implying:

$$\limsup_{k\to\infty} \mathcal{L}_{\text{joint}}(g_{\theta^*(\sigma_k)}, t; \sigma_k) \leq -\lambda \|\mathbf{S}\|_0. \tag{45}$$

Combining the matching upper and lower bounds confirms that the optimal value bracket collapses to $\lim_{k\to\infty} \mathcal{L}_{\text{joint}}(g_{\theta^*(\sigma_k)}, t; \sigma_k) = -\lambda\|\mathbf{S}\|_0$. Because $\mathcal{L}_{\text{joint}} = \mathcal{L} - \lambda\mathbb{E}[\|\mathbf{\Phi}^{(\sigma)}(\mathbf{\Delta})\|_1]$ where $\lim \mathcal{L} \geq 0$ and $\lim \mathbb{E}[\|\mathbf{\Phi}^{(\sigma)}(\mathbf{\Delta})\|_1] \leq \|\mathbf{S}\|_0$, attaining $-\lambda\|\mathbf{S}\|_0$ strictly requires:

$$\lim_{k\to\infty} \mathcal{L}(g_{\theta^*(\sigma_k)}, \mathbf{\Phi}^{(\sigma_k)}) = 0 \quad \text{and} \quad \lim_{k\to\infty} \mathbb{E}\left[\|\mathbf{\Phi}^{(\sigma_k)}(\mathbf{\Delta})\|_1\right] = \|\mathbf{S}\|_0. \tag{46}$$

Therefore, every accumulation point θ^* of the minimizer sequence achieves the zero inner InfoMax infimum ($g_{\theta^*} \in \mathcal{G}^\circ_{0^+}$) and maximizes the outer regularizer reward over $\mathcal{G}^\circ_{0^+}$, proving $g_{\theta^*} \in \mathcal{G}^*_{0^+}$ and $\theta^* \in \Theta^*_{0^+}$.

3. Conclusion (Block-Identifiability): By Step 2, every accumulation point θ^* belongs to $\Theta^*_{0^+}$. While $\Theta^*_{0^+}$ is a solution manifold containing multiple equivalent parameter configurations due to network and coordinate symmetries (meaning the global optimal encoder parameter is not unique), Sec. B.1 guarantees that *every* encoder parameterized by $\theta^* \in \Theta^*_{0^+}$ block-identifies the shared causal variables $\mathbf{C}_S$. Hence, the limiting encoder g_{θ^*} induced by any accumulation point θ^* block-identifies the shared causal variables $\mathbf{C}_S$. □

While Thm. 2 guarantees that the theoretical global optimum of the joint objective reliably recovers the bi-level optimum and block-identifies the shared variables along an annealing schedule, the empirical loss landscape of $\mathcal{L}_{\text{joint}}$ is highly non-convex. Standard gradient descent initialized from scratch is prone to getting trapped in local minima, failing to find the optimal representations. This challenge motivates our use of BayesDisc, which leverages probabilistic modeling to reliably navigate this non-convex space and successfully recover the theoretical optimum.

B.3 Dynamic Identifiability Algebra

While Thm. 2 establishes block-identifiability for the shared content $\mathbf{C}_S$ dynamically defined for any single paired observation, the overarching goal in our setting is to identify finer individual blocks by leveraging varying observability patterns across the dataset. The foundational concept of intersecting identified views was established in multi-view CRL (Yao et al., 2023). We build upon this by formulating a *Dynamic Identifiability Algebra*, establishing how intersection and union operations natively achieve strict block-identifiability on dynamically shifting subsets, and analyzing why relative complements are not automatically block-identified in general under our Disc-InfoMax objective.

Let $A, B \in \text{supp}(p(\mathbf{S}))$ be two binary masks from the support of the true shared mask distribution $\mathbf{S} = \mathbf{M}^{(1)} \odot \mathbf{M}^{(2)} \sim p(\mathbf{S})$. Since the single global optimal encoder g_{θ^*} simultaneously identifies all shared patterns across the dataset (Sec. B.1), the causal subvectors $\mathbf{C}_A$ and $\mathbf{C}_B$ are both block-identified. By Def. 1, there exist binary latent masks $A', B' \in \{0,1\}^m$ with $\|A'\|_0 = \|A\|_0$

and $\|B'\|_0 = \|B\|_0$, and diffeomorphisms $\psi_A : \mathcal{C}_A \to \hat{\mathcal{Z}}_{A'}$ and $\psi_B : \mathcal{C}_B \to \hat{\mathcal{Z}}_{B'}$ such that $\hat{\mathbf{Z}}_{A'} = \psi_A(\mathbf{C}_A)$ and $\hat{\mathbf{Z}}_{B'} = \psi_B(\mathbf{C}_B)$. In particular, the latent representations $\hat{\mathbf{Z}}_{A'}$ and $\hat{\mathbf{Z}}_{B'}$ depend exclusively on $\mathbf{C}_A$ and $\mathbf{C}_B$, respectively.

To enhance readability and maintain consistency with multi-view CRL (Yao et al., 2023), we adopt standard set notation for active coordinate supports: we write $A \cap B$ for coordinate-wise intersection ($A \odot B$), $A \setminus B$ for relative complement ($A \odot (\mathbf{1} - B)$), and $A \cup B$ for union ($A \vee B$), with analogous conventions for latent masks A', B'.

Corollary 1 *(Intersection).* The intersection of the causal variables $\mathbf{C}_{A\cap B}$ is strictly block-identified by the latent intersection mask $A' \cap B'$.

Proof. We prove the claim in three steps:

1. Exclusive Functional Dependence: By Def. 1 and Sec. B.1, $\hat{\mathbf{Z}}_{A'} = \psi_A(\mathbf{C}_A)$ and $\hat{\mathbf{Z}}_{B'} = \psi_B(\mathbf{C}_B)$ are smooth functions depending exclusively on $\mathbf{C}_A$ and $\mathbf{C}_B$, respectively. Hence, the partial derivatives with respect to inactive causal coordinates vanish identically: $\frac{\partial \hat{\mathbf{Z}}_{A'}}{\partial \mathbf{C}_{\setminus A}} = \mathbf{0}$ and $\frac{\partial \hat{\mathbf{Z}}_{B'}}{\partial \mathbf{C}_{\setminus B}} = \mathbf{0}$, where $\setminus A \coloneqq \mathbf{1} - A$. For any latent coordinate $j \in A' \cap B'$, its representation $\hat{Z}_j$ belongs to both subvectors $\hat{\mathbf{Z}}_{A'}$ and $\hat{\mathbf{Z}}_{B'}$. Thus, $\frac{\partial \hat{Z}_j}{\partial \mathbf{C}_{\setminus A}} = \mathbf{0}$ and $\frac{\partial \hat{Z}_j}{\partial \mathbf{C}_{\setminus B}} = \mathbf{0}$. By De Morgan's laws, $(\setminus A)\cup(\setminus B) = \setminus(A\cap B)$. Therefore, $\hat{\mathbf{Z}}_{A'\cap B'}$ has zero partial derivatives with respect to all extraneous causal variables $\mathbf{C}_{\setminus(A\cap B)}$. Consequently, $\hat{\mathbf{Z}}_{A'\cap B'}$ depends exclusively on $\mathbf{C}_{A\cap B}$, defining a smooth mapping:

$$\psi_\cap : \mathcal{C}_{A\cap B} \to \hat{\mathcal{Z}}_{A'\cap B'} \quad \text{such that} \quad \hat{\mathbf{Z}}_{A'\cap B'} = \psi_\cap(\mathbf{C}_{A\cap B}). \tag{47}$$

2. Sub-Jacobian Cardinality and Local Diffeomorphism: We first establish that the latent intersection mask satisfies $\|A' \cap B'\|_0 = \|A \cap B\|_0$ exactly. By Step 1, $\hat{\mathbf{Z}}_{A'\cap B'}$ depends exclusively on $\mathbf{C}_{A\cap B}$; by Lem. 2, the maximal entropy condition strictly requires $\|A' \cap B'\|_0 \le \|A \cap B\|_0$. Symmetrically, the combined subvector $\hat{\mathbf{Z}}_{A'\cup B'}$ depends exclusively on $\mathbf{C}_{A\cup B}$, forcing $\|A' \cup B'\|_0 \le \|A \cup B\|_0$. By the inclusion-exclusion principle on active coordinates, $\|A' \cup B'\|_0 = \|A'\|_0 + \|B'\|_0 - \|A' \cap B'\|_0$. Since $\|A'\|_0 = \|A\|_0$ and $\|B'\|_0 = \|B\|_0$ by Lem. 3, we have $\|A\|_0 + \|B\|_0 - \|A' \cap B'\|_0 \le \|A \cup B\|_0 = \|A\|_0 + \|B\|_0 - \|A \cap B\|_0$, which immediately implies $\|A' \cap B'\|_0 \ge \|A \cap B\|_0$. Combining both inequalities yields $\|A' \cap B'\|_0 = \|A \cap B\|_0$ and $\|A' \cup B'\|_0 = \|A \cup B\|_0$. Consequently, the sub-Jacobian $\frac{\partial \hat{\mathbf{Z}}_{A'\cap B'}}{\partial \mathbf{C}_{A\cap B}}$ is a square matrix of size $\|A \cap B\|_0 \times \|A \cap B\|_0$.

Now consider the full block diffeomorphism $\psi_A : \mathcal{C}_A \to \hat{\mathcal{Z}}_{A'}$. Decomposing $\mathbf{C}_A$ into $(\mathbf{C}_{A\cap B}, \mathbf{C}_{A\setminus B})$ and $\hat{\mathbf{Z}}_{A'}$ into $(\hat{\mathbf{Z}}_{A'\cap B'}, \hat{\mathbf{Z}}_{A'\setminus B'})$, its Jacobian J_{ψ_A} is a square matrix of size $\|A\|_0 \times \|A\|_0$. Because $\hat{\mathbf{Z}}_{A'\cap B'}$ has zero partial derivatives with respect to $\mathbf{C}_{A\setminus B}$ (since $A \setminus B \subseteq \setminus(A \cap B)$), J_{ψ_A} is strictly block lower-triangular:

$$J_{\psi_A} = \begin{pmatrix} \frac{\partial \hat{\mathbf{Z}}_{A'\cap B'}}{\partial \mathbf{C}_{A\cap B}} & \mathbf{0} \\ \frac{\partial \hat{\mathbf{Z}}_{A'\setminus B'}}{\partial \mathbf{C}_{A\cap B}} & \frac{\partial \hat{\mathbf{Z}}_{A'\setminus B'}}{\partial \mathbf{C}_{A\setminus B}} \end{pmatrix}. \tag{48}$$

Because ψ_A is a diffeomorphism (Sec. B.1), its Jacobian is everywhere non-singular: $\det(J_{\psi_A}) \neq 0$ for all $\mathbf{C}_A \in \mathcal{C}_A$. The determinant of a block lower-triangular matrix is the product of the determinants of its diagonal blocks:

$$\det(J_{\psi_A}) = \det\left(\frac{\partial \hat{\mathbf{Z}}_{A'\cap B'}}{\partial \mathbf{C}_{A\cap B}}\right) \cdot \det\left(\frac{\partial \hat{\mathbf{Z}}_{A'\setminus B'}}{\partial \mathbf{C}_{A\setminus B}}\right) \neq 0. \tag{49}$$

This strictly requires that the top-left block satisfies $\det\left(\frac{\partial \hat{\mathbf{Z}}_{A'\cap B'}}{\partial \mathbf{C}_{A\cap B}}\right) \neq 0$ everywhere on $\mathcal{C}_{A\cap B}$. By the Inverse Function Theorem, $\psi_\cap$ is a local diffeomorphism around every point.

3. Global Bijectivity and Diffeomorphism: To prove global injectivity, suppose for contradiction that there exist distinct points $\mathbf{c}_1 \neq \mathbf{c}_2 \in \mathcal{C}_{A\cap B}$ such that $\psi_\cap(\mathbf{c}_1) = \psi_\cap(\mathbf{c}_2) = \mathbf{z}^* \in \hat{\mathcal{Z}}_{A'\cap B'}$.

