BLOCK DISENTANGLEMENT IN CRL: BRIDGING IDENTIFIABILITY AND VISUAL STATE ESTIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Causal representation learning (CRL) is the process of recovering causally-related latent variables from high-dimensional observations. As a label-free inference method, CRL is particularly attractive for applications where data labels are unavailable or impractical to obtain. While there has been significant progress in understanding the identifiability guarantees of CRL, such guarantees often hold under highly stylized assumptions, which temper the direct application to real-world problems. This paper has a two-fold objective for *interventional* CRL. First, it establishes identifiability guarantees for substantially weaker interventional assumptions, resulting in *block disentanglement* of the causal variables, where the block structure depends on the realistically available intervention mechanisms. Secondly, the block disentanglement framework is used for embodied visual state estimation, in which the objective is to recover the latent physical variables of a robotic system directly from visual data (images and videos) *without labeled data*. These two components are critically complementary. The block disentanglement theory delineates identifiability guarantees under weakened assumptions, and the application demonstrates that the resulting objective remains effective in a controlled embodied setting despite further assumption violations, providing a theory-to-practice bridge needed to translate the promise of **label-free** CRL into practical problems.

1 INTRODUCTION

Physical systems are generally described by a set of state variables needed to characterize the systems' condition, in conjunction with the physical laws that delineate how these variables causally influence one another. These variables and causal mechanisms determine how the system evolves and responds to different forms of external stimulation or interventions. In many applications, however, these variables and mechanisms are not observed directly. Instead, the system is monitored or perceived through high-dimensional sensory systems (e.g., images and videos). This creates a fundamental need to recover hidden variables and causal mechanisms from high-dimensional observable data. Such recovery is critical for meaningful reasoning about the system (e.g., how the system responds to hypothetical stimuli) and for transferring models across environments with distributional shifts.

The recovered latent models and mechanisms must be provably correct, generating system representations that specify the true underlying system state variables. Supervised learning offers a direct solution to this, but it is label-intensive and requires large-scale per-sample labels, which are often unavailable to expense and complexity of curating annotated and labeled data. To address this challenge, *identifiable representation learning* circumvents dependence on labeled data while providing provably correct recovery without per-sample supervision.

Causal representation learning (CRL) has emerged as a form of identifiable representation learning apt for physical systems. Specifically, it is formed at the confluence of three primary research directions: (i) disentangling generative factors embedded in high-dimensional data, (ii) causal inference, and (iii) representation learning. Because a system's variables are not necessarily statistically independent (they are coupled through complex interactions), CRL aims to jointly disentangle them and learn the causal mechanisms that connect them (Schölkopf et al., 2021). Such causal mechanisms represent stable conditional laws relating causes to effects. Extensive studies in recent years have established the identifiability limits of CRL, establishing when latent causal factors are recoverable; for representative studies see (Varıcı et al., 2025; Ahuja et al., 2023; Yao et al., 2025; Ng et al., 2025). In parallel, there has been an active interest in translating these CRL guarantees into applications in genomics (Tejada-Lapuerta et al., 2025), biomedical (Sun et al., 2025) and dynamical systems (Yao et al., 2024).

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.

A particularly important application with growing interest in which CRL offers a natural solution is the analysis of video observations of physical systems (e.g., robots), where the objective is recovering the latent state variables from image sequences without requiring frame-by-frame labels (Lippe et al., 2023a;b; Chen et al., 2025). A video provides a high-dimensional record of how the physical system evolves and responds to external inputs, while the true variables (e.g., a robot's positions, orientations, and velocity) are hidden and embedded in the video record. In such applications, while videos are readily and inexpensively available, annotating each frame is prohibitively laborious and costly. Therefore, this is a setting in which CRL can offer significant potential by identifying and recovering latent variables and their causal dynamics directly from unlabeled video.

Theory-to-practice gap. Achieving these objectives in videos, however, necessitates operationalizing CRL in regimes beyond what the current CRL theory covers. Specifically, while identifiability results in CRL often provide strong guarantees, they also rely on *collectively* strong assumptions such as access to counterfactual observations; atomic (single-node) intervention targets; stochastic hard or *do* interventions; independent samples; exhaustive interventional diversity; and an exactly known latent dimensionality. While each of these assumptions can be justifiably valid in certain settings, in real-world settings, such as systems evolving continuously over time, they might not all be satisfied *simultaneously*. Challenges that violate this set of assumptions include sequential access to information and partial or imperfect interventions with unknown targets. As a result, the gap between identifiable CRL theory and label-free state recovery from video is fundamentally a gap in the joint validity of the assumptions required by existing guarantees. For more discussion on the existing CRL guarantees and their attendant assumptions, we refer to Appendix A.

To bridge the gap between current CRL theory and practical visual state estimation, we take two complementary steps, one theoretical and one empirical. Both are critically needed, as they address different but complementary obstacles.

- **Theoretical bridge:** The theoretical step weakens the intervention assumptions required for identifiability. Specifically, current CRL theory focuses on disentangling *individual* variables, which in turn often necessitates intervention structures that isolate one variable at a time. To address this, we consider *block disentanglement*, and establish conditions under which an intervened *block* of variables remains identifiable even when individual variables cannot be disentangled separately.
- **Empirical bridge:** On the empirical side, we further weaken the assumptions and investigate whether the resulting block disentanglement principle remains effective when applied to embodied visual data, where temporal dependence, imperfect intervention effects, and uncertainty in the latent representation arise naturally. For this purpose, robotics is a natural target and has long motivated CRL (Schölkopf et al., 2021): an effective robotic world model must capture the causal factors governing the robot's state and its interaction with the environment. In our experiments, we show that even under extensive assumption violations, the proposed method retains significant information about the "active block" of variables in a simulated robot pose estimation task.

Collectively, our results demonstrate that interventional CRL remains highly effective beyond the stringent regimes under stylized assumptions. We achieve this by relaxing the identifiable object target from individual causal variables to an intervention-defined block under weaker conditions; the resulting block disentanglement principle remains effective even when the experimental data further depart from the theoretical requirements. Accordingly, our results point to a step-by-step transition in which progressively weaker theoretical *and* empirical conditions still yield useful representation learning objectives.

2 PRELIMINARIES

Latent causal model. Consider a latent causal space consisting of n causal random variables $Z \triangleq [Z_1, \dots, Z_n]^\top$. An *unknown* transformation $f : \mathbb{R}^n \to \mathbb{R}^d$ maps Z to high-dimensional observed random variables $X \triangleq [X_1, \dots, X_d]^\top$ according to $X = f(Z)$, where $d \geq n$. As is standard in CRL literature, we assume that f is injective, and a smooth diffeomorphism onto its image. Denote the probability distributions of Z and X by p and p_X, respectively. Following causal Bayesian network formalism (Pearl, 2009), the distribution p of Z factorizes according to a directed acyclic graph (DAG) $\mathcal{G}$ on n nodes, where node i represents Z_i:

$$p(z) = \prod_{i=1}^{n} p_i(z_i \mid z_{\mathrm{pa}(i)}) \,, \tag{1}$$

where $\mathrm{pa}(i)$ denotes the parents of node i, and conditional distribution $p_i(z_i \mid z_{\mathrm{pa}(i)})$ encodes its *causal mechanism*. We use $\overline{\mathrm{pa}}(i)$ to denote $\mathrm{pa}(i) \cup \{i\}$. We assume p has full support on $\mathbb{R}^n$ and that all causal mechanisms $\{p_i(z_i \mid z_{\mathrm{pa}(i)}) : i \in [n]\}$ are continuously differentiable with respect to all latent variables.

Interventions. Interventions are a defining ingredient of causal modeling and play a central role in CRL. We begin by describing the most restrictive intervention model and then progressively relax it. The strongest form is an atomic do-intervention on a latent variable Z_i, which deterministically sets that variable to a prescribed value, $Z_i = z_i$, while removing its dependence on its usual causal mechanism. In this paper, intervention assumptions are weakened in two directions:

- **Unknown interventions:** We assume access to K intervention models such that, under each model, an *unknown* subset of the variables is intervened on. We denote the subset of variables intervened by intervention $k \in [K]$ by $\mathcal{I}_k \subseteq [n]$.
- **Stochastic interventions:** We assume we lack access to surgical and deterministic do interventions and can perform only stochastic interventions, in which only probability distributions can be changed. Specifically, an interventional environment that intervenes on Z_i changes its causal mechanism $p_i(z_i \mid z_{\mathrm{pa}(i)})$ to another distinct (conditional) distribution, while leaving the mechanisms of all other variables intact. We consider both *stochastic hard and stochastic soft interventions*: a soft intervention on Z_i replaces the observational mechanism with a different $q_i(z_i \mid z_{\mathrm{pa}(i)})$, whereas a hard intervention removes the influence of the parents altogether and changes the conditional distribution of Z_i to the marginal distribution $q_i(z_i)$. Thus, applying intervention k, which impacts nodes $\mathcal{I}_k$ creates an *interventional environment* that changes the distribution of the latent variables to

$$\text{hard intervention:} \quad q^{(k)}(z) = \prod_{j \notin \mathcal{I}_k} p_j(z_j \mid z_{\mathrm{pa}(j)}) \cdot \prod_{i \in \mathcal{I}_k} q_i^{(k)}(z_i)\,, \tag{2}$$

$$\text{soft intervention:} \quad q^{(k)}(z) = \prod_{j \notin \mathcal{I}_k} p_j(z_j \mid z_{\mathrm{pa}(j)}) \cdot \prod_{i \in \mathcal{I}_k} q_i^{(k)}(z_i \mid z_{\mathrm{pa}(i)})\,. \tag{3}$$

CRL block disentanglement. The main objective in CRL is to use observational data X to recover the latent variables Z and the latent causal graph $\mathcal{G}$. Identifiability in CRL pertains to establishing conditions under which a unique equivalence class of models explains the observed data. Existing CRL identifiability results mainly focus on *elementwise* disentanglement of Z, that is, recovering the variables $\{Z_i : i \in [n]\}$ individually (Squires et al., 2023; Buchholz et al., 2023; Zhang et al., 2023; von Kügelgen et al., 2023; Varıcı et al., 2025; Chen et al., 2024). However, these results generally require strong assumptions. In contrast, a more relaxed identifiability target, *block disentanglement*, has received relatively less attention (Ahuja et al., 2023; Liang et al., 2023; Yao et al., 2025). In our theoretical analysis, we aim to prove block disentanglement under weaker conditions on data availability. For completeness, we formally define both types of disentanglement as follows.

Definition 1 (Disentanglement classes)**.** *Let $\hat{Z}$ be an estimate of the latent variables Z.*

*(i) We say that $\hat{Z}$ achieves **elementwise disentanglement** if there exist a permutation π of $[n]$ and diffeomorphisms $\phi_i : \mathbb{R} \to \mathbb{R}$ such that $\hat{Z}_{\pi_i} = \phi_i(Z_i)$ for all $i \in [n]$.*

*(ii) We say that $\hat{Z}$ achieves **block disentanglement** for $\mathcal{I} \subseteq [n]$ if there exist a subset $\hat{\mathcal{I}} \subseteq [n]$ with size $|\mathcal{I}|$ and a diffeomorphism $\phi_{\mathcal{I}} : \mathbb{R}^{|\mathcal{I}|} \to \mathbb{R}^{|\mathcal{I}|}$ such that $\hat{Z}_{\hat{\mathcal{I}}} = \phi_{\mathcal{I}}(Z_{\mathcal{I}})$.*

Hence, elementwise disentanglement states that all latent variables are recovered up to a permutation and elementwise transformations, whereas block disentanglement states that the estimate $\hat{Z}_{\hat{\mathcal{I}}}$ of block $\mathcal{I}$ is an invertible function of the true block $Z_{\mathcal{I}}$ without leakage from other variables or noise.

Weakened assumptions. In this paper, we consider a substantially more relaxed setting than those typically studied in the existing literature. Before reviewing how individual works combine different assumptions, we first organize the commonly used assumptions into six broad categories. This categorical view provides a common reference point for comparing the assumptions adopted across existing CRL studies.

- *(C1) Atomic targets.* Each intervention alters the mechanism of exactly one latent coordinate.
- *(C2) do or hard interventions.* These interventions break the dependence of the intervened variables on their parents.
- *(C3) Pairing.* Some studies adopt more than one intervention environment per node. In such cases, the environments associated with Z_i come in matched pairs $(q^i, \bar{q}^i)$.

- *(C4) Interventional discrepancy.* When a node undergoes a pair of interventions, the mechanisms $q_i(z_i)$ and $\bar{q}_i(z_i)$ are sufficiently distinct: $\nabla_{z_i} \log \big(q_i(z_i)/\bar{q}_i(z_i)\big)$ is nonzero almost everywhere.
- *(C5) Sample independence.* Samples within each environment are i.i.d.
- *(C6) Latent dimension.* The latent dimension n is assumed to be known.

Existing results obtain identifiability under different tradeoffs between intervention structure and assumptions on the latent model or observation map. In the general nonlinear setting closest to ours, (von Kügelgen et al., 2023; Varıcı et al., 2025) establish elementwise disentanglement under (C1)–(C6): interventions are atomic and hard, with paired interventions satisfying the node-wise discrepancy condition. Other works in the literature operate under additional structural assumptions, such as linear transformations, linear causal models, or known causal graphs (Squires et al., 2023; Buchholz et al., 2023; Zhang et al., 2023; Ahuja et al., 2023; Liang et al., 2023; Yao et al., 2025).

In this paper, we relax assumptions (C1)–(C6) in two stages. In the first stage, which forms the *theoretical bridge*, we replace assumptions (C1)–(C4) with weaker ones, and establish identifiability guarantees together with the corresponding block-disentanglement results. These four assumptions collectively imply requiring less stringent conditions on the power, coverage, and diversity of the interventions. In the second stage, which forms the *empirical bridge*, we move further beyond the theory by considering settings that additionally violate (C5) and (C6). Combined, these two stages allow us to provide empirical evidence for disentanglement from visual data under conditions that are substantially weaker than those considered in the existing literature.

3 Theoretical Bridge: Weakening Intervention Diversity

Next, we develop the theoretical framework for establishing *block disentanglement* under the weakened assumptions by removing (C1)–(C4). We begin by introducing the definitions needed to formalize the problem, methodology, and analysis used throughout the remainder of the paper. Following the score-based interventional CRL framework of (Varıcı et al., 2025), we denote the score functions associated with PDFs p and p_X by $s(z) \triangleq \nabla_z \log p(z)$ and $s_X(x) \triangleq \nabla_x \log p_X(x)$, respectively. Similarly, the score of interventional PDFs $q^{(k)}$ and its observational counterpart are denoted by $s^{(k)}$ and $s_X^{(k)}$, respectively. Accordingly, we define the score difference function under intervention $\mathcal{I}_k$ with respect to the base observational environment as

$$d^{(k)}(z) \triangleq s^{(k)}(z) - s(z) = \sum_{i \in \mathcal{I}_k} \nabla_z \log \frac{q_i^{(k)}(z_i | z_{\mathrm{pa}(i)})}{p_i(z_i | z_{\mathrm{pa}(i)})} \,. \tag{4}$$

Since the ratio term depends only on $\overline{\mathrm{pa}}(i)$, for the gradient we have

$$[d^{(k)}(z)]_j = 0 \,, \qquad \forall j \notin \overline{\mathrm{pa}}(\mathcal{I}_k) \,. \tag{5}$$

Furthermore, the score differences in the latent space and the observed space are related by the Jacobian of f according to (Varıcı et al., 2025):

$$d_X^{(k)}(x) = \big[J_f(z)^\dagger\big]^\top \cdot d^{(k)}(z) \,. \tag{6}$$

Next, we define the **active block** as the union of the intervention targets $\mathcal{I} \triangleq \bigcup_{k \in [K]} \mathcal{I}_k$, with size $m \triangleq |\mathcal{I}|$, and denote the complementary **nuisance** coordinates by $\mathcal{I}^c \triangleq [n] \setminus \mathcal{I}$. We emphasize that the sets $\{\mathcal{I}_k : k \in [K]\}$, and subsequently $\mathcal{I}$ and $\mathcal{I}^c$ are *unknown*. Collecting the latent score differences for each environment $d^{(k)}(z)$, we define the latent score difference matrix as

$$\mathbf{D}(z) \;\triangleq\; \big[\, d^{(1)}(z), \ldots, d^{(K)}(z) \,\big] \;\in\; \mathbb{R}^{n \times K} \,. \tag{7}$$

Similarly, we denote the score difference matrix in observation domain by $\mathbf{D}_X(x)$. Next, we define the key conditions to formalize sufficient conditions that ensure block disentanglement.