Because $\psi_A : \mathcal{C}_A \to \hat{\mathcal{Z}}_{A'}$ is a global diffeomorphism, it possesses a smooth global inverse $\psi_A^{-1} : \hat{\mathcal{Z}}_{A'} \to \mathcal{C}_A$. By the inverse function rule, its Jacobian is:

$$J_{\psi_A^{-1}} = (J_{\psi_A})^{-1} = \begin{pmatrix} \left(\frac{\partial \hat{\mathbf{Z}}_{A'\cap B'}}{\partial \mathbf{C}_{A\cap B}}\right)^{-1} & \mathbf{0} \\ * & \left(\frac{\partial \hat{\mathbf{Z}}_{A'\setminus B'}}{\partial \mathbf{C}_{A\setminus B}}\right)^{-1} \end{pmatrix}, \tag{50}$$

which is also block lower-triangular. Thus, $\frac{\partial (\psi_A^{-1})_{A\cap B}}{\partial \hat{\mathbf{Z}}_{A'\setminus B'}} = \mathbf{0}$. Because the domain $\hat{\mathcal{Z}}_{A'}$ is connected (as the continuous diffeomorphic image of the connected open set $\mathcal{C}_A$), the recovered coordinates $(\psi_A^{-1})_{A\cap B}$ depend exclusively on $\hat{\mathbf{Z}}_{A'\cap B'}$, defining a smooth left-inverse $\eta : \hat{\mathcal{Z}}_{A'\cap B'} \to \mathcal{C}_{A\cap B}$ such that:

$$\eta(\psi_\cap(\mathbf{c})) = (\psi_A^{-1})_{A\cap B}(\psi_A(\mathbf{c}, \mathbf{c}_\setminus)) = \mathbf{c} \tag{51}$$

for any $\mathbf{c} \in \mathcal{C}_{A\cap B}$ and any $\mathbf{c}_\setminus \in \mathcal{C}_{A\setminus B}$. Applying η to the assumption $\psi_\cap(\mathbf{c}_1) = \psi_\cap(\mathbf{c}_2)$ yields:

$$\mathbf{c}_1 = \eta(\psi_\cap(\mathbf{c}_1)) = \eta(\psi_\cap(\mathbf{c}_2)) = \mathbf{c}_2, \tag{52}$$

directly contradicting $\mathbf{c}_1 \neq \mathbf{c}_2$. Hence, $\psi_\cap$ is globally injective. Surjectivity onto the codomain $\hat{\mathcal{Z}}_{A'\cap B'} \coloneqq (g_{\theta^*}(\mathcal{X}))_{A'\cap B'}$ holds by definition of the coordinate projection of the encoder's range.

Because any bijective local diffeomorphism between connected smooth manifolds without boundary is a global diffeomorphism, $\psi_\cap : \mathcal{C}_{A\cap B} \to \hat{\mathcal{Z}}_{A'\cap B'}$ is a global diffeomorphism, successfully block-identifying the intersection $\mathbf{C}_{A\cap B}$. □

Corollary 2 *(Union).* Let $A, B \in \mathrm{supp}(p(\mathbf{S}))$ be two identified observability patterns. Then the union of the causal variables $\mathbf{C}_{A\cup B}$ is strictly block-identified by the latent union mask $A' \cup B'$.

Proof. Partition the coordinate indices of $A \cup B$ into three mutually disjoint subsets:

$$A \cup B = (A \cap B) \cup (A \setminus B) \cup (B \setminus A). \tag{53}$$

By Lem. 3, the active latent coordinate support partitions analogously into mutually disjoint subsets:

$$A' \cup B' = (A' \cap B') \cup (A' \setminus B') \cup (B' \setminus A'). \tag{54}$$

Because these subsets form a disjoint partition of $A \cup B$, the causal coordinate domain decomposes as the Cartesian product $\mathcal{C}_{A\cup B} = \mathcal{C}_{A\cap B} \times \mathcal{C}_{A\setminus B} \times \mathcal{C}_{B\setminus A}$, with total dimension $\|A \cup B\|_0 = \|A \cap B\|_0 + \|A \setminus B\|_0 + \|B \setminus A\|_0 = \|A' \cup B'\|_0$ (as established in Step 2 of Corollary 1).

We establish the diffeomorphism $\psi_\cup : \mathcal{C}_{A\cup B} \to \hat{\mathcal{Z}}_{A'\cup B'}$ in three steps:

1. Exclusive Functional Dependence: By Sec. B.1, $\hat{\mathbf{Z}}_{A'}$ depends exclusively on $\mathbf{C}_A = (\mathbf{C}_{A\cap B}, \mathbf{C}_{A\setminus B})$ and $\hat{\mathbf{Z}}_{B'}$ depends exclusively on $\mathbf{C}_B = (\mathbf{C}_{A\cap B}, \mathbf{C}_{B\setminus A})$. Since $\mathbf{C}_A \subseteq \mathbf{C}_{A\cup B}$ and $\mathbf{C}_B \subseteq \mathbf{C}_{A\cup B}$, the combined latent subvector $\hat{\mathbf{Z}}_{A'\cup B'}$ has vanishing partial derivatives with respect to all extraneous causal variables $\mathbf{C}_{\setminus(A\cup B)}$:

$$\frac{\partial \hat{\mathbf{Z}}_{A'\cup B'}}{\partial \mathbf{C}_{\setminus(A\cup B)}} = \mathbf{0}. \tag{55}$$

Thus, $\hat{\mathbf{Z}}_{A'\cup B'}$ defines a smooth mapping $\psi_\cup : \mathcal{C}_{A\cup B} \to \hat{\mathcal{Z}}_{A'\cup B'}$ depending exclusively on $\mathbf{C}_{A\cup B}$.

2. Block Lower-Triangular Jacobian and Local Diffeomorphism: By Corollary 1, $\hat{\mathbf{Z}}_{A'\cap B'}$ depends solely on $\mathbf{C}_{A\cap B}$. Furthermore, $\hat{\mathbf{Z}}_{A'\setminus B'}$ is a coordinate subvector of $\hat{\mathbf{Z}}_{A'}$, so it has zero partial derivatives with respect to $\mathbf{C}_{B\setminus A}$ (since $B \setminus A \subseteq \setminus A$). Symmetrically, $\hat{\mathbf{Z}}_{B'\setminus A'}$ is a subvector of $\hat{\mathbf{Z}}_{B'}$, so it has zero partial derivatives with respect to $\mathbf{C}_{A\setminus B}$ (since $A \setminus B \subseteq \setminus B$). Arranging the causal coordinates as $(\mathbf{C}_{A\cap B}, \mathbf{C}_{A\setminus B}, \mathbf{C}_{B\setminus A})$ and the latent coordinates as $(\hat{\mathbf{Z}}_{A'\cap B'}, \hat{\mathbf{Z}}_{A'\setminus B'}, \hat{\mathbf{Z}}_{B'\setminus A'})$, the Jacobian $J_{\psi_\cup}$ is block lower-triangular:

$$J_{\psi_\cup} = \begin{pmatrix} \frac{\partial \hat{\mathbf{Z}}_{A'\cap B'}}{\partial \mathbf{C}_{A\cap B}} & \mathbf{0} & \mathbf{0} \\ \frac{\partial \hat{\mathbf{Z}}_{A'\setminus B'}}{\partial \mathbf{C}_{A\cap B}} & \frac{\partial \hat{\mathbf{Z}}_{A'\setminus B'}}{\partial \mathbf{C}_{A\setminus B}} & \mathbf{0} \\ \frac{\partial \hat{\mathbf{Z}}_{B'\setminus A'}}{\partial \mathbf{C}_{A\cap B}} & \mathbf{0} & \frac{\partial \hat{\mathbf{Z}}_{B'\setminus A'}}{\partial \mathbf{C}_{B\setminus A}} \end{pmatrix}. \tag{56}$$

The determinant of this block lower-triangular matrix is the product of its diagonal blocks:

$$\det(J_{\psi_\cup}) = \det\left(\frac{\partial \hat{\mathbf{Z}}_{A'\cap B'}}{\partial \mathbf{C}_{A\cap B}}\right) \cdot \det\left(\frac{\partial \hat{\mathbf{Z}}_{A'\setminus B'}}{\partial \mathbf{C}_{A\setminus B}}\right) \cdot \det\left(\frac{\partial \hat{\mathbf{Z}}_{B'\setminus A'}}{\partial \mathbf{C}_{B\setminus A}}\right). \tag{57}$$

Recall from Corollary 1 that the full diffeomorphism $\psi_A : \mathcal{C}_A \to \hat{\mathcal{Z}}_{A'}$ has Jacobian determinant:

$$\det(J_{\psi_A}) = \det\left(\frac{\partial \hat{\mathbf{Z}}_{A'\cap B'}}{\partial \mathbf{C}_{A\cap B}}\right) \cdot \det\left(\frac{\partial \hat{\mathbf{Z}}_{A'\setminus B'}}{\partial \mathbf{C}_{A\setminus B}}\right) \neq 0. \tag{58}$$

Similarly, the diffeomorphism $\psi_B : \mathcal{C}_B \to \hat{\mathcal{Z}}_{B'}$ satisfies:

$$\det(J_{\psi_B}) = \det\left(\frac{\partial \hat{\mathbf{Z}}_{A'\cap B'}}{\partial \mathbf{C}_{A\cap B}}\right) \cdot \det\left(\frac{\partial \hat{\mathbf{Z}}_{B'\setminus A'}}{\partial \mathbf{C}_{B\setminus A}}\right) \neq 0. \tag{59}$$

Hence, all three diagonal sub-Jacobians are everywhere non-singular on their respective domains. It follows that $\det(J_{\psi_\cup}) \neq 0$ everywhere on $\mathcal{C}_{A\cup B}$. By the Inverse Function Theorem, $\psi_\cup$ is a local diffeomorphism around every point.

3. Global Bijectivity and Diffeomorphism: To prove global injectivity, suppose $\psi_\cup(\mathbf{c}) = \psi_\cup(\tilde{\mathbf{c}})$ for points $\mathbf{c} = (\mathbf{c}_{A\cap B}, \mathbf{c}_{A\setminus B}, \mathbf{c}_{B\setminus A})$ and $\tilde{\mathbf{c}} = (\tilde{\mathbf{c}}_{A\cap B}, \tilde{\mathbf{c}}_{A\setminus B}, \tilde{\mathbf{c}}_{B\setminus A})$ in $\mathcal{C}_{A\cup B}$. Matching coordinates:

(i) Since the subvector $\hat{\mathbf{Z}}_{A'}$ is identical under both points, the global invertibility of ψ_A yields $(\mathbf{c}_{A\cap B}, \mathbf{c}_{A\setminus B}) = \psi_A^{-1}(\hat{\mathbf{Z}}_{A'}) = (\tilde{\mathbf{c}}_{A\cap B}, \tilde{\mathbf{c}}_{A\setminus B})$, implying $\mathbf{c}_{A\cap B} = \tilde{\mathbf{c}}_{A\cap B}$ and $\mathbf{c}_{A\setminus B} = \tilde{\mathbf{c}}_{A\setminus B}$.

(ii) Similarly, since the subvector $\hat{\mathbf{Z}}_{B'}$ is identical under both points, the global invertibility of ψ_B yields $(\mathbf{c}_{A\cap B}, \mathbf{c}_{B\setminus A}) = \psi_B^{-1}(\hat{\mathbf{Z}}_{B'}) = (\tilde{\mathbf{c}}_{A\cap B}, \tilde{\mathbf{c}}_{B\setminus A})$, implying $\mathbf{c}_{B\setminus A} = \tilde{\mathbf{c}}_{B\setminus A}$.

Therefore, $\mathbf{c} = \tilde{\mathbf{c}}$, establishing that $\psi_\cup$ is globally injective. Surjectivity onto the codomain $\hat{\mathcal{Z}}_{A'\cup B'} := (g_{\theta^*}(\mathcal{X}))_{A'\cup B'}$ holds by construction.

Because $\psi_\cup$ is a bijective local diffeomorphism between connected smooth manifolds without boundary, it is a global diffeomorphism. Thus, the union $\mathbf{C}_{A\cup B}$ is strictly block-identified by $\hat{\mathbf{Z}}_{A'\cup B'}$. □

Remark *(Identifiability of Relative Complements).* While the intersection $\mathbf{C}_{A\cap B}$ and union $\mathbf{C}_{A\cup B}$ achieve strict block-identifiability, the relative complement subvector $\hat{\mathbf{Z}}_{A'\setminus B'}$ is *not automatically block-identified* in general. Specifically, $\hat{\mathbf{Z}}_{A'\setminus B'}$ is not guaranteed to be functionally independent of the shared variables $\mathbf{C}_{A\cap B}$; that is, the off-diagonal Jacobian block $\frac{\partial \hat{\mathbf{Z}}_{A'\setminus B'}}{\partial \mathbf{C}_{A\cap B}}$ may be non-zero. Furthermore, unlike multi-view CRL (Yao et al., 2023), we cannot apply private component extraction (Lyu et al., 2021). This is because we neither assume statistical independence between the true causal blocks $\mathbf{C}_{A\setminus B}$ and $\mathbf{C}_{A\cap B}$, nor do we enforce statistical independence between their corresponding learned representations $\hat{\mathbf{Z}}_{A'\setminus B'}$ and $\hat{\mathbf{Z}}_{A'\cap B'}$. Because our self-supervised learning setup operates dynamically without view labels or mask labels, it is unrealistic to enforce such representation independence or to assume strict independence among underlying causal variables. Consequently, the relative complement retains conditional dependence on the intersection and is not strictly isolated as an independent block.