Definition 2 (Ancestral closure of the active block)**.** *The active block $\mathcal{I}$ is said to be* ancestrally closed *in $\mathcal{G}$ if* $\mathrm{pa}(i) \subseteq \mathcal{I}$, $\forall i \in \mathcal{I}$. *Equivalently,* $\overline{\mathrm{pa}}(\mathcal{I}) = \mathcal{I}$.

Definition 3 (Collective rank diversity)**.** *The collection of interventions is said to satisfy* collective rank diversity *over the active block $\mathcal{I}$ if the associated score differences jointly span the active subspace, i.e., for almost every z, we have* $\mathrm{rank}\big(\mathbf{D}(z)\big) = |\mathcal{I}|$.

Definitions 2 and 3 replace the canonical interventional assumptions (C1)–(C4). Rather than constraining each intervention individually to be atomic, hard, paired, or sufficiently distinct, these

definitions impose only aggregate structural conditions on the collection of interventional environments. Definition 2 ensures that intervention-induced score differences remain confined to the active block $\mathcal{I}$, while Definition 3 requires that these score differences collectively span that block.

Specifically, Definition 3 is the block-level counterpart of the interventional discrepancy condition (C4). This condition is satisfied by $K \geq |\mathcal{I}|$ multi-node interventional environments as long as their score difference directions span the active subspace almost everywhere. Furthermore, this condition is strictly weaker relative to the canonical full-coverage setting considered by (C1)–(C4). In particular, under (C1)–(C4), taking score differences between the paired hard-intervention environments produces one nonzero direction per intervened coordinate, and hence satisfies Definition 3. The converse does not hold: collective rank diversity permits unpaired, multi-node, and soft interventions, provided their score difference directions jointly span the active block. Hence, the two definitions replace intervention-by-intervention restrictions with collective conditions on support and span. This shows that the strong canonical intervention structure can be relaxed substantially while still retaining a nontrivial identifiability guarantee. Leveraging these definitions, we provide sufficient conditions for the block-identifiability result formalized next.

Theorem 1 (Block identifiability under collective rank diversity, (cf. Appendix B for proof))**.** *Let (h, g) be a smooth encoder–decoder pair that perfectly reconstructs X, i.e., $g(h(X)) = X$. If the estimated matrix*

$$\mathbf{D}_{\hat{z}}(\hat{z}) = \left[J_h(x)^\dagger\right]^\top \mathbf{D}_X(x) \tag{8}$$

is supported on a row set $\hat{\mathcal{I}} \subseteq [n]$ with $|\hat{\mathcal{I}}| = |\mathcal{I}|$, then under the ancestral-closure and collective-rank-diversity conditions of Definitions 2 and 3, together with assumptions (C5)–(C6), the estimate $\hat{z} = h(X)$ block-disentangles the true active block $Z_{\mathcal{I}}$.

Comparison to existing block disentanglement results. Existing block-identifiability results rely on substantially different structural information. von Kügelgen et al. (2021) established the concept under counterfactual CRL setting. In interventional CRL, (Ahuja et al., 2023) obtains a block-affine identification result by exploiting intervention-induced support independence, and (Liang et al., 2023) assumes that the causal graph and intervention targets are known and require multiple perfect interventions on the same target block satisfying a block-interventional discrepancy condition. Most closely related setting is (Yao et al., 2025) which provides a general invariance-based framework for block identifiability: latent blocks that *remain invariant* can be identified. In contrast, our block disentanglement target $\mathcal{I}$ is precisely the intervention-affected (i.e., *varying*) block.

4 Empirical Bridge: Block Disentanglement to Embodied Video

The theoretical bridge in Section 3 already relaxes assumptions (C1)–(C4), while retaining (C5)–(C6), and establishes block-disentanglement guarantees under this weaker setting. The next question is whether the resulting CRL principles remain effective when we move further toward practical visual state estimation and additionally relax (C5)–(C6). This constitutes our *empirical bridge*.

To evaluate this progression, we consider two complementary empirical settings. The first, which we call the *assumption-matched benchmark*, satisfies the full set of assumptions (C1)–(C6). It provides a reference point for the performance attainable when the canonical CRL conditions hold, and allows us to assess the overall effectiveness of the methodology in its most favorable identifiable setting. The second, which we call the *embodied setting*, additionally dispenses with (C5)–(C6), on top of the relaxation of (C1)–(C4) already established by the block-disentanglement theory. Comparing these two settings therefore serves two complementary purposes: benchmarking performance under the canonical assumptions, and assessing how far the combined theoretical and empirical bridges can carry CRL toward visual state estimation under substantially weaker conditions.

4.1 Setup and Data Generation

We consider estimating a robot's pose and joint configuration from visual observations as an important instance of visual state estimation. This capability is particularly relevant when a robot's internal state is unavailable, unreliable, or cannot be directly communicated to other agents. Accordingly, visual recovery of articulated robot state has received growing attention in computer vision and robotics. We refer to Appendix A for a thorough review of the related literature.

For our study, the latent physical state of interest is the robot arm's joint configuration. The rendered images depend on these joint angles, as well as other intrinsic and extrinsic factors such as link geometry, camera pose, lighting, and background. To isolate the state estimation problem, we vary

only the joint angles while holding the remaining generative factors fixed. We generate the visual observations using the Panda-Gym platform (Gallouédec et al., 2021), which provides controlled actuation of the robotic arm and corresponding rendered images. To preserve observability of the full joint configuration and support the injective rendering assumption on f, we record each pose from two distinct camera viewpoints.

As briefly mentioned earlier, in our analysis we consider the following empirical regimes.

- **Assumption-matched benchmark.** This setting satisfies (C1)–(C6). Interventions are atomic (C1), replacing one joint's sampling distribution with a *hard* intervention while leaving the others untouched (C2); environments come in matched pairs $(q^i, \bar{q}^i)$ per joint as (C3) requires, and are designed to satisfy (C4); joint angles are sampled i.i.d. and rendered independently, so there is no temporal structure (C5); and the true number of varying joints is known (C6). This regime validates whether disentanglement and variable estimation are numerically stable on rendered 128×128 images, and whether the canonical objective recovers the intended coordinates at this visual scale.
- **Embodied setting.** This setting moves beyond the conditions of Theorem 1 and additionally violates (C5)–(C6). Specifically, it replaces i.i.d. rendering with continuous videos. A proportional controller drives the arm from a sampled start to a sampled goal configuration, and it retains every frame along the way. Episodes terminate and reset, limiting dependence across trajectories but not within one. Each environment k corresponds to a distinct goal-sampling distribution that *prioritizes* displacement of a designated target joint while letting the remaining joints move as the dynamics dictate. Importantly, no observational dataset is distinguished. These departures from the theoretical setting arise through three distinct mechanisms: temporal dependence in the observed data, imperfect and coupled intervention effects, and mismatch between the learned bottleneck dimension and the true latent dimension. We summarize each violation below and indicate the corresponding theoretical assumption that it relaxes.
  - *V1: Temporal dependence* (violates C5). Consecutive frames differ by one controller step. Therefore, samples within a video are strongly correlated. We in fact leverage the controlled nature of the videos to compute score differences without an observational environment by treating the two halves of each video as observational and interventional data.
  - *V2: Off-target effects* (violates C1–C4 and Definitions 1 and 2). The control task itself induces small but nonzero motion in off-target joints, so $\mathbf{D}$ carries small nonzero entries on rows in $\mathcal{I}^c$ instead of exact zeros. The active/nuisance partition becomes approximate. Furthermore, environments are unpaired, and interventions are not hard.
  - *V3: Overcomplete bottleneck* (violates C6). We train autoencoders with $d_b \in \{8, 12, 16, 32\}$ dimensions against 6 true latent factors. Increasing d_b increases the autoencoder's reconstruction capacity. However, without a disentanglement signal, nothing forces the active factor information to concentrate. This violation makes the active/nuisance contrast a valid metric even when $\mathcal{I} = [n]$.

4.2 Methodology

In this section, we describe how we implement a CRL algorithm that enjoys the theoretical identifiability results. In practice, we train an encoder–decoder pair (h, g) where $h : \mathbb{R}^d \to \mathbb{R}^{d_b}$ encodes the visual data into d_b entries. Our approach achieves disentanglement by estimating latent score functions. This can be implemented efficiently by leveraging a fundamental connection between latent score estimates and the score estimates of the observational data in (6): by estimating the score differences d^i_X once on the visual domain X, one can compute the estimated score difference $d^i_{\hat{Z}}$ via $d^i_{\hat{Z}}(\hat{Z}) = [J_g(\hat{Z})]^\top \cdot d^i_X(x)$.

Assumption-matched benchmark. In the assumption-matched benchmark introduced above, where assumptions (C1)–(C6) are all satisfied, Varıcı et al. (2025, Theorem 22) shows that elementwise disentanglement can be achieved by learning an encoder–decoder pair that minimizes a score-based regularization objective together with a reconstruction loss. Specifically, the objective penalizes deviations from the expected structure of score differences across interventional environments, while the reconstruction term ensures that the learned representation preserves the observed data. Under the stated assumptions, minimizing this objective guarantees *elementwise disentanglement* of the latent variables:

$$\mathcal{L}(h, g) = \underbrace{\mathbb{E}\big[\|g(h(x)) - x\|^2\big]}_{\text{Reconstruction Loss}} + \lambda \sum_{i=1}^{n} \underbrace{\big\|\mathbb{E}\big[\big|J_g(\hat{z})^\top \cdot d^i_X(x)\big|\big] - e_i\big\|^2}_{\text{Sparsity Loss}}, \tag{9}$$

where e_i denotes the i-th standard unit vector. This result follows from the observation that the score difference *between* the paired interventions q^i and $\bar{q}^i$ is *one-sparse* on the true latent domain.

Embodied setting. In the embodied setting, where C1–C6 no longer jointly hold, neither the pairing nor the atomicity is available. Therefore, the score difference does not fit to a single coordinate i (unlike (9)). Following Theorem 1, we instead confine the score differences to a fixed block. For each environment $k \in [K]$ we estimate $d_X^{(k)}$ directly from trajectories, use change-of-variables formula (6) to obtain $d_Z^{(k)}$, and penalize parts falling outside the estimated active block $\hat{\mathcal{I}}$:

$$\mathcal{L}_{\text{blk}}(h, g) = \underbrace{\mathbb{E}\big[\|g(h(x)) - x\|^2\big]}_{\text{Reconstruction}} + \lambda \underbrace{\Big\| \Big[\mathbb{E}\big[\big| J_g(\hat{z})^\top \cdot \mathbf{D}_X(x) \big| \big] \Big]_{\hat{\mathcal{I}}^c,:} \Big\|_F^2}_{\text{Block sparsity}} + \lambda_s \underbrace{\big(\mathbb{E}\big[\|h(x)\|\big] - 1\big)^2}_{\text{Scale}} . \quad (10)$$

Importantly, we do not need to know the latent dimension n. We assume that the active-block size $m = |\mathcal{I}|$ is known and designate m coordinates of the d_b-dimensional representation as $\hat{\mathcal{I}}$. The remaining $d_b - m$ coordinates are nuisance and can encode whatever information is needed for reconstruction.

5 Empirical Results and Analysis

This section empirically demonstrates that, despite extensive relaxation of interventional assumptions, we see only graceful degradation compared to CRL approaches designed under those assumptions. All experimental details, additional experiments and ablations are given in Appendices C and D.

Experiment pipeline. The implementation consists of three stages, summarized in Figure 1. The input consists of high-dimensional visual observations X collected under different interventional environments, and the objective is to obtain a low-dimensional estimate $\hat{Z}$ of the physical state.

Stage 1 (AE1) performs dimensionality reduction. Directly estimating score functions and score differences in the image domain is computationally prohibitive. We therefore use a convolutional autoencoder (E_1, D_1) to map each grayscale image X to a lower-dimensional $8 \times 8 \times 1$ feature representation $Y = E_1(X)$ by minimizing the reconstruction loss $\mathbb{E}_X[\|X - D_1(E_1(X))\|_2^2]$. After training, we freeze E_1 and use it as a feature extractor. This compression substantially reduces the dimension in which we must estimate score differences, at the potential cost of discarding latent information. We therefore separately evaluate how much state information is retained by Y.

Stage 2 (LDR) estimates intervention-induced score differences in the feature space. We exploit the fact that a binary classifier trained with cross-entropy loss estimates the log-density ratio between the two classes (Gutmann & Hyvärinen, 2012). For each interventional environment k, we train a log-density-ratio estimator $f_{\text{LDR}}^{(k)}$ to distinguish samples between chosen distributions. The gradient of its logit, $\nabla_y f_{\text{LDR}}^{(k)}(y)$, then provides an estimate of the corresponding feature-space score difference.

Stage 3 (AE2) learns the disentangled representation. A second autoencoder (E_2, D_2) is trained on Y using the objectives developed in Section 4.2, producing $\hat{z} \triangleq E_2(Y) \in \mathbb{R}^{d_b}$. The training objective depends on the empirical regime: in the assumption-matched benchmark, we use the canonical score sparsity objective (9) that promotes elementwise disentanglement, whereas in the embodied setting, we use the block sparsity objective (10) as motivated by our block-disentanglement theory.

Empirical objectives. The experiments are designed to determine how far the CRL principles developed in the preceding sections carry from the assumption-matched setting toward embodied visual data. In particular, we address the following questions:

- *Can CRL recover physical state from visual observations without state labels?* In the assumption-matched benchmark, we test whether the canonical score-based CRL objective can recover individual joint variables from high-dimensional images when its identifiability assumptions are satisfied, and compare its performance with supervised pose-estimation methods.
- *Does the block-disentanglement principle remain effective beyond the theory?* When we additionally violate the assumptions underlying the block identifiability theorem, we test whether our approach still concentrates information about the relevant physical variables in the target block?
- *Where in the pipeline is state information retained or lost?* We compare how much information about the true state is retained after visual compression, in the complete learned latent representation, and in the recovered active block. This lets us distinguish limitations from feature extraction from those from representation learning or block selection.
- *How sensitive is the resulting representation to departures from the idealized setting?* We study the effect of factors such as the bottleneck dimension and other experimental variations to assess whether the observed disentanglement behavior persists as the representation and data-generating process become less tightly matched to the theoretical assumptions.