C Methodology Details

C.1 Remark on Distributional Symmetry

One might question whether the strict assumption of identical marginals $p(\mathbf{X}^{(1)}) = p(\mathbf{X}^{(2)})$ limits applicability to practical setups where the two augmented views might be sampled from distinct, asymmetric augmentation distributions (e.g., $X^{(1)}$ from a geometry-heavy augmentation and $X^{(2)}$ from a color-heavy augmentation). However, this assumption introduces no practical loss of generality. By construction, our Disc-InfoMax objective is perfectly symmetric with respect to the two augmented

views. The discernment loss depends on the unoriented magnitude of the residual $\mathbf{\Delta}$, and the entropy regularization is a symmetric summation. Because the objective remains identically unchanged if the order of the pair is swapped, the effective data distribution optimized by the network is the symmetrized mixture of the two distinct augmentation distributions. Thus, we can theoretically treat the practical $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ as being identically drawn from this combined symmetrized distribution, perfectly satisfying the identical marginal assumption required to mathematically justify the use of a single shared encoder g_θ and projection t.

C.2 Connection to Practical InfoMax SSL and the Entropy Term

Our theoretical Disc-InfoMax objective natively connects to the practical implementations of standard InfoMax SSL (Balestriero & LeCun, 2025; Bardes et al., 2021). Because our problem formulation establishes that the two paired observations share an identical marginal distribution ($p(\mathbf{X}^{(1)}) = p(\mathbf{X}^{(2)}) = p(\mathbf{X})$), the entropy summation $\sum_{i\in\{1,2\}} H(t(\hat{\mathbf{Z}}^{(i)}))$ elegantly simplifies to $2H(t(\hat{\mathbf{Z}}))$. In practical self-supervised learning, this allows us to compute a single unified entropy (or variance) regularization term over the concatenated batch of both views, perfectly mirroring the standard InfoMax objective while fulfilling our theoretical requirement for a single shared projection t.

C.3 The Identifiable Index Set Q

Definition 5 ***(Identifiable Index Set Q).*** Let $Q \subseteq \{1, \dots, n\}$ be the index set of a subset of true causal variables. Q is an *identifiable index set* if it can be algebraically isolated by iteratively applying the rules of intersection and union to the collection of all initial shared basis blocks present within the dataset's support. Consequently, the block $\mathbf{C}_Q$ is guaranteed to be block-identified by the optimal encoder.

What decides Q? The granularity of Q is determined entirely by the available set of initial basis blocks. Because these basis sets are explicitly defined by the shared content mask $\mathbf{S} = \mathbf{M}^{(1)} \odot \mathbf{M}^{(2)}$, the identifiable blocks Q are purely dictated by the underlying random mask distribution $p(\mathbf{M}^{(1)}, \mathbf{M}^{(2)})$. In practice, this distribution is governed by the dataset's innate pairing process or the chosen data augmentations. To change the granularity of the identified blocks, one must change this pairing process. Under optimal conditions with sufficiently diverse observability patterns, this algebraic refinement resolves Q down to individual dimensions, achieving element-wise identifiability.

C.4 Information Capacity: Beyond Standard InfoMax SSL

According to (von Kügelgen et al., 2021, Theorem 4.4), standard InfoMax SSL is theoretically limited to identifying only the "content" variables, defined as the subset of causal variables that strictly survive the full intersection of all observability patterns across the entire dataset. In a paired partially observed setting with dynamically shifting views, standard SSL is penalized for retaining variant features, forcing it to discard any causal variable that is not universally shared across all augmentations.

In contrast, BayesDisc evaluates shared content dynamically on a per-pair basis. By capturing varying pairwise overlaps and leveraging the dynamic identifiability algebra (Sec. B.3), our method can theoretically identify up to *all* causal variables, provided the dataset contains sufficient coverage of overlapping observability patterns.

C.5 Training Algorithm Pseudocode and Computational Efficiency

Alg. 1 summarizes the end-to-end training procedure for Disc-InfoMax using the BayesDisc bi-level proxy. Since the Proximity Mask (the Bayesian posterior) is not the target we aim to optimize, its computation is detached from the gradient tree.

A major practical advantage of BayesDisc is its negligible computational footprint and minimal implementation complexity:

- **Strictly Element-Wise Analytic Form:** Evaluating the posterior proximity mask $\mathbf{\Phi}^{(\sigma)}(\mathbf{\Delta})$ amounts to computing the closed-form density ratio in Eq. (9). Since this operation is element-

wise, the computational cost is negligible compared to the most forward and backward passes of a deep neural network.

- **Seamless Drop-in Implementation:** As shown in Alg. 1, adapting standard InfoMax SSL frameworks to BayesDisc requires only adding a few lines of code into the loss calculation, integrating seamlessly into existing deep learning training pipelines.

Algorithm 1 Bi-Level Optimization of Disc-InfoMax with BayesDisc

Require: Dataset $\mathcal{D}$, encoder g_θ, initial bandwidth σ_{init}, final bandwidth σ_{final}, total epochs E.

for epoch $e = 1$ **to** E **do**
  $\sigma \leftarrow \text{Anneal}(\sigma_{\text{init}}, \sigma_{\text{final}}, e, E)$ [**Outer loop:** Regularizer penalty bounds via σ tracking]
  for batch $(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) \in \mathcal{D}$ **do**
    $\hat{\mathbf{z}}^{(1)}, \hat{\mathbf{z}}^{(2)} \leftarrow g_\theta(\mathbf{x}^{(1)}), g_\theta(\mathbf{x}^{(2)})$ [Forward pass]
    $\mathbf{\Delta} \leftarrow \hat{\mathbf{z}}^{(1)} - \hat{\mathbf{z}}^{(2)}$ [Compute differences]
    $p(\mathbf{\Delta}|a) \leftarrow \mathcal{N}(\mathbf{\Delta}|\mathbf{0}, \sigma^2\mathbf{I})$ [Compute conditional density]
    $p(\mathbf{\Delta}) \leftarrow \mathcal{N}(\mathbf{\Delta}|\mathbf{0}, 2.0\mathbf{I})$ [Compute marginal density]
    $w \leftarrow \sigma/\sqrt{2.0}$ [Prior Constraint]
    $\mathbf{\Phi} \leftarrow (p(\mathbf{\Delta}|a) \cdot w)/p(\mathbf{\Delta})$ [Infer Proximity Mask (Eq. (9))]
    $\mathbf{\Phi} \leftarrow \text{stop_gradient}(\mathbf{\Phi})$ [Detach from gradient tree]
    $\tilde{\mathbf{z}}^{(1)}, \tilde{\mathbf{z}}^{(2)} \leftarrow \mathbf{\Phi} \odot \hat{\mathbf{z}}^{(1)}, \mathbf{\Phi} \odot \hat{\mathbf{z}}^{(2)}$ [Apply Proximity Mask]
    $\mathcal{L}_{\text{joint}} \leftarrow \mathcal{L}_{\text{disc}}(\tilde{\mathbf{z}}^{(1)}, \tilde{\mathbf{z}}^{(2)}) + \lambda \cdot \mathcal{L}_{\text{entropy}}(\hat{\mathbf{z}}^{(1)}, \hat{\mathbf{z}}^{(2)})$ [Disc-InfoMax loss]
    Update θ by minimizing $\mathcal{L}_{\text{joint}}$ [**Inner loop:** Optimize encoder for current σ]
  end for
end for

D Remarks on the Alignment Bandwidth σ

D.1 Asymptotic Stability of the Bayesian Posterior

In the BayesDisc formulation, the analytic prior is defined as $w = p(a) = \sigma/\sqrt{2.0}$. At first glance, taking the theoretical limit $\sigma \to 0^+$ implies that the prior probability of sharing a feature approaches zero ($p(a) \to 0$), which seems counterintuitive. However, for successfully aligned dimensions where the residual approaches zero ($\Delta_j \to 0$), the conditional density $p(\Delta_j|a)$ behaves as $\mathcal{N}(0, \sigma^2)$, which scales by $1/\sigma$. As $\sigma \to 0^+$, this conditional density diverges to infinity. Consequently, the σ terms cancel out in the numerator of the posterior calculation, $p(\Delta_j|a) \cdot w$. This mathematical balance ensures that the inferred posterior mask $p(a|\Delta_j)$ remains stable, converging properly to 1 for perfectly aligned shared content and 0 for variant content, without diverging or collapsing.

D.2 Practical Heuristic for Tuning σ

Finding the optimal alignment bandwidth σ in practice follows a systematic empirical heuristic. Setting $\sigma = \sqrt{2.0}$, which yields a prior of $w = 1$, makes the mask fully permissive and the objective mathematically identical to vanilla SSL. Thus, we naturally use this as our starting point. From here, we initially anneal σ toward a smaller target value (e.g., 0.1) throughout the training. This initial sweep dictates our next steps: if the model trains successfully, we attempt to anneal σ even further to more closely approximate the strict theoretical limit ($\sigma \to 0^+$). Conversely, if the model collapses mid-training because the mask drops too low, becomes overly strict, and halts learning, we pinpoint the exact σ value where the failure occurred. We then set the end value of our annealing schedule safely above this collapse threshold and restart the training. Finally, once a stable terminal σ is established, we systematically try lowering the initial starting value below $\sqrt{2.0}$. While not universally effective, this adjustment can sometimes accelerate the overall training convergence.

E Numerical Dataset Details

To empirically evaluate the ability of our model to identify causal variables under controlled, non-linear mixing scenarios, we generate a synthetic numerical dataset. This dataset simulates entangled

observations where high-dimensional, dense observations are generated from a set of underlying causal variables via a complex, nonlinear generative mapping, modeling a Causal Representation Learning (CRL) problem under paired partial observability.

E.1 Structural Causal Model and DAG Generation

We first establish the ground-truth causal relationships among the n true latent variables. We model these dependencies using a Structural Causal Model (SCM) defined over a Directed Acyclic Graph (DAG), $\mathcal{G}_{\text{DAG}} = (\mathcal{V}, \mathcal{E})$, where the vertices $\mathcal{V}$ represent the n causal variables (i.e., $|\mathcal{V}| = n$) and the directed edges $\mathcal{E}$ represent direct causal influences.

The structure of the DAG and the corresponding non-linear structural equations are generated through the following procedural steps:

Graph Topology Generation. To simulate diverse causal structures, the DAG is sampled according to an Erdős-Rényi (ER) random graph model. The sparsity of the graph is controlled by a density parameter, k, such that the expected number of total edges in the graph is $n \times k$. We first generate an undirected ER graph, extract its adjacency matrix, and then apply a random acyclic orientation to ensure strict acyclicity. Finally, the node indices are randomly permuted to avoid any artificial diagonal structure in the adjacency matrix.

Non-linear Structural Equations. Once the DAG topology is fixed, the values of the causal variables $\mathbf{c} \in \mathbb{R}^n$ are generated following a topological sort of the graph. The value of each causal variable C_j is determined by a non-linear function of its direct parents in the DAG, denoted as $\mathbf{C}_{\text{pa}(j)}$, plus an independent exogenous noise term N_j:

$$C_j = \rho_j(\mathbf{C}_{\text{pa}(j)}) + N_j, \tag{60}$$

In our non-linear setting, the mechanism ρ_j is parameterized as a single-hidden-layer Multi-Layer Perceptron (MLP) with a hidden dimension of $2n$. Specifically, the mechanism takes the form:

$$\rho_j(\mathbf{x}) = \text{sigmoid}\left(\mathbf{x}\mathbf{W}_1^{(j)}\right)\mathbf{W}_2^{(j)}, \tag{61}$$

where $\text{sigmoid}(\cdot)$ denotes the logistic sigmoid activation function. To ensure robust non-linear signal propagation, the weight matrices $\mathbf{W}_1^{(j)} \in \mathbb{R}^{|\text{pa}(j)| \times 2n}$ and $\mathbf{W}_2^{(j)} \in \mathbb{R}^{2n}$ are sampled from a uniform distribution bounded between 0.5 and 2.0 in absolute value, with the sign of each weight randomly assigned with equal probability. If a node C_j has no parents (i.e., $|\text{pa}(j)| = 0$), the non-linear mechanism ρ_j evaluates to 0.