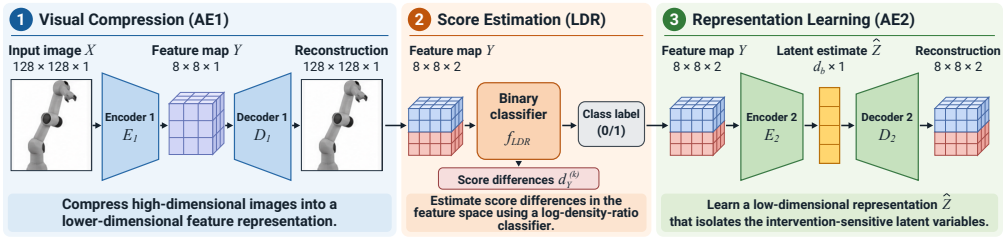


Figure 1: **Training pipeline.** A shared Autoencoder 1 (AE1) compresses each $128 \times 128 \times 1$ image X into an $8 \times 8 \times 1$ feature vector Y. Features from both cameras are concatenated along the channel axis, forming an $8 \times 8 \times 2$ input to the LDR and AE2. AE2 then outputs the final $d_b \times 1$ latent estimate $\hat{Z}$.

We show that the assumption-matched setting exhibits accurate elementwise state recovery, and in the embodied setting, the active block retains most of the state information available in the learned representation even though exact identifiability is no longer guaranteed.

Evaluation metrics. In order to answer the questions above, we specify appropriate metrics for both assumption-matched and embodied settings. In the assumption-matched setting, we consider Mean Squared Error (MSE) as our main metric. To account for the elementwise affine ambiguity of the recovered variables, we calibrate each estimated component using an affine transformation fitted on a separate set of 1000 samples, and report MSE on the held-out evaluation set.

To evaluate the embodied setting, we need a block-level measure of whether the learned representation preserves the information in the original latent. By Theorem 1, the representation within the active block is identifiable only up to an arbitrary diffeomorphism $\hat{Z}_{\hat{\mathcal{I}}} = \phi_{\mathcal{I}}(Z_{\mathcal{I}})$. Thus, we assess whether the learned active block retains sufficient information to recover $Z_{\mathcal{I}}$, which we denote *active predictability*, and measure it using explained variance $R^2(Z_{\mathcal{I}} \mid \hat{Z}_{\hat{\mathcal{I}}})$ under nonlinear kernel ridge regression. High active predictability indicates that the active block retains substantial information about the corresponding ground-truth latent. To localize where information is lost, we compare active predictability with two corresponding quantities at earlier stages of the data pipeline $Z \to X \to Y \to \hat{Z} \to \hat{Z}_{\hat{\mathcal{I}}}$. Specifically, we report

$$\underbrace{R^2(Z_{\mathcal{I}} \mid \hat{Z}_{\hat{\mathcal{I}}})}_{\text{active predictability}}, \qquad \underbrace{R^2(Z_{\mathcal{I}} \mid \hat{Z})}_{\text{all-latent predictability}}, \qquad \underbrace{R^2(Z_{\mathcal{I}} \mid Y)}_{\text{feature predictability}}. \tag{11}$$

The first term measures recovery from the active block, the second the full learned latent, and the third how much information about $Z_{\mathcal{I}}$ is contained in the feature map after the initial dimensionality reduction stage. These three quantities diagnose information loss at successive stages of the pipeline: feature extraction, representation learning, and construction of the active block.

Assumption-matched setting results. The setting for this experiment satisfies (C1)–(C6), and we use canonical CRL techniques. In this setting, we compare the performance of our approach against a recent supervised baseline, RoboPEPP (Goswami et al., 2025), in Table 1. For extensive ablations and additional results, see Appendix D. RoboPEPP is among the few methods that can estimate pose when joint angles are unknown at test time (along with HPE (Ban et al., 2024) and RoboPOSE (Labbé et al., 2021)), and reports stronger performance than these alternatives. RoboPEPP uses a two-stage training procedure. First, a self-supervised I-JEPA backbone (Assran et al., 2023) is trained on the entire training dataset. In the second stage, a "JointNet" is trained to predict the six joint angles from the concatenated I-JEPA embeddings of the two camera views, using an L_2 loss and ground-truth joint-angle labels. *In contrast, our CRL approach requires no labeled state data for training.*

As expected, RoboPEPP's MSE decreases as the amount of labeled training data increases. The important observation is that our approach achieves performance comparable to RoboPEPP when

Table 1: Comparison of MSE (rad^2) for RoboPEPP and our method in assumption-matched regime.

Model	Training Labels	Joint 1	Joint 2	Joint 3	Joint 4	Joint 5	Joint 6
Ours	none during training	0.083	0.015	0.217	0.035	0.198	0.080
RoboPEPP	1% of dataset	0.136	0.039	0.237	0.053	0.253	0.200
	5% of dataset	0.075	0.017	0.134	0.024	0.153	0.081
	10% of dataset	0.030	0.010	0.072	0.022	0.091	0.063
	100% of dataset	0.003	0.001	0.007	0.003	0.010	0.011

Table 2: Embodied setting results across score regularization weight λ and bottleneck dimension d_b.

λ	d_b	Active $R^2(Z_{\mathcal{I}} \mid \hat{Z}_{\hat{\mathcal{I}}})$	All-latent $R^2(Z_{\mathcal{I}} \mid \hat{Z})$	Feature $R^2(Z_{\mathcal{I}} \mid Y)$
0	16	0.33	0.72	0.77 (AE1 frozen)
50	16	0.58	0.63	
100	16	0.63	0.66	
0	32	0.23	0.77	
50	32	0.61	0.67	
100	32	0.55	0.66	
50	8	0.58	0.60	
50	12	0.63	0.66	
50	16	0.58	0.63	
50	32	0.61	0.67	

RoboPEPP is trained using labels for 5% of the dataset (approximately 6.5K labeled poses, corresponding to 13K labeled images from two camera orientations). The key point is therefore not only the resulting estimation accuracy, but that we achieve comparable performance while learning the underlying state representation **without** training on labeled data.

Embodied setting results. We now move beyond the theoretically-guaranteed regime, with assumption violations V1–V3 all active simultaneously. We apply our block-disentanglement approach and report the results in Table 2. For detailed plots during training, refer to Appendix C.4. Specifically, we answer two main questions:

- **Does the approach remain effective under V1–V3?** With score regularization, active predictability approaches all-latent predictability, indicating that the regularizer retains much of the state information already present in $\hat{z}$ within the active block $\hat{Z}_{\hat{\mathcal{I}}}$. In comparison, a reconstruction-only objective ($\lambda = 0$) results in significantly lower active predictability: for $d_b = 32$, $\lambda = 0$ yields 0.23 whereas $\lambda = 50$ yields 0.61, retaining most of the all-latent predictability of 0.67.
- **How sensitive is the representation to the assumption violations?** V1–V2 arise from the video modality and are therefore not cleanly ablatable without changing the embodied setting itself. V3, in contrast, can be varied directly through the bottleneck dimension. Increasing d_b substantially changes the capacity and information content of the full representation, but has little effect on active predictability: across $d_b \in \{8, 12, 16, 32\}$, the active block recovery remains at essentially the same level. Thus, the active/nuisance contrast persists under substantial overparameterization, while the full representation continues to learn additional features that purely help reconstruction.

In summary, our main empirical observation is that the structural principles motivating CRL remain effective well beyond their validity regime. In the embodied setting, despite extensive violations of the theoretical conditions that underlie it, the block regularizer retains substantially more recoverable state information within the designated active coordinates than the unregularized baseline.

Finally, we provide additional results in Appendix C, including different episode generation setups and training curves for Table 2.

6 Concluding Remarks

We have proposed a two-step path from CRL identifiability theory toward label-free visual state estimation. In this first step, we weaken canonical intervention assumptions and establish block-disentanglement guarantees, showing that an active group of causal variables can remain identifiable even when individual variables cannot be recovered separately. In the second step, we show that the resulting principle remains effective for retaining relevant information in embodied visual settings that further violate the theoretical assumptions. Empirically, score-based CRL recovers robot joint states from images without state labels in the assumption-matched setting, and the block-based objective continues to be informative about the relevant physical variables under temporal dependence, imperfect interventions, and overcomplete representations. These two components are complementary. The theory characterizes what remains identifiable under weaker assumptions, while the experiments assess whether those principles remain effective as CRL moves toward practical visual data. To further narrow the theory-practice gap and realize CRL's important potential, a key next step is to extend the identifiability theory to settings with temporally dependent observations, approximate intervention effects, and unknown or overcomplete latent dimensionality.