Exogenous Noise. To simulate realistic, non-Gaussian independent variations for each causal mechanism, the exogenous noise terms N_j are sampled independently from a standard Gumbel distribution (scale = 1). The resulting vector $\mathbf{c} = (C_1, \ldots, C_n)^\top$ serves as the fully realized, structurally dependent "base" causal vector for a given data sample, before any partial observability masks or the final invertible mixing function are applied.

E.2 Causal Latent Space and Generative Mapping

The data generation process begins by defining a true latent causal space of dimension n. The mapping from this underlying latent space to the corresponding observation space of dimension d is performed via a deterministic mixing function. To ensure that the generative mixing function is nonlinear, it is instantiated as a randomly initialized, invertible Multi-Layer Perceptron (MLP).

Specifically, the invertible mixing MLP is constructed with the following architectural properties:

- **Input/Output Matching:** Because the MLP must be invertible, its internal dimensionality is set to $\texttt{mlp_dim} = \max(d, n)$.
- **Network Depth and Activations:** The network consists of multiple dense linear layers (the exact depth of which is evaluated in our ablation studies) interlaced with LeakyReLU activation functions ($\alpha = 0.2$), providing a non-linear yet strictly monotonic mapping.

- **Well-Conditioned Weight Initialization:** To guarantee strict invertibility and numerical stability while maintaining complex entanglement, the dense weight matrices of each linear layer are initialized as random orthogonal matrices sampled uniformly from the orthogonal group $O(n)$. Because orthogonal matrices perfectly preserve vector norms (condition number exactly equal to 1), this initialization strictly prevents the mixing function from squashing dimensions together, thereby securing the theoretical identifiability of the generative process.
- **Frozen Parameters:** Once initialized, the parameters of the mixing function are completely frozen (`requires_grad` = False) for the dataset generation and model training phases.

E.3 Paired Generation and Partial Observability

For each data sample, we generate a paired set of views, denoted as $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$. The true underlying masked latent variables corresponding to these views, $\mathbf{Z}^{(1)}$ and $\mathbf{Z}^{(2)}$, are constructed to share only a randomly selected subset of their dimensions. This simulates a realistic scenario where only certain causal factors remain invariant across pairs while others vary.

Identifiable Mask Generation. The observation of dimensions across pairs is governed by a set of unique binary observation masks. To guarantee theoretical element-level identifiability through the dynamic identifiability algebra (Sec. B.3), these masks are iteratively generated to satisfy three strict structural conditions:

1. **Full Coverage:** Every causal dimension must be present (unmasked) in at least one observation pattern.
2. **Uniqueness:** All generated masks must be distinct.
3. **Pairwise Identifiability:** For any pair of distinct causal variables $i \neq j$, there must exist at least one mask where dimension i is observed while dimension j is masked.

Mask Sampling and Fill Values. For each sample in the dataset, a fully realized "base" causal vector $\mathbf{c} \in \mathbb{R}^n$ is generated. We then randomly sample two distinct mask vectors. The binary pattern of each chosen mask is applied to $\mathbf{c}$ to construct the masked latent views $\mathbf{Z}^{(1)}$ and $\mathbf{Z}^{(2)}$. As formalized in the main text, dimensions marked as occluded ($M_j = 0$) are replaced by uninformative noise.

In our experimental setting, we employ a *mask filling probability* of $p = 0.5$ to determine the type of nuisance noise:

- With p probability, an occluded dimension is replaced by a completely new, random sample drawn independently from the marginal distribution of that specific causal variable.
- With $1 - p$ probability, the occluded dimension is replaced by a static baseline constant of 0.

To ensure that unobserved dimensions filled with the static constant 0 are unambiguously distinguishable from the active variables, we introduce a deterministic gap in the active latent space. Before applying the static 0 fill, we push all active causal variables and randomly filled variables away from the origin by adding their sign (e.g., $+1$ for positive values, -1 for negative values). This transformation successfully empties the latent space strictly around 0, ensuring that no naturally occurring causal variable can accidentally mimic the unobserved static baseline before being passed through the frozen invertible mixing MLP.

E.4 Dataset Configuration and Ablation Setup

The complete generated numerical dataset consists of 90,112 total samples, strictly partitioned into three disjoint splits to ensure valid evaluation:

- **Training Set:** 65,536 samples
- **Validation Set:** 8,192 samples
- **Test Set:** 16,384 samples

To rigorously establish the robustness, consistency, and statistical significance of our evaluations on this benchmark, we conduct extensive ablation studies. We systematically evaluate the model across various model and dataset configurations, specifically sweeping over the causal latent dimension size (n), the causal graph density parameter (k), and the number of layers in the generative mixing MLP (L). Furthermore, all numerical experiments are independently repeated across 16 random seeds, each controlling data generation, mixing MLP weights, observability masks, and model initialization.

E.5 Evaluation Protocol and Out-of-Distribution Test Environments

Single-View and Fully Observed Testing Protocol. Standard self-supervised representation learning protocols (e.g., InfoNCE) train encoders on partial views (such as random crops or coordinate dropouts) but evaluate representations on the complete, uncropped observation. Following this established practice, our numerical benchmark distinguishes between training and testing:

- **Training and Validation:** Encoders train on paired observations with partial observability masks $\mathbf{M}^{(1)}, \mathbf{M}^{(2)} \in \{0,1\}^n$ generated to satisfy element-level coverage under the dynamic identifiability algebra (Sec. B.3).
- **Testing (Fully Observed):** All test evaluations are performed on a single observation with fully opened masks ($\mathbf{M} = \mathbf{1}$), such that $\mathbf{Z}^{(1)} = \mathbf{C}$ and $\mathbf{X}^{(1)} = f(\mathbf{C})$. Evaluating on complete observations without coordinate dropouts ensures that evaluation metrics reflect the true recovery of the underlying causal system.

Systematic Definition of the 5 Out-of-Distribution Environments. To evaluate whether learned representations capture modular causal mechanisms or merely exploit statistical correlations, we introduce 5 systematic distribution shifts generated from the underlying Structural Causal Model (SCM). The generative mixing MLP $f(\cdot)$ remains strictly identical across all environments, isolating distribution shifts entirely to the latent causal space:

1. **Cause Intervention:** Evaluates upstream interventional invariance. We deterministically select a non-isolated root node $r \in \mathcal{V}_{\text{root}} = \{v \in \mathcal{V} \mid \text{pa}(v) = \emptyset,\ \text{ch}(v) \neq \emptyset\}$ and perform an intervention by setting $C_r \leftarrow 2.0$. Downstream descendants $\text{desc}(r)$ are recomputed according to the structural equations, while non-descendants $\text{nd}(r)$ remain invariant.
2. **Effect Intervention:** Evaluates local modularity and effect isolation. We select a non-isolated leaf node $l \in \mathcal{V}_{\text{leaf}} = \{v \in \mathcal{V} \mid \text{ch}(v) = \emptyset,\ \text{pa}(v) \neq \emptyset\}$, sever its incoming edges $\text{pa}(l)$, and set $C_l \leftarrow 2.0$. All other variables C_{-l} remain strictly unchanged.
3. **Mechanism Shift:** Evaluates invariance under autonomous mechanism changes. We select an edge $r \to c$ and negate its structural weight: $W_{rc} \to -W_{rc}$. Descendants $\text{desc}(c)$ shift under the altered mechanism, while non-descendants remain invariant.
4. **Edge Addition:** Evaluates topological graph robustness. We sample two previously unconnected nodes u, v such that $v \notin \text{desc}(u)$ and add a directed edge $u \to v$ to the DAG ($B_{uv} \leftarrow 1$), introducing a new structural dependency.
5. **Support Extrapolation:** Evaluates representation disentanglement outside the compact training support. We scale the exogenous noise standard deviation for all causal nodes by a factor of two: $\text{std}(N_j) \leftarrow 2.0 \cdot \text{std}(N_j), \forall j \in \{1, \ldots, n\}$, quadrupling the variance while preserving graph topology and structural equations.

F Experimental Details

F.1 Network Architectures and Optimization

Across all experiments, our models were optimized using the AdamW optimizer with BF16 (bfloat16) mixed precision to accelerate training while preserving numerical stability. For fair comparison, the primary hyperparameter for the LeJEPA baseline (λ) was tuned and optimized for each modality.

For the image-based Causal3DIdent experiments, to perfectly map the backbone's output to our desired latent dimensionality ($m = 32$), we append a custom projection head consisting of a LeakyReLU activation followed by a Linear layer. The specific network architectures, training durations, batch

sizes, and hyperparameter schedules utilized for both the numerical and Causal3DIdent image datasets are summarized in Tab. 6.

Table 6: Architectural and optimization details for numerical and Causal3DIdent datasets.

Parameter	Numerical Dataset	Causal3DIdent (Image)
Encoder Architecture	4-Layer MLP	ResNet-18
Projection Head	None	LeakyReLU $\rightarrow$ Linear ($m = 32$)
Observation Dim (d)	64	$224 \times 224 \times 3$
Hidden Units	256	Standard ResNet dimensions
Activation Function	LeakyReLU ($\alpha = 0.2$)	ReLU (Backbone)
Optimizer	AdamW	AdamW
Batch Size	64	256
Total Epochs	16	128
Peak Learning Rate	3×10^{-4} (Constant)	5×10^{-4}
Weight Decay	5×10^{-3}	5×10^{-4}
LeJEPA Weight (λ)	0.1	0.02
NT-Xent Temperature (τ)	0.5	0.5
LR Schedule	None	Linear Warmup ($10^{-3}\times$ peak)
LR Decay	None	Cosine Annealing (min 5×10^{-7})
Mixed Precision	BF16	BF16

F.2 Causal3DIdent Dataset Details

Train/Validation Split A random 10% subset of samples from the train dataset were used as the validation split. The rest were used as the training split.

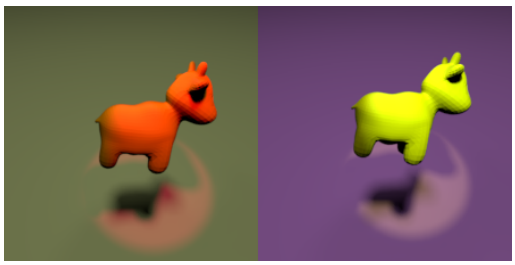


Figure 2: Sample paired observations from Causal3DIdent. Color variables vary between the pair, while other variables are invariant.

Paired Observation Generation To generate paired observations representing partial observability on the discrete Causal3DIdent dataset, we adapt the Latent Transformation (LT) approach from (von Kügelgen et al., 2021). For an original image with causal factors $\mathbf{c}$, we perturb the unshared variables by sampling from a uniform distribution on $[-1, 1]$ to yield a target vector $\mathbf{c}'$. Because Causal3DIdent is pre-rendered, we then compute L_2 distances across the dataset's latent manifold and retrieve the image corresponding to the nearest available neighbor $\mathbf{c}_{\text{nn}}$ to serve as the paired view.

F.3 Evaluation Metrics

We evaluate representations via:

- **Mean Correlation Coefficient (MCC):** Measures linear coordinate identifiability up to permutation and coordinate-wise scaling via Hungarian bipartite matching between learned latents and ground-truth causal factors.
- R^2 **Score:** Measures non-linear decodability by training an MLP regressor on frozen representations to predict ground-truth generative factors.

F.4 BayesDisc Bandwidth Annealing (σ)

As explained in Sec. 3.2.2, we adjust the bandwidth (σ) of BayesDisc to mediate the balance between the discernment objective and the continuous regularizer. Depending on the complexity, this is achieved either through a calibrated constant or a progressive annealing schedule.

Numerical Dataset Schedule For the numerical dataset, we empirically found that a constant bandwidth of $\sigma = 0.3$ is sufficient for all experiments.