AI USE STATEMENT

In this work, we used generative AI tools to implement methods, and provide feedback on research methodology or experiments. We have not used generative AI tools to generate synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, support qualitative and thematic data analysis, and interpret results; and clean and reformat dataset, and assist with translation, are not applicable to this work. Additionally, we used generative AI tools to create or edit software code, draft parts of a research paper, summarize or analyze existing literature, brainstorming, and edit a research paper to improve readability. We have reviewed all AI-assisted work. LLM-generated code was verified and tested for correctness. LLMs were used to assist with high-level structural planning and localized refinements. All substantive claims, arguments, technical content, and the core draft were produced manually by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The relevant loss functions are fully specified in Section 4.2 and all implementation details are included in Appendices C to E.

REFERENCES

Emre Acartürk, Burak Varıcı, Karthikeyan Shanmugam, and Ali Tajer. Sample complexity of interventional causal representation learning. In *Proc. Advances in Neural Information Processing Systems*, Vancouver, Canada, December 2024.

Kartik Ahuja, Divyat Mahajan, Yixin Wang, and Yoshua Bengio. Interventional causal representation learning. In *Proc. International Conference on Machine Learning*, Honolulu, HI, July 2023.

Kartik Ahuja, Amin Mansouri, and Yixin Wang. Multi-domain causal representation learning via weak distributional invariances. In *Proc. International Conference on Artificial Intelligence and Statistics*, Valencia, Spain, May 2024.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. *arXiv:2301.08243*, 2023.

Shikun Ban, Juling Fan, Xiaoxuan Ma, Wentao Zhu, Yu Qiao, and Yizhou Wang. Real-time holistic robot pose estimation with unknown states. In *European Conference on Computer Vision*, pp. 1–17. Springer, 2024.

Simon Buchholz, Goutham Rajendran, Elan Rosenfeld, Bryon Aragam, Bernhard Schölkopf, and Pradeep Ravikumar. Learning linear causal representations from interventions under general nonlinear mixing. In *Proc. Advances in Neural Information Processing Systems*, New Orleans, LA, December 2023.

Guangyi Chen, Yunlong Deng, Peiyuan Zhu, Yan Li, Yifan Shen, Zijian Li, and Kun Zhang. Causalverse: Benchmarking causal representation learning with configurable high-fidelity simulations. In *Proc. Advances in Neural Information Processing Systems*, San Diego, CA, December 2025.

Tianyu Chen, Kevin Bello, Francesco Locatello, Bryon Aragam, and Pradeep Ravikumar. Identifying general mechanism shifts in linear causal representations. In *Proc. Advances in Neural Information Processing Systems*, Montreal, Canada, December 2024.

Quentin Gallouédec, Nicolas Cazin, Emmanuel Dellandréa, and Liming Chen. panda-gym: Open-source goal-conditioned environments for robotic learning. *arXiv:2106.13687*, 2021.

Pablo García-Ruiz, Francisco J Romero-Ramirez, Rafael Muñoz-Salinas, Manuel J Marín-Jiménez, and Rafael Medina-Carnicer. Fiducial objects: Custom design and evaluation. *Sensors*, 23(24): 9649, 2023.

Sergio Garrido-Jurado, Rafael Muñoz-Salinas, Francisco José Madrid-Cuevas, and Manuel Jesús Marín-Jiménez. Automatic generation and detection of highly reliable fiducial markers under occlusion. *Pattern Recognition*, 47(6):2280–2292, 2014.

Raktim Gautam Goswami, Prashanth Krishnamurthy, Yann LeCun, and Farshad Khorrami. Robopepp: Vision-based robot pose and joint angle estimation through embedding predictive pre-training. In *Proc. Conference on Computer Vision and Pattern Recognition Conference*, Nashville, TN, June 2025.

Michael U Gutmann and Aapo Hyvärinen. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *Journal of Machine Learning Research*, 13 (2), 2012.

Aapo Hyvärinen and Petteri Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 12(3):429–439, April 1999.

Ilyes Khemakhem, Ricardo Monti, Diederik Kingma, and Aapo Hyvärinen. Ice-beem: Identifiable conditional energy-based deep models based on nonlinear ICA. In *Proc. Advances in Neural Information Processing Systems*, virtual, December 2020.

Yann Labbé, Justin Carpentier, Mathieu Aubry, and Josef Sivic. Single-view robot pose and joint angle estimation via render & compare. In *Proc. Conference on Computer Vision and Pattern Recognition*, pp. 1654–1663, virtual, June 2021.

Inbeom Lee, Tongtong Jin, and Bryon Aragam. Beyond identifiability: Learning causal representations with few environments and finite samples. *arXiv:2603.25796*, 2026.

Timothy E Lee, Jonathan Tremblay, Thang To, Jia Cheng, Terry Mosier, Oliver Kroemer, Dieter Fox, and Stan Birchfield. Camera-to-robot pose estimation from a single image. In *Proc. IEEE International Conference on Robotics and Automation*, pp. 9426–9432, virtual, 2020.

Wendong Liang, Armin Kekić, Julius von Kügelgen, Simon Buchholz, Michel Besserve, Luigi Gresele, and Bernhard Schölkopf. Causal component analysis. In *Proc. Advances in Neural Information Processing Systems*, New Orleans, LA, December 2023.

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 *Proc. International Conference on Learning Representations*, Kigali, Rwanda, May 2023a.

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M Asano, Taco Cohen, and Efstratios Gavves. Biscuit: Causal representation learning from binary interactions. In *Proc. Uncertainty in Artificial Intelligence*, Pittsburgh, PA, August 2023b.

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 *Proc. International Conference on Machine Learning*, Long Beach, CA, June 2019.

Ignavier Ng, Shaoan Xie, Xinshuai Dong, Peter Spirtes, and Kun Zhang. Causal representation learning from general environments under nonparametric mixing. In *Proc. International Conference on Artificial Intelligence and Statistics*, Mai Khao, Thailand, May 2025.

Judea Pearl. *Causality*. Cambridge University Press, Cambridge, UK, 2009.

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.

Chandler Squires, Anna Seigal, Salil S. Bhate, and Caroline Uhler. Linear causal disentanglement via interventions. In *Proc. International Conference on Machine Learning*, Honolulu, HI, July 2023.

Yuewen Sun, Lingjing Kong, Guangyi Chen, Loka Li, Gongxu Luo, Zijian Li, Yixuan Zhang, Yujia Zheng, Mengyue Yang, Petar Stojanov, Eran Segal, Eric P. Xing, and Kun Zhang. Causal representation learning from multimodal biomedical observations. In *Proc. International Conference on Learning Representations*, Singapore, April 2025.

Alejandro Tejada-Lapuerta, Paul Bertin, Stefan Bauer, Hananeh Aliee, Yoshua Bengio, and Fabian J Theis. Causal machine learning for single-cell genomics. *Nature Genetics*, pp. 1–12, 2025.

Burak Varıcı, Emre Acartürk, Karthikeyan Shanmugam, and Ali Tajer. Linear causal representation learning from unknown multi-node interventions. In *Proc. Advances in Neural Information Processing Systems*, Vancouver, Canada, December 2024.

Burak Varıcı, Emre Acartürk, Karthikeyan Shanmugam, Abhishek Kumar, and Ali Tajer. Score-based causal representation learning: Linear and general transformations. *Journal of Machine Learning Research*, 26(112):1–90, 2025.

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 *Proc. Advances in Neural Information Processing Systems*, virtual, December 2021.

Julius von Kügelgen, Michel Besserve, Wendong Liang, Luigi Gresele, Armin Kekić, Elias Bareinboim, David M Blei, and Bernhard Schölkopf. Nonparametric identifiability of causal representations from unknown interventions. In *Proc. Advances in Neural Information Processing Systems*, New Orleans, LA, December 2023.