Causal3DIdent Schedule For the highly complex pixel representations of the Causal3DIdent dataset, the network requires a longer duration to establish initial structural alignments. To accommodate this, we employ a slow, continuous annealing schedule across the entire 128-epoch training run. The parameter σ is initialized at $\sqrt{2.0}$ (matching the marginal prior variance, implying complete uncertainty at initialization) and is exponentially annealed down to a final value of 1.38 by the final epoch. This gradual tightening ensures steady gradient flow from the mask while gently forcing the network to perfect its invariant causal alignments.

F.5 Computational Resources

All experiments were conducted on NVIDIA A100 Tensor Core GPUs (40GB VRAM).

- **Numerical Experiments:** The lightweight MLP architecture allowed for high parallelization. We executed 16 concurrent runs (distinct random seeds) on a single GPU. Each run, on average, consumed ∼1.2GB VRAM and completed 16 epochs in approximately 15 minutes.
- **Causal3DIdent Experiments:** Processing high-dimensional images ($224 \times 224 \times 3$) via a ResNet-18 encoder required allocating one full GPU per run. Each run utilized ∼16GB VRAM and completed the 128-epoch training schedule in roughly 5 hours.

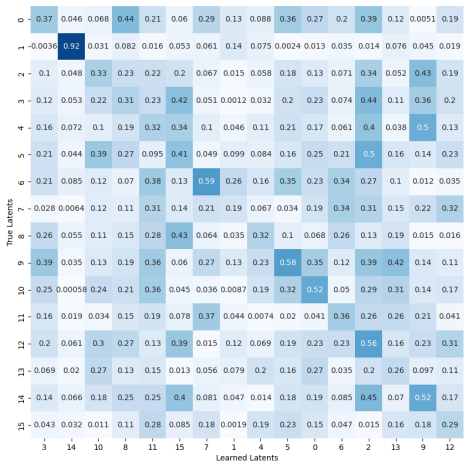
G Extra Experiment Results

G.1 Training (Masked) Distribution Numerical Results

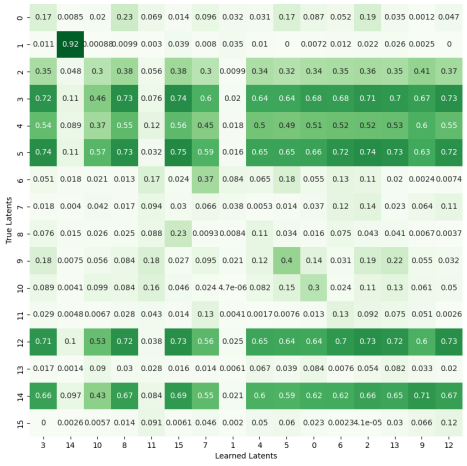
In this section, we report the evaluation of representations on the training (masked) distribution. In this regime, observations are partially observed via random coordinate masks following the identical sampling distribution used during training ($\mathbf{m} \sim \mathcal{M}_{\text{train}}$). As shown in Tab. 7, BayesDisc vastly outperforms both vanilla self-supervised baselines (InfoNCE and LeJEPA) across the configurations, achieving near-perfect identifiability (MCC ≈ 0.96–0.99 and $R^2 \approx 0.98$–0.99), except for $n = 16$ similar to the test distribution results. Significantly, BayesDisc matches the performance of the True Mask baseline (Yao et al., 2023), demonstrating that our Bayesian mask inference mechanism effectively resolves partial observability without requiring supervision.

Table 7: **Training (Masked) Distribution Numerical Results** across different parameters (MCC ↑, R^2 ↑). Default setting is $k = 1, L = 4, n = 8$. Best self-supervised results are bolded.

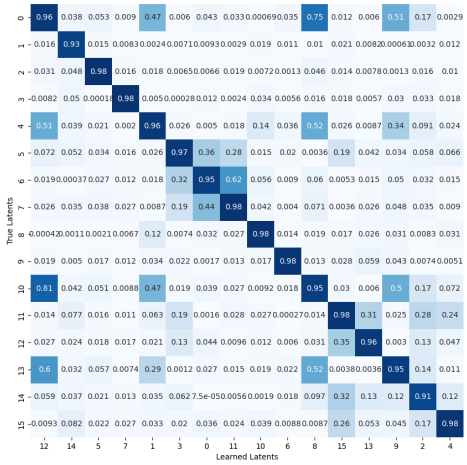
	InfoNCE		LeJEPA		BayesDisc (Ours)		True Mask	
Setting	MCC	R^2	MCC	R^2	MCC	R^2	MCC	R^2
Default	$0.343_{\pm 0.036}$	$0.621_{\pm 0.054}$	$0.351_{\pm 0.048}$	$0.636_{\pm 0.051}$	$\mathbf{0.963_{\pm 0.031}}$	$\mathbf{0.983_{\pm 0.008}}$	$0.973_{\pm 0.029}$	$0.988_{\pm 0.019}$
$k = 0$	$0.442_{\pm 0.061}$	$0.775_{\pm 0.051}$	$0.500_{\pm 0.061}$	$0.836_{\pm 0.060}$	$\mathbf{0.987_{\pm 0.001}}$	$\mathbf{0.989_{\pm 0.001}}$	$0.987_{\pm 0.001}$	$0.994_{\pm 0.000}$
$k = 2$	$0.329_{\pm 0.042}$	$0.582_{\pm 0.032}$	$0.323_{\pm 0.035}$	$0.595_{\pm 0.032}$	$\mathbf{0.957_{\pm 0.027}}$	$\mathbf{0.984_{\pm 0.003}}$	$0.976_{\pm 0.003}$	$0.992_{\pm 0.001}$
$k = 3$	$0.340_{\pm 0.051}$	$0.575_{\pm 0.054}$	$0.331_{\pm 0.054}$	$0.583_{\pm 0.062}$	$\mathbf{0.958_{\pm 0.030}}$	$\mathbf{0.982_{\pm 0.011}}$	$0.975_{\pm 0.004}$	$0.992_{\pm 0.001}$
$L = 2$	$0.346_{\pm 0.046}$	$0.627_{\pm 0.049}$	$0.354_{\pm 0.039}$	$0.637_{\pm 0.045}$	$\mathbf{0.972_{\pm 0.015}}$	$\mathbf{0.986_{\pm 0.001}}$	$0.980_{\pm 0.003}$	$0.993_{\pm 0.001}$
$L = 6$	$0.338_{\pm 0.038}$	$0.628_{\pm 0.054}$	$0.347_{\pm 0.043}$	$0.635_{\pm 0.053}$	$\mathbf{0.965_{\pm 0.034}}$	$\mathbf{0.981_{\pm 0.014}}$	$0.980_{\pm 0.003}$	$0.993_{\pm 0.001}$
$L = 8$	$0.357_{\pm 0.044}$	$0.632_{\pm 0.055}$	$0.361_{\pm 0.046}$	$0.641_{\pm 0.054}$	$\mathbf{0.973_{\pm 0.006}}$	$\mathbf{0.985_{\pm 0.001}}$	$0.978_{\pm 0.003}$	$0.992_{\pm 0.000}$
$n = 4$	$0.421_{\pm 0.108}$	$0.697_{\pm 0.062}$	$0.397_{\pm 0.078}$	$0.681_{\pm 0.060}$	$\mathbf{0.969_{\pm 0.019}}$	$\mathbf{0.990_{\pm 0.015}}$	$0.965_{\pm 0.039}$	$0.995_{\pm 0.004}$
$n = 12$	$0.349_{\pm 0.019}$	$0.615_{\pm 0.025}$	$0.342_{\pm 0.021}$	$0.624_{\pm 0.027}$	$\mathbf{0.966_{\pm 0.021}}$	$\mathbf{0.979_{\pm 0.009}}$	$0.979_{\pm 0.004}$	$0.991_{\pm 0.001}$
$n = 16$	$0.348_{\pm 0.030}$	$0.602_{\pm 0.048}$	$0.348_{\pm 0.033}$	$0.640_{\pm 0.047}$	$\mathbf{0.802_{\pm 0.163}}$	$\mathbf{0.932_{\pm 0.077}}$	$0.981_{\pm 0.005}$	$0.988_{\pm 0.001}$



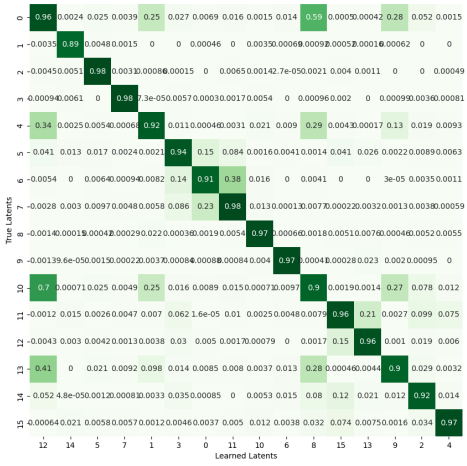
(a) Pearson Correlation Matrix with `seed=0`



(b) Element-wise R^2 with `seed=0`



(c) Pearson Correlation Matrix with `seed=12`



(d) Element-wise R^2 with `seed=12`

Figure 3: Element-wise result of numerical experiment with $n = 16$. Darker color represents higher correlation. Note that the order of learned latent variables is permuted to match the true latent variables, and non-zero off-diagonal correlations reflect the underlying causal dependencies among the true latents.

G.2 Numerical Dataset Investigations

G.2.1 Local Minima and Identification Failure

Analysis of Optimization in High Dimensions As discussed in the main text, scaling the latent dimensionality to $n = 16$ introduces severe optimization challenges. Fig. 3 visualizes the correlation matrices and component-wise R^2 scores for two identical training setups that differ only by their random initialization seeds.

Despite operating under the exact same hyperparameter configuration, the outcomes diverge drastically. With `seed=12`, BayesDisc successfully navigates the landscape, isolating and identifying all 16 true causal variables (yielding strong diagonal alignment). The non-zero off-diagonal values in this successful run are expected: because the true causal variables are generated from an SCM over a DAG, they are inherently partially dependent, and properly identified representations naturally reflect these ground-truth causal correlations. Conversely, under `seed=0`, the optimization collapses into a local minimum: the encoder successfully identifies only a single variable, leaving the remaining

15 dimensions highly entangled and poorly predictive. This discrepancy demonstrates that while BayesDisc effectively mitigates non-convexity in moderate regimes, it does not completely eliminate the risk of catastrophic local minima in highly complex, high-dimensional spaces.

G.2.2 Violation of Invertible Mixing Function Assumption

Our theoretical identifiability guarantees strictly assume that the mapping from the causal space to the observation space is a smooth diffeomorphism. To analyze what happens when this assumption breaks, we evaluate models under extreme generative complexities: extreme network depth ($L = 16$) and extreme dimensionality ($n = 32$).

To quantify the degree of invertibility, we train a supervised MLP regressor directly on the true causal variables (i.e., predicting true $\mathbf{Z}$ from observations $\mathbf{X}$); an R^2 score near 1.0 indicates that the mapping remains practically invertible. Tab. 8 provides a comparative evaluation across InfoNCE, LeJEPA, BayesDisc (Ours), the True Mask baseline (Yao et al., 2023), and the Supervised benchmark.

Table 8: **Violation of Invertible Mixing Function Assumption.** Impact of high mixing depth ($L = 16$) and latent dimensionality ($n = 32$) on representation identifiability and invertibility (Supervised regression).

	Default ($L = 4, n = 8$)		$L = 16$ (High Depth)		$n = 32$ (High Dim)	
Method	MCC	R^2	MCC	R^2	MCC	R^2
InfoNCE	$0.456_{\pm 0.039}$	$0.821_{\pm 0.055}$	$0.491_{\pm 0.035}$	$0.710_{\pm 0.037}$	$0.296_{\pm 0.015}$	$0.366_{\pm 0.028}$
LeJEPA	$0.457_{\pm 0.045}$	$0.834_{\pm 0.049}$	$0.490_{\pm 0.038}$	$0.693_{\pm 0.028}$	$0.294_{\pm 0.012}$	$0.507_{\pm 0.032}$
BayesDisc (Ours)	$0.963_{\pm 0.024}$	$0.986_{\pm 0.006}$	$0.452_{\pm 0.035}$	$0.661_{\pm 0.042}$	$0.366_{\pm 0.023}$	$0.739_{\pm 0.045}$
True Mask	$0.973_{\pm 0.031}$	$0.993_{\pm 0.002}$	$0.789_{\pm 0.031}$	$0.876_{\pm 0.020}$	$0.828_{\pm 0.028}$	$0.864_{\pm 0.021}$
Supervised	$0.999_{\pm 0.000}$	$0.998_{\pm 0.000}$	$0.881_{\pm 0.022}$	$0.913_{\pm 0.015}$	$0.985_{\pm 0.002}$	$0.974_{\pm 0.002}$

Testing the Limits of Invertibility The results in Tab. 8 demonstrate that strict invertibility is fundamental to causal identifiability. At extreme depth ($L = 16$), the generative mapping loses numerical invertibility (Supervised R^2 drops to 0.913 and MCC to 0.881). Consequently, all unsupervised methods suffer severe drops: BayesDisc achieves MCC 0.452, comparable to vanilla SSL (0.490 for LeJEPA and 0.491 for InfoNCE), while the True Mask drops to 0.789. This confirms that when invertibility is violated, unshared information cannot be reliably disentangled regardless of the selector mechanism.

At high dimensionality ($n = 32$), supervised decodability remains high ($R^2 = 0.974$), yet unsupervised identifiability collapses for all SSL methods (BayesDisc MCC drops to 0.366, LeJEPA to 0.294, InfoNCE to 0.296). In this regime, the combinatorial space of possible partial masks (2^{32}) combined with non-linear entanglement creates a severe optimization bottleneck. The True Mask baseline retains an MCC of 0.828, confirming that while information is theoretically preserved, unsupervised mask inference in high dimensions requires stronger inductive biases or additional structural constraints.

G.2.3 Full Out-of-Distribution and Interventional Sweeps

In this section, we provide the full evaluation across all 10 configurations and all 6 evaluation environments: Standard, Cause Intervention, Effect Intervention, Mechanism Shift, Edge Addition, and Support Extrapolation.

Analysis of OOD Robustness The comprehensive results across all sweeps in Tab. 9 corroborate the findings from the main text. Across varying graph densities ($k \in \{0, 1, 2, 3\}$), network depths ($L \in \{2, 4, 6, 8\}$), and latent dimensions ($n \in \{4, 8, 12\}$), BayesDisc consistently achieves high MCC (0.88–0.97) across all shifts. Remarkably, under interventional environments (Cause and Effect Interventions) and structural shifts (Mechanism Shift and Edge Addition), BayesDisc preserves its coordinate disentanglement, demonstrating genuine causal modularity. In contrast, vanilla SSL methods remain stalled at MCC 0.42–0.52 across all shifts due to representational entanglement.

Table 9: **Comprehensive Out-of-Distribution Evaluation Across All Sweeps** (MCC ↑, mean ± std across 16 seeds).

Config	Model	Standard	Cause	Effect	Mech.	Edge	Extrap.
Default	InfoNCE	$0.456_{\pm 0.039}$	$0.431_{\pm 0.038}$	$0.431_{\pm 0.045}$	$0.445_{\pm 0.051}$	$0.470_{\pm 0.047}$	$0.423_{\pm 0.037}$
	LeJEPA	$0.457_{\pm 0.045}$	$0.429_{\pm 0.035}$	$0.428_{\pm 0.041}$	$0.436_{\pm 0.046}$	$0.467_{\pm 0.050}$	$0.421_{\pm 0.034}$
	BayesDisc (Ours)	$\mathbf{0.963_{\pm 0.024}}$	$\mathbf{0.922_{\pm 0.019}}$	$\mathbf{0.900_{\pm 0.033}}$	$\mathbf{0.962_{\pm 0.024}}$	$\mathbf{0.958_{\pm 0.023}}$	$\mathbf{0.947_{\pm 0.027}}$
	True Mask	$0.973_{\pm 0.031}$	$0.857_{\pm 0.008}$	$0.852_{\pm 0.030}$	$0.973_{\pm 0.031}$	$0.972_{\pm 0.031}$	$0.962_{\pm 0.026}$
$k = 0$	InfoNCE	$0.451_{\pm 0.060}$	$0.426_{\pm 0.071}$	$0.433_{\pm 0.050}$	$0.451_{\pm 0.060}$	$0.472_{\pm 0.062}$	$0.415_{\pm 0.052}$
	LeJEPA	$0.508_{\pm 0.061}$	$0.481_{\pm 0.060}$	$0.485_{\pm 0.073}$	$0.508_{\pm 0.061}$	$0.508_{\pm 0.061}$	$0.478_{\pm 0.051}$
	BayesDisc (Ours)	$\mathbf{0.987_{\pm 0.001}}$	$\mathbf{0.939_{\pm 0.005}}$	$\mathbf{0.940_{\pm 0.007}}$	$\mathbf{0.987_{\pm 0.001}}$	$\mathbf{0.981_{\pm 0.005}}$	$\mathbf{0.969_{\pm 0.002}}$
	True Mask	$0.987_{\pm 0.000}$	$0.864_{\pm 0.000}$	$0.864_{\pm 0.000}$	$0.987_{\pm 0.000}$	$0.984_{\pm 0.003}$	$0.971_{\pm 0.001}$
$k = 2$	InfoNCE	$0.474_{\pm 0.040}$	$0.452_{\pm 0.040}$	$0.458_{\pm 0.026}$	$0.447_{\pm 0.036}$	$0.473_{\pm 0.042}$	$0.436_{\pm 0.029}$
	LeJEPA	$0.437_{\pm 0.033}$	$0.434_{\pm 0.046}$	$0.439_{\pm 0.034}$	$0.435_{\pm 0.030}$	$0.445_{\pm 0.034}$	$0.424_{\pm 0.044}$
	BayesDisc (Ours)	$\mathbf{0.955_{\pm 0.025}}$	$\mathbf{0.911_{\pm 0.028}}$	$\mathbf{0.880_{\pm 0.025}}$	$\mathbf{0.955_{\pm 0.025}}$	$\mathbf{0.952_{\pm 0.026}}$	$\mathbf{0.938_{\pm 0.026}}$
	True Mask	$0.977_{\pm 0.004}$	$0.855_{\pm 0.004}$	$0.857_{\pm 0.003}$	$0.977_{\pm 0.004}$	$0.975_{\pm 0.004}$	$0.965_{\pm 0.008}$
$k = 3$	InfoNCE	$0.486_{\pm 0.036}$	$0.484_{\pm 0.044}$	$0.458_{\pm 0.037}$	$0.480_{\pm 0.032}$	$0.497_{\pm 0.057}$	$0.458_{\pm 0.035}$
	LeJEPA	$0.460_{\pm 0.037}$	$0.463_{\pm 0.058}$	$0.440_{\pm 0.040}$	$0.455_{\pm 0.033}$	$0.470_{\pm 0.050}$	$0.421_{\pm 0.039}$
	BayesDisc (Ours)	$\mathbf{0.951_{\pm 0.026}}$	$\mathbf{0.902_{\pm 0.028}}$	$\mathbf{0.881_{\pm 0.031}}$	$\mathbf{0.950_{\pm 0.027}}$	$\mathbf{0.948_{\pm 0.026}}$	$\mathbf{0.937_{\pm 0.032}}$
	True Mask	$0.976_{\pm 0.005}$	$0.856_{\pm 0.005}$	$0.856_{\pm 0.005}$	$0.976_{\pm 0.005}$	$0.975_{\pm 0.005}$	$0.966_{\pm 0.009}$
$L = 2$	InfoNCE	$0.449_{\pm 0.048}$	$0.428_{\pm 0.045}$	$0.420_{\pm 0.047}$	$0.436_{\pm 0.054}$	$0.476_{\pm 0.049}$	$0.429_{\pm 0.045}$
	LeJEPA	$0.454_{\pm 0.035}$	$0.439_{\pm 0.051}$	$0.425_{\pm 0.027}$	$0.445_{\pm 0.038}$	$0.475_{\pm 0.047}$	$0.438_{\pm 0.035}$
	BayesDisc (Ours)	$\mathbf{0.969_{\pm 0.012}}$	$\mathbf{0.931_{\pm 0.013}}$	$\mathbf{0.904_{\pm 0.022}}$	$\mathbf{0.969_{\pm 0.012}}$	$\mathbf{0.966_{\pm 0.011}}$	$\mathbf{0.956_{\pm 0.015}}$
	True Mask	$0.981_{\pm 0.004}$	$0.859_{\pm 0.004}$	$0.860_{\pm 0.003}$	$0.981_{\pm 0.004}$	$0.979_{\pm 0.006}$	$0.970_{\pm 0.007}$
$L = 6$	InfoNCE	$0.446_{\pm 0.033}$	$0.434_{\pm 0.055}$	$0.420_{\pm 0.031}$	$0.430_{\pm 0.028}$	$0.456_{\pm 0.046}$	$0.424_{\pm 0.043}$
	LeJEPA	$0.446_{\pm 0.044}$	$0.429_{\pm 0.050}$	$0.426_{\pm 0.037}$	$0.438_{\pm 0.045}$	$0.452_{\pm 0.044}$	$0.420_{\pm 0.042}$
	BayesDisc (Ours)	$\mathbf{0.959_{\pm 0.037}}$	$\mathbf{0.921_{\pm 0.022}}$	$\mathbf{0.892_{\pm 0.036}}$	$\mathbf{0.964_{\pm 0.021}}$	$\mathbf{0.954_{\pm 0.037}}$	$\mathbf{0.946_{\pm 0.032}}$
	True Mask	$0.981_{\pm 0.004}$	$0.859_{\pm 0.004}$	$0.859_{\pm 0.003}$	$0.981_{\pm 0.004}$	$0.979_{\pm 0.005}$	$0.969_{\pm 0.007}$
$L = 8$	InfoNCE	$0.463_{\pm 0.045}$	$0.445_{\pm 0.049}$	$0.437_{\pm 0.044}$	$0.444_{\pm 0.047}$	$0.474_{\pm 0.043}$	$0.430_{\pm 0.032}$
	LeJEPA	$0.458_{\pm 0.049}$	$0.439_{\pm 0.042}$	$0.435_{\pm 0.041}$	$0.451_{\pm 0.047}$	$0.472_{\pm 0.048}$	$0.433_{\pm 0.037}$
	BayesDisc (Ours)	$\mathbf{0.968_{\pm 0.013}}$	$\mathbf{0.926_{\pm 0.008}}$	$\mathbf{0.903_{\pm 0.021}}$	$\mathbf{0.967_{\pm 0.012}}$	$\mathbf{0.963_{\pm 0.015}}$	$\mathbf{0.955_{\pm 0.014}}$
	True Mask	$0.980_{\pm 0.004}$	$0.858_{\pm 0.003}$	$0.858_{\pm 0.003}$	$0.980_{\pm 0.004}$	$0.978_{\pm 0.005}$	$0.968_{\pm 0.006}$
$n = 4$	InfoNCE	$0.516_{\pm 0.090}$	$0.478_{\pm 0.082}$	$0.467_{\pm 0.058}$	$0.482_{\pm 0.098}$	$0.521_{\pm 0.096}$	$0.430_{\pm 0.062}$
	LeJEPA	$0.489_{\pm 0.055}$	$0.462_{\pm 0.064}$	$0.465_{\pm 0.064}$	$0.456_{\pm 0.060}$	$0.495_{\pm 0.065}$	$0.459_{\pm 0.057}$
	BayesDisc (Ours)	$\mathbf{0.965_{\pm 0.025}}$	$\mathbf{0.906_{\pm 0.030}}$	$\mathbf{0.875_{\pm 0.057}}$	$\mathbf{0.965_{\pm 0.022}}$	$\mathbf{0.963_{\pm 0.023}}$	$\mathbf{0.943_{\pm 0.029}}$
	True Mask	$0.974_{\pm 0.032}$	$0.731_{\pm 0.015}$	$0.739_{\pm 0.006}$	$0.973_{\pm 0.035}$	$0.973_{\pm 0.031}$	$0.964_{\pm 0.022}$
$n = 12$	InfoNCE	$0.465_{\pm 0.027}$	$0.441_{\pm 0.034}$	$0.452_{\pm 0.030}$	$0.455_{\pm 0.033}$	$0.471_{\pm 0.028}$	$0.436_{\pm 0.027}$
	LeJEPA	$0.445_{\pm 0.025}$	$0.428_{\pm 0.033}$	$0.435_{\pm 0.021}$	$0.433_{\pm 0.021}$	$0.450_{\pm 0.025}$	$0.427_{\pm 0.024}$
	BayesDisc (Ours)	$\mathbf{0.961_{\pm 0.021}}$	$\mathbf{0.927_{\pm 0.017}}$	$\mathbf{0.915_{\pm 0.026}}$	$\mathbf{0.960_{\pm 0.023}}$	$\mathbf{0.959_{\pm 0.022}}$	$\mathbf{0.952_{\pm 0.018}}$
	True Mask	$0.981_{\pm 0.002}$	$0.899_{\pm 0.002}$	$0.900_{\pm 0.002}$	$0.981_{\pm 0.002}$	$0.981_{\pm 0.003}$	$0.973_{\pm 0.002}$
$n = 16$	InfoNCE	$0.470_{\pm 0.037}$	$0.457_{\pm 0.035}$	$0.451_{\pm 0.033}$	$0.457_{\pm 0.034}$	$0.474_{\pm 0.038}$	$0.433_{\pm 0.032}$
	LeJEPA	$0.447_{\pm 0.021}$	$0.440_{\pm 0.024}$	$0.436_{\pm 0.021}$	$0.438_{\pm 0.027}$	$0.451_{\pm 0.025}$	$0.421_{\pm 0.022}$
	BayesDisc (Ours)	$\mathbf{0.799_{\pm 0.150}}$	$\mathbf{0.779_{\pm 0.139}}$	$\mathbf{0.775_{\pm 0.135}}$	$\mathbf{0.803_{\pm 0.149}}$	$\mathbf{0.795_{\pm 0.150}}$	$\mathbf{0.792_{\pm 0.150}}$
	True Mask	$0.982_{\pm 0.003}$	$0.920_{\pm 0.003}$	$0.920_{\pm 0.003}$	$0.982_{\pm 0.003}$	$0.980_{\pm 0.005}$	$0.976_{\pm 0.002}$

G.2.4 Full Downstream Causal Discovery Sweeps

To evaluate whether identified causal representations enable downstream causal structure discovery, we apply three standard algorithms (LiNGAM, PC, GES) to the learned representations across all 10 configurations. We compare the Structural Hamming Distance (SHD) against the true causal DAG.