Dingling Yao, Caroline Muller, and Francesco Locatello. Marrying causal representation learning with dynamical systems for science. In *Proc. Advances in Neural Information Processing Systems*, Vancouver, BC, Canada, December 2024.

Dingling Yao, Dario Rancati, Riccardo Cadei, Marco Fumero, and Francesco Locatello. Unifying causal representation learning with the invariance principle. In *Proc. International Conference on Learning Representations*, Singapore, April 2025.

Soheun Yi, Yizhou Lu, Chandler Squires, and Pradeep Ravikumar. Concept modulation models: A unified framework for identifiability and extrapolation. *arXiv:2606.18509*, 2026.

Jiaqi Zhang, Chandler Squires, Kristjan Greenewald, Akash Srivastava, Karthikeyan Shanmugam, and Caroline Uhler. Identifiability guarantees for causal disentanglement from soft interventions. In *Proc. Advances in Neural Information Processing Systems*, New Orleans, LA, December 2023.

TABLE OF CONTENTS

A Related Work and Limitations **13**

B Proof of Theorem 1 **14**

C Embodied Regime **15**

C.1 Data Generation 16

C.2 Score Difference Estimation 16

C.3 Training and Evaluation Details 17

C.4 Sensitivity and Additional Results 17

D Assumption-Matched Regime **19**

D.1 Data Generation 19

D.2 Training Details 23

D.3 Additional Quantitative Results 23

D.4 Supervised Comparison and Robustness 25

D.5 Ablations 29

D.6 Qualitative Examples 32

E Architecture Details **32**

A RELATED WORK AND LIMITATIONS

Identifiable causal representation learning. Without additional structure or distributional diversity, recovering latent variables from observations under an unknown nonlinear mixing function is generally impossible (Hyvärinen & Pajunen, 1999; Locatello et al., 2019). Causal representation learning (CRL) addresses this ambiguity by exploiting causal structure and changes across environments. A substantial literature establishes identifiability from interventions under different assumptions on the observation map, latent causal model, and intervention mechanisms (Squires et al., 2023; Buchholz et al., 2023; Zhang et al., 2023; von Kügelgen et al., 2023; Varıcı et al., 2024; 2025; Acartürk et al., 2024; Ng et al., 2025; Lee et al., 2026). Other approaches include counterfactual CRL (von Kügelgen et al., 2021), unified frameworks to analyze interventional CRL along with nonlinear ICA and perturbation models (Yi et al., 2026), and temporal and multi-view CRL literature (Yao et al., 2025). These results range from linear latent representations or observation models to general nonlinear transformations, and consider settings including hard, soft, known, and unknown interventions.

Interventional diversity is particularly relevant to our work. Existing nonlinear CRL guarantees commonly obtain elementwise disentanglement by exploiting sufficiently localized and diverse changes across interventional environments (von Kügelgen et al., 2023; Varıcı et al., 2025). Our result instead asks what remains identifiable when such intervention structure is unavailable, replacing intervention-by-intervention requirements with collective conditions on the support and span of score differences. The precise comparison between these assumptions and our conditions is given in Section 3.

Block disentanglement. When individual latent coordinates cannot be identified, a weaker objective is to recover groups of variables up to an invertible transformation within each group. von Kügelgen et al. (2021) identify the invariant content block from paired observations (also called counterfactual setting) whose augmentations preserve content while sufficiently varying the remaining style variables, allowing causal dependence between content and style.

Under polynomial transformation, Ahuja et al. (2023) exploit intervention-induced support independence to obtain block-affine guarantees. For general nonlinear transformation, Liang et al. (2023) identify intervention-target blocks when the causal graph and target sets are known and sufficiently diverse perfect interventions satisfy a block score-rank condition. Ahuja et al. (2024) use ancestrally closed soft intervention target sets (with proper distributional assumptions) to disentangle the non-intervened variables from the intervened ones. Yao et al. (2025) establish a broader *invariant-block* identification result by combining information preservation with an invariance condition. Our Theorem 1 establishes recovery of the intervention-affected block through collective score diversity, even when its individual coordinates cannot be separated. The discussion following the theorem compares the required intervention structure and diversity conditions with these earlier guarantees.

Visual robot pose estimation. Visual robot pose estimation is most commonly formulated as a supervised prediction problem. DREAM (Lee et al., 2020) estimates robot pose through image-space keypoints, while RoboPose (Labbé et al., 2021) iteratively refines pose and joint-angle estimates. More recent methods include HPE (Ban et al., 2024) and RoboPEPP (Goswami et al., 2025), the latter combining self-supervised visual pretraining with supervised joint-angle regression. These methods demonstrate strong pose-estimation performance, but ultimately rely on pose or joint-state supervision for the prediction task. In contrast, our objective is to recover the underlying physical state representation without per-frame state labels, using variation across intervention environments as the learning signal.

Classical alternatives include geometry- and marker-based systems, such as fiducial-marker approaches (Garrido-Jurado et al., 2014; García-Ruiz et al., 2023). These methods provide a different tradeoff, relying on known geometric structure, calibration, or instrumentation rather than learned latent representations. Our experiments are not intended to replace these specialized pose-estimation pipelines; rather, robot pose provides a controlled embodied setting in which to study whether identifiable-CRL principles can be carried from high-dimensional observations toward state estimation under progressively weaker assumptions.

Limitations. We list the main limitations of our paper.

- Our empirical analysis is limited to simulation studies in a controlled environment. This is a conscious choice in this paper to maximize controllability of the experimental setting. However, in order to demonstrate practicality, CRL field needs to start using real world data as well. This paper shows that such a transition doesn't have to be done abruptly.
- Proving robust estimation for both score estimation and disentanglement stages against using non-i.i.d. data would address V1 (temporal dependence) directly, and are likely a prerequisite for any theoretical treatment of video modality from the interventional CRL perspective.
- An approximate version of Theorem 1 that bounds the degradation of block recovery in terms of the magnitude of off-block score deviations would be required to treat V2 (off-target effects) from a theoretically grounded perspective.
- There exists CRL approaches that can not only tolerate $d_b > n$ but can also identify n. While this would help with addressing V3 (overcomplete bottleneck), our empirical analysis implies that our method is currently robust against using an overcomplete learned representation.
- As shown in Appendix D, violations of the injectivity of the transformation f are very impactful: a consistent occlusion of a joint from an unfavorable camera angle impedes recovery of that joint, and requires a multi-camera setup to resolve.

B Proof of Theorem 1

Proof. Since (h, g) achieve perfect reconstruction and f is an injective diffeomorphism, the composition

$$\tau \triangleq h \circ f : \mathbb{R}^n \to \mathbb{R}^n \tag{12}$$

is a diffeomorphism with $\hat{Z} = \tau(Z)$. The proof consists of showing that τ cannot mix the block $\mathcal{I}$ with its complement.

Due to the score difference function sparsity property in (5), the score difference $d^{(k)}$ is zero outside $\overline{\mathrm{pa}}(\mathcal{I}_k)$. Definition 2 posits $\overline{\mathrm{pa}}(i) \subseteq \mathcal{I}$, which implies $[d^{(k)}(z)]_j = 0$ for any $j \notin \overline{\mathrm{pa}}(\mathcal{I}_k) \subseteq \mathcal{I}$.

Therefore, for any $j \in \mathcal{I}^c$ and k, we have $[d^{(k)}(z)]_j = 0$. Accordingly, matrix $\mathbf{D}(z)$ can be written in block format,

$$\mathbf{D}(z) = \begin{bmatrix} [\mathbf{D}(z)]_{\mathcal{I},:} \\ 0 \end{bmatrix} , \tag{13}$$

where the "diagonal" part $[\mathbf{D}(z)]_{\mathcal{I},:}$ has size $m \times K$. Due to the collective rank diversity condition Definition 3, this diagonal term has full row rank — i.e., $\mathrm{rank}([\mathbf{D}(z)]_{\mathcal{I},:}) = m$ — at almost every z. We note that (6) implies the matrix identity

$$\mathbf{D}_{\hat{Z}}(\hat{Z}) = [J_\tau(z)]^{-\top} \cdot \mathbf{D}(z) . \tag{14}$$

Therefore,

$$\mathbf{D}_{\hat{Z}}(\hat{Z}) = \begin{bmatrix} [[J_\tau(z)]^{-\top}]_{:,\mathcal{I}} & [[J_\tau(z)]^{-\top}]_{:,\mathcal{I}^c} \end{bmatrix} \cdot \begin{bmatrix} [\mathbf{D}(z)]_{\mathcal{I},:} \\ 0 \end{bmatrix} = [[J_\tau(z)]^{-\top}]_{:,\mathcal{I}} \cdot [\mathbf{D}(z)]_{\mathcal{I},:} . \tag{15}$$

The theorem statement enforces the $\hat{\mathcal{I}}^c$ rows of $\mathbf{D}_{\hat{Z}}(\hat{Z})$ to be zero. That is,

$$\mathbf{0}_{(n-m)\times K} = [\mathbf{D}_{\hat{Z}}(\hat{Z})]_{\hat{\mathcal{I}}^c,:} = [[J_\tau(z)^{-\top}]_{\hat{\mathcal{I}}^c,\mathcal{I}} \cdot [\mathbf{D}(z)]_{\mathcal{I},:} . \tag{16}$$

Applying transpose,

$$\mathbf{0}_{K\times(n-m)} = [\mathbf{D}(z)^\top]_{:,\mathcal{I}}[J_\tau(z)^{-1}]_{\mathcal{I},\hat{\mathcal{I}}^c}. \tag{17}$$

Since $[\mathbf{D}(z)]_{\mathcal{I},:}$ has full row rank, $[\mathbf{D}(z)^\top]_{:,\mathcal{I}}$ has full column rank. Hence

$$[J_\tau(z)^{-1}]_{\mathcal{I},\hat{\mathcal{I}}^c} = \mathbf{0}_{m\times(n-m)}. \tag{18}$$

This holds almost everywhere by the collective rank diversity condition, and therefore everywhere by continuity of J_τ^{-1}.

Partitioning the rows of J_τ^{-1} according to $(\mathcal{I}, \mathcal{I}^c)$ and its columns according to $(\hat{\mathcal{I}}, \hat{\mathcal{I}}^c)$, we thus have

$$J_\tau^{-1} = \begin{bmatrix} [J_\tau^{-1}]_{\mathcal{I},\hat{\mathcal{I}}} & \mathbf{0} \\ [J_\tau^{-1}]_{\mathcal{I}^c,\hat{\mathcal{I}}} & [J_\tau^{-1}]_{\mathcal{I}^c,\hat{\mathcal{I}}^c} \end{bmatrix}. \tag{19}$$

Since J_τ^{-1} is invertible, both diagonal blocks are invertible. Consequently, by block matrix inversion,

$$J_\tau = \begin{bmatrix} [J_\tau]_{\hat{\mathcal{I}},\mathcal{I}} & \mathbf{0} \\ [J_\tau]_{\hat{\mathcal{I}}^c,\mathcal{I}} & [J_\tau]_{\hat{\mathcal{I}}^c,\mathcal{I}^c} \end{bmatrix}, \tag{20}$$

where $[J_\tau]_{\hat{\mathcal{I}},\mathcal{I}}$ is invertible. Therefore,

$$\frac{\partial \hat{z}_j}{\partial z_k} = 0 \qquad \forall j \in \hat{\mathcal{I}}, \quad \forall k \in \mathcal{I}^c. \tag{21}$$

Hence $\hat{z}_{\hat{\mathcal{I}}}$ depends only on $Z_{\mathcal{I}}$.

Likewise, the upper-right zero block of J_τ^{-1} implies that $Z_{\mathcal{I}}$ depends only on $\hat{z}_{\hat{\mathcal{I}}}$. Thus, there exists a map $\tau_{\mathcal{I}}$ such that

$$\hat{z}_{\hat{\mathcal{I}}} = \tau_{\mathcal{I}}(Z_{\mathcal{I}}), \tag{22}$$

whose inverse is given by the corresponding restriction of τ^{-1}. Moreover, $[J_\tau]_{\hat{\mathcal{I}},\mathcal{I}}$ is invertible. Therefore $\tau_{\mathcal{I}}$ is a diffeomorphism, proving block disentanglement. □

C Embodied Regime

This appendix provides additional implementation and evaluation details for the embodied setting of Section 4. Unless otherwise stated, all reported experiments use $K = n = 6$ environments, with one *nominally* targeted joint per environment. Hence, the union of the intervention targets is $\mathcal{I} = [n] = \{1, \ldots, 6\}$. We designate the corresponding first six learned coordinates as $\hat{\mathcal{I}} = \{1, \ldots, 6\}$; when $d_b > 6$, the remaining coordinates $\hat{\mathcal{I}}^c = \{7, \ldots, d_b\}$ form the nuisance portion of the learned representation.

Ground-truth joint angles are not used to train AE1, the LDR estimators, or AE2. They are used only for evaluation through the R^2 metrics defined in Eqn. (11). All results and training curves reported in this section correspond to a single fixed random seed and to the final training iterate.

C.1 Data Generation

Data collection replaces i.i.d. rendering with continuous trajectories. We construct six environments, one for each robot joint. Each environment is generated from 4000 closed-loop trajectories and differs through the goal-sampling distribution used for its nominally targeted joint.

Each episode proceeds as follows.

1. Sample a start configuration and a goal configuration. In environment k, the goal distribution prioritizes displacement of joint k (sampled uniformly at random with minimum displacement of 0.5 radians), while the remaining joint targets are changed in a small amount (shifted by a uniformly random value less than 0.05 radians).
2. Drive the arm from its start configuration toward the sampled goal using a proportional controller. Every simulator step is rendered and retained, so consecutive observations differ by one controller step.
3. Terminate the trajectory once the goal is reached within a prescribed tolerance or after a fixed step budget (max. 40 frames), and reset the environment before generating the next trajectory.

We use the same two-camera rendering configuration: one at 45 and other 135 degrees with respect to the arm's natural position, with observations from the two views concatenated after AE1 feature extraction.

Start states are sampled uniformly at random from the valid angles. The simulator itself rejects malformed configurations (e.g., where the robot arm intersects the table), so no additional processing during generation were needed.

The target distribution at the default configuration (the default "embodied regime" setup) is chosen such that in environment k, the k-th joint angle is set to be significantly different (difference ≥ 0.5 radians) from the initial state. The off-target joints are instead targeted with relatively smaller perturbations (< 0.05 radians).

The proportional controller gain is set to 1; however, the trajectories were stable and smooth across the configuration space. Maximum trajectory length was set at 40 frames.

The resulting data exhibit strong temporal dependence within each trajectory. Resets limit dependence across episodes but do not remove the correlation between consecutive frames within one episode. Moreover, targeting one joint does not yield an atomic intervention in the resulting trajectory: closed-loop motion generally induces simultaneous motion of other joints. Thus, the environment label indicates the joint whose displacement is emphasized rather than an intervention that changes only a single latent coordinate.

In the finalized experiments we use all six environments jointly, so $\mathcal{I} = [n]$. Consequently, off-target motion makes the individual environments non-atomic and couples their intervention effects.

C.2 Score Difference Estimation

Unlike the assumption-matched setting, the embodied regime does not contain a separately generated observational dataset. Instead, we exploit the directional structure of each controlled trajectory to construct the two distributions required by the log-density-ratio estimator.

For every trajectory in environment k, we divide the sequence into two non-overlapping temporal halves. Samples from the first half form one classifier class and samples from the second half form the other. We then train one binary LDR classifier $f^{(k)}_{\mathrm{LDR}}$ for each environment. As described in the main pipeline, its logit gradient provides the corresponding feature-space score difference estimate.

In the embodied setting, this temporal-half contrast