Table 10: **Structural Hamming Distance** (SHD ↓, mean ± std across 16 seeds) for downstream causal discovery across all numerical configurations on fully observed test set. Best self-supervised results are bolded.

Config	Model	LiNGAM (SHD ↓)	PC (SHD ↓)	GES (SHD ↓)
Default	SSL (LeJEPA)	$20.25_{\pm 8.56}$	$19.94_{\pm 9.31}$	$21.62_{\pm 15.61}$
	BayesDisc (Ours)	$\mathbf{15.75_{\pm 5.44}}$	$\mathbf{14.50_{\pm 4.75}}$	$\mathbf{16.38_{\pm 5.98}}$
	True Mask	$13.62_{\pm 6.11}$	$13.56_{\pm 12.00}$	$12.69_{\pm 13.46}$
	True Causal Variables	$8.50_{\pm 9.25}$	$7.31_{\pm 4.34}$	$7.19_{\pm 7.78}$
$k = 0$	SSL (LeJEPA)	$17.06_{\pm 5.18}$	$19.81_{\pm 3.40}$	$17.69_{\pm 9.46}$
	BayesDisc (Ours)	$\mathbf{12.00_{\pm 6.12}}$	$\mathbf{17.25_{\pm 6.56}}$	$\mathbf{13.19_{\pm 6.78}}$
	True Mask	$11.50_{\pm 7.62}$	$17.06_{\pm 6.43}$	$12.56_{\pm 6.62}$
	True Causal Variables	$0.00_{\pm 0.00}$	$4.00_{\pm 0.00}$	$0.00_{\pm 0.00}$
$k = 2$	SSL (LeJEPA)	$20.19_{\pm 7.15}$	$18.88_{\pm 4.23}$	$20.12_{\pm 15.61}$
	BayesDisc (Ours)	$\mathbf{15.38_{\pm 15.23}}$	$\mathbf{13.94_{\pm 7.06}}$	$\mathbf{16.56_{\pm 8.00}}$
	True Mask	$14.50_{\pm 13.25}$	$13.12_{\pm 9.48}$	$16.00_{\pm 7.88}$
	True Causal Variables	$13.31_{\pm 22.71}$	$11.25_{\pm 8.56}$	$14.12_{\pm 23.98}$
$k = 3$	SSL (LeJEPA)	$16.62_{\pm 5.61}$	$18.00_{\pm 14.38}$	$19.50_{\pm 6.12}$
	BayesDisc (Ours)	$\mathbf{14.50_{\pm 14.62}}$	$\mathbf{16.06_{\pm 5.18}}$	$\mathbf{17.19_{\pm 9.78}}$
	True Mask	$13.50_{\pm 12.12}$	$17.00_{\pm 15.88}$	$16.81_{\pm 14.03}$
	True Causal Variables	$10.44_{\pm 16.87}$	$16.88_{\pm 9.48}$	$17.00_{\pm 24.75}$
$L = 2$	SSL (LeJEPA)	$20.94_{\pm 4.43}$	$20.81_{\pm 3.90}$	$22.44_{\pm 6.75}$
	BayesDisc (Ours)	$\mathbf{16.06_{\pm 19.31}}$	$\mathbf{13.81_{\pm 6.53}}$	$\mathbf{15.88_{\pm 16.11}}$
	True Mask	$13.12_{\pm 8.86}$	$11.44_{\pm 4.62}$	$12.69_{\pm 5.09}$
	True Causal Variables	$8.50_{\pm 9.25}$	$7.31_{\pm 4.34}$	$7.19_{\pm 7.78}$
$L = 6$	SSL (LeJEPA)	$20.50_{\pm 4.00}$	$21.62_{\pm 3.86}$	$21.50_{\pm 9.38}$
	BayesDisc (Ours)	$\mathbf{14.25_{\pm 7.56}}$	$\mathbf{14.12_{\pm 11.36}}$	$\mathbf{15.06_{\pm 11.43}}$
	True Mask	$13.81_{\pm 7.15}$	$12.12_{\pm 3.73}$	$13.69_{\pm 6.21}$
	True Causal Variables	$8.50_{\pm 9.25}$	$7.31_{\pm 4.34}$	$7.19_{\pm 7.78}$
$L = 8$	SSL (LeJEPA)	$19.88_{\pm 5.86}$	$19.50_{\pm 5.00}$	$21.44_{\pm 10.00}$
	BayesDisc (Ours)	$\mathbf{16.25_{\pm 6.06}}$	$\mathbf{14.06_{\pm 7.93}}$	$\mathbf{16.38_{\pm 4.61}}$
	True Mask	$16.44_{\pm 10.25}$	$13.69_{\pm 2.71}$	$16.56_{\pm 8.87}$
	True Causal Variables	$8.50_{\pm 9.25}$	$7.31_{\pm 4.34}$	$7.19_{\pm 7.78}$
$n = 4$	SSL (LeJEPA)	$4.12_{\pm 1.36}$	$5.62_{\pm 3.23}$	$5.75_{\pm 2.31}$
	BayesDisc (Ours)	$\mathbf{3.00_{\pm 2.00}}$	$\mathbf{4.44_{\pm 3.00}}$	$\mathbf{4.56_{\pm 1.75}}$
	True Mask	$2.31_{\pm 2.09}$	$4.50_{\pm 4.88}$	$4.12_{\pm 3.36}$
	True Causal Variables	$3.06_{\pm 3.43}$	$4.31_{\pm 7.96}$	$3.81_{\pm 7.40}$
$n = 12$	SSL (LeJEPA)	$51.00_{\pm 11.62}$	$45.12_{\pm 13.48}$	$49.88_{\pm 22.61}$
	BayesDisc (Ours)	$\mathbf{36.31_{\pm 44.21}}$	$\mathbf{28.44_{\pm 12.12}}$	$\mathbf{32.88_{\pm 27.48}}$
	True Mask	$33.88_{\pm 23.86}$	$25.06_{\pm 9.81}$	$28.31_{\pm 23.96}$
	True Causal Variables	$17.69_{\pm 47.71}$	$12.25_{\pm 13.06}$	$10.88_{\pm 11.86}$
$n = 16$	SSL (LeJEPA)	$95.81_{\pm 12.53}$	$69.06_{\pm 23.06}$	$92.75_{\pm 41.69}$
	BayesDisc (Ours)	$\mathbf{81.38_{\pm 131.98}}$	$\mathbf{53.88_{\pm 88.86}}$	$\mathbf{74.12_{\pm 257.11}}$
	True Mask	$57.62_{\pm 68.98}$	$38.62_{\pm 34.61}$	$47.94_{\pm 58.68}$
	True Causal Variables	$24.94_{\pm 52.81}$	$15.81_{\pm 17.90}$	$16.00_{\pm 30.00}$

Analysis of Causal Discovery Trends As shown in Tab. 10, the relative performance across methods exhibits the same consistent tendency across all three discovery algorithms (LiNGAM, PC, and GES). BayesDisc consistently outperforms the vanilla SSL baseline across all configurations. Since BayesDisc identifies the underlying causal variables much better, this confirms that it learns not only individual variables but also preserves their inter-variable causal dependencies to a meaningful

degree. While BayesDisc is generally slightly behind the True Mask baseline, its performance is comparable and in some cases even superior (e.g., on $k = 3$ with PC), which is a strong result given that BayesDisc completely drops the view-label requirement.

At the same time, evaluating causal discovery directly on the true causal variables yields substantially better results across all settings, though even on the true variables the results are not perfect due to inherent statistical testing errors and finite-sample limitations. Nonetheless, the gap indicates that even minor identification defects in the representation stage can have a magnified impact on downstream causal discovery. Importantly, this gap is not unique to BayesDisc; the True Mask baseline suffers a similar drop compared to the true variables. Finally, as the latent dimension n increases (most visibly at $n = 16$), the standard deviation across runs explodes. This high variance is directly explained by the local minima and identification failures explored in Sec. G.2.1, where certain random initializations fail to identify all latent factors.

G.3 CAUSAL3DIDENT DATASET

Table 11: Element-wise Correlation and R^2 on Causal3DIdent. Best values are bolded.

Method	Metric	X	Y	Z	α	β	γ
BayesDisc (Ours)	Corr.	$0.444_{\pm 0.021}$	$\mathbf{0.637_{\pm 0.051}}$	$\mathbf{0.949_{\pm 0.049}}$	$0.546_{\pm 0.108}$	$\mathbf{0.613_{\pm 0.084}}$	$0.563_{\pm 0.105}$
	R^2	$0.328_{\pm 0.058}$	$\mathbf{0.453_{\pm 0.054}}$	$\mathbf{0.914_{\pm 0.083}}$	$\mathbf{0.360_{\pm 0.143}}$	$\mathbf{0.482_{\pm 0.132}}$	$\mathbf{0.372_{\pm 0.130}}$
LeJEPA	Corr.	$0.580_{\pm 0.039}$	$0.546_{\pm 0.062}$	$0.454_{\pm 0.072}$	$0.529_{\pm 0.060}$	$0.513_{\pm 0.051}$	$0.510_{\pm 0.066}$
	R^2	$0.353_{\pm 0.051}$	$0.313_{\pm 0.067}$	$0.214_{\pm 0.074}$	$0.293_{\pm 0.068}$	$0.275_{\pm 0.053}$	$0.270_{\pm 0.069}$
InfoNCE	Corr.	$\mathbf{0.646_{\pm 0.075}}$	$0.606_{\pm 0.081}$	$0.458_{\pm 0.075}$	$\mathbf{0.572_{\pm 0.053}}$	$0.569_{\pm 0.064}$	$\mathbf{0.581_{\pm 0.057}}$
	R^2	$\mathbf{0.442_{\pm 0.101}}$	$0.385_{\pm 0.106}$	$0.217_{\pm 0.075}$	$0.336_{\pm 0.063}$	$0.332_{\pm 0.077}$	$0.348_{\pm 0.065}$

Method	Metric	Spot P.	Obj C.	Spot C.	BG C.	Shape
BayesDisc (Ours)	Corr.	$\mathbf{0.637_{\pm 0.051}}$	$\mathbf{0.642_{\pm 0.107}}$	$\mathbf{0.896_{\pm 0.118}}$	$\mathbf{0.806_{\pm 0.067}}$	$\mathbf{0.560_{\pm 0.018}}$
	R^2	$\mathbf{0.457_{\pm 0.070}}$	$\mathbf{0.469_{\pm 0.152}}$	$\mathbf{0.854_{\pm 0.171}}$	$\mathbf{0.787_{\pm 0.072}}$	$\mathbf{0.366_{\pm 0.012}}$
LeJEPA	Corr.	$0.508_{\pm 0.060}$	$0.502_{\pm 0.043}$	$0.516_{\pm 0.050}$	$0.502_{\pm 0.063}$	$0.488_{\pm 0.052}$
	R^2	$0.272_{\pm 0.067}$	$0.260_{\pm 0.045}$	$0.278_{\pm 0.056}$	$0.269_{\pm 0.065}$	$0.255_{\pm 0.055}$
InfoNCE	Corr.	$0.570_{\pm 0.077}$	$0.561_{\pm 0.064}$	$0.524_{\pm 0.029}$	$0.542_{\pm 0.054}$	$0.529_{\pm 0.054}$
	R^2	$0.346_{\pm 0.101}$	$0.333_{\pm 0.076}$	$0.282_{\pm 0.035}$	$0.306_{\pm 0.061}$	$0.295_{\pm 0.063}$

Element-wise Identifiability Analysis Tab. 11 provides a granular, component-wise breakdown of the Mean Correlation Coefficient (MCC) and R^2 scores for the Causal3DIdent dataset. BayesDisc demonstrates comparable or superior identification performance across almost all causal variables compared to the standard InfoMax SSL baselines. Interestingly, the degree to which BayesDisc successfully identifies variables is not uniform. Certain factors, such as the Z-coordinate and specific color variables (Spotlight Color and Background Color), are identified with exceptionally high accuracy. Conversely, other variables show more moderate results. Because all continuous variables are treated equally during the nearest-neighbor pair matching process, the precise mechanisms driving this unequal variable identification remain an open question for future investigation.

Handling Unknown Latent Dimensionality via Overparameterization ($m > n$) As visualized in Fig. 4 (seed 0), BayesDisc isolates the 11 true causal variables into distinct latent coordinates with strong diagonal correlation. The remaining 21 excess dimensions do not suffer from dimensional collapse nor do they contaminate the identified causal variables. Instead, they remain uninformative components carrying residual variance, demonstrating that BayesDisc robustly accommodates overparameterized representations without requiring a priori knowledge of the true causal dimension. Notably, it successfully learns the dependency between variables (Y and Spotlight Position).

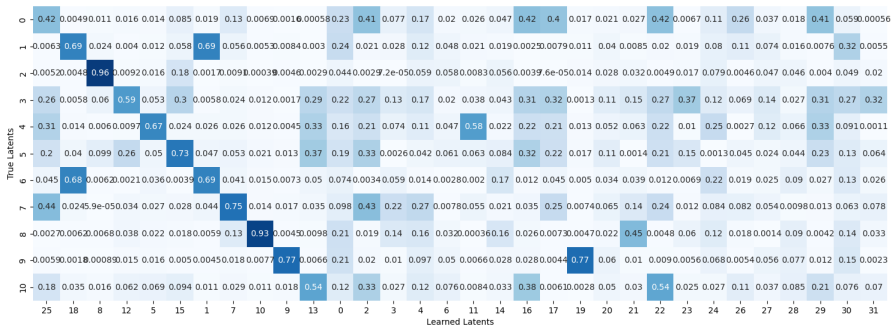


Figure 4: Element-wise result of latent overparameterization ($m = 32 > n = 11$) on Causal3DIdent (seed 0). Darker color represents higher correlation. The 11 ground-truth causal factors are distinctly identified, while the 21 excess latent dimensions remain uninformative without collapsing. Note that the order of learned latent variables is permuted to match the true latent variables.

H Discussion on Limitations and Theory-to-Practice Gaps

While our theoretical results and empirical validations demonstrate the effectiveness of BayesDisc for causal representation learning under paired partial observability, our approach exhibits several theoretical boundaries and practical trade-offs. In this section, we provide an in-depth discussion of these limitations and clarify the gaps between our formal theory and empirical implementation.

H.1 Finite-Sample Consistency and Optimization Dynamics

While our theoretical results establish identifiability at the level of the population objective and its continuous optimization trajectory, they do not provide an end-to-end gradient descent consistency guarantee for the finite-sample estimator. This gap between population-level guarantees and empirical optimization highlights an important boundary of our work and the broader field.

First, establishing the identifiability of the population objective is a necessary and logically prior step to analyzing finite-sample estimator consistency. A central theoretical question in our setup is whether the underlying causal variables can be identified from the objective at all without explicit view labels. Our proofs demonstrate that this is indeed possible, achieving the same block-identifiability guarantees as prior multi-view methods that relied on explicit labels. This population-level result serves as a foundational prerequisite that justifies the subsequent empirical optimization of the objective.

Second, the lack of finite-sample consistency guarantees under highly non-convex deep learning objectives remains an open theoretical challenge across the subfield of identifiability-based causal representation learning. Because standard gradient-based optimization in these landscapes is not mathematically guaranteed to converge to the global optimum, our theoretical guarantees are complemented by extensive empirical evaluations. As detailed in Sec. G.2.1, while BayesDisc successfully navigates the loss landscape in most practical regimes, it is not immune to catastrophic local minima in highly complex, high-dimensional spaces (e.g., $n = 16$). BayesDisc's core practical contribution is thus providing a tractable surrogate objective that enables standard optimizers to reach the theoretically guaranteed optimum across a wide variety of regimes, even in the absence of an end-to-end finite-sample consistency proof.

H.2 Inference-Time Deployment and Downstream Task Execution

A central property of our framework is that the proximity mask $\mathbf{\Phi}^{(\sigma)}(\mathbf{\Delta})$ is defined as an element-wise function of the representation residual across paired observations:

$$\mathbf{\Delta} = g_\theta(\mathbf{X}^{(1)}) - g_\theta(\mathbf{X}^{(2)}). \tag{62}$$

During self-supervised pre-training, these paired observations provide the necessary differential signal for BayesDisc to compute the analytic posterior $P(a \mid \Delta_j)$ and selectively align shared variables.

At inference time, however, the encoder g_θ is typically deployed on individual, unpaired test observations $\mathbf{X}$. Because the mask is strictly a function of the pair residual $\mathbf{\Delta}$, a single input $\mathbf{X}$ provides no residual. Consequently, the model cannot evaluate the proximity mask at test time to distinguish which latent coordinates represent observed causal factors versus unobserved or corrupted nuisance noise for that specific sample.

How downstream tasks are executed in light of this characteristic depends on the nature of the evaluation:

1. **Fully Observed Downstream Tasks (Standard Benchmarks):** In standard causal representation learning and downstream benchmark evaluations (e.g., downstream causal discovery in Tab. 3, linear probing, or regression on fully observed test sets), all underlying causal variables are present at inference time. Because BayesDisc block-identifies the true causal variables into distinct, coordinate-wise latent blocks, the entire representation vector $\hat{\mathbf{Z}} = g_\theta(\mathbf{X})$ captures the complete set of generative factors. Downstream models (such as graph discovery algorithms or linear classifiers) operate directly on this unmasked representation vector, fully benefiting from the disentangled causal structure without needing any inference-time mask.
2. **Partially Observed Single-Sample Tasks:** If an individual test instance is partially observed (i.e., some causal factors are missing or corrupted), the encoder embeds the unobserved coordinates as arbitrary unaligned representations. Because the encoder preserves structural variance rather than collapsing it, downstream supervised models trained on these representations can naturally learn task-specific weighting to focus on predictive coordinates while ignoring uninformative dimensions.
3. **Comparative and Counterfactual Downstream Tasks:** For downstream tasks that inherently operate on pairs of inputs—such as counterfactual reasoning, change detection, style transfer, or action effect prediction—the test pair $(\mathbf{X}, \mathbf{X}')$ can directly evaluate the BayesDisc posterior $\mathbf{\Phi}^{(\sigma)}(g_\theta(\mathbf{X}) - g_\theta(\mathbf{X}'))$. In this setting, the inferred mask explicitly identifies which causal variables underwent transformation between the two observations.

H.3 Theory-to-Practice Gap: Entropy Maximization and Bounded Support

In our theoretical formulation (Sec. 3.1 and Sec. B), the inner objective of Disc-InfoMax incorporates an entropy maximization term evaluated over the unit hypercube $(0,1)^m$ via an auxiliary projection t:

$$-H(t(g_\theta(\mathbf{X}))) = D_{\mathrm{KL}}\left(p_{t(g_\theta(\mathbf{X}))} \parallel \mathrm{Uniform}((0,1)^m)\right). \tag{63}$$

In our empirical implementation of BayesDisc (Sec. 3.2), however, we replace this theoretical term with Sketched Isotropic Gaussian Regularization (SIGReg) (Balestriero & LeCun, 2025), which regularizes the unprojected representations $\hat{\mathbf{Z}}$ toward an isotropic standard normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$ in unbounded Euclidean space $\mathbb{R}^m$.

This formulation mirrors the foundational theoretical-empirical bridge established by (von Kügelgen et al., 2021) in discriminative SSL:

- **Why the Uniform Hypercube in Theory?** From a measure-theoretic standpoint, formulating maximal differential entropy over a compact domain with bounded support (such as the unit hypercube $(0,1)^m$) is mathematically useful. On $(0,1)^m$, the uniform distribution is the unique maximum-entropy distribution, where the density is identically constant ($p(u) = 1$) and the negative entropy achieves an exact infimum of zero ($D_{\mathrm{KL}} = 0$).
- **Why Isotropic Gaussian Regularization in Practice?** Practically, as recognized across modern SSL frameworks (e.g., SimCLR (Chen et al., 2020), Barlow Twins (Zbontar et al., 2021), VICReg (Bardes et al., 2021), and LeJEPA (Balestriero & LeCun, 2025)), the core functional utility of the entropy maximization term is simply *collapse prevention*. The objective serves to prevent the representations from collapsing onto a single point or a degenerate low-dimensional subspace. Isotropic Gaussian regularizers like SIGReg achieve this collapse prevention with minimal computational overhead and numerical stability, providing an effective practical proxy for the theoretical entropy term.

H.4 Causal Dependencies, the Projection t, and the Collapse Dilemma

A closely related theoretical nuance involves statistical and causal dependencies among latent variables. In real-world SCMs, causal variables are frequently dependent (e.g., $C_1 \to C_2$).

In theory, the auxiliary projection $t \in \mathcal{T}$ is introduced precisely to accommodate these dependencies (Yao et al., 2023). In practice, however, parameterizing t as a flexible learnable function (e.g., an MLP) creates a fatal optimization failure mode: excessive flexibility in the projection t sabotages collapse prevention. If t is parameterized by a flexible neural network, the encoder g_θ can collapse the representation $\hat{\mathbf{Z}}$ onto a nearly lower-dimensional manifold, while the projection head t simply re-expands the collapsed dimensions to satisfy the entropy criterion on $t(\hat{\mathbf{Z}})$. The encoder thus achieves low total loss while completely failing to learn informative representations.

To prevent this collapse, the capacity of t must be severely restricted. Indeed, Multi-View CRL (Yao et al., 2023) encountered this exact dilemma: although their formal proofs relied on an auxiliary projection t, their official implementation omitted the projection head entirely and set $x = t(x)$.

When $x = t(x)$ and an isotropic Gaussian regularizer like SIGReg is applied directly to $\hat{\mathbf{Z}}$, an apparent contradiction arises: the regularizer penalizes cross-covariance between coordinates ($\mathrm{Cov}(\hat{Z}_i, \hat{Z}_j) \to 0$ for $i \neq j$), which could theoretically suppress the true causal dependencies between variables.

In practice, this theoretical tension is reconciled by empirical optimization dynamics:

1. **Non-Zero Optimization Equilibrium:** Deep neural networks optimized via gradient descent with finite steps and weight decay never drive the training loss to its theoretical infimum of zero.
2. **Loss Balancing:** The objective consists of two competing forces: the Discernment loss (which anchors the representations to the true generative factors and their inherent correlations) and the entropy loss (which prevents dimensional collapse). At empirical convergence, the model reaches a balanced trade-off where coordinates remain uncollapsed while preserving non-zero cross-variable correlations.
3. **Empirical Evidence:** As evidenced by our numerical experiments (Tab. 3) and the correlation heatmaps in Sec. G, BayesDisc preserves the true off-diagonal causal dependencies and achieves superior downstream causal discovery without suffering from dimensional collapse.

We emphasize that our primary contribution is the *Discernment mechanism* (Disc-InfoMax and BayesDisc), which resolves the mask inference and selective alignment challenge under view-label-free partial observability. Developing an entropy regularizer that simultaneously prevents dimensional collapse while natively supporting arbitrary non-linear copula dependencies remains an open challenge for the broader representation learning community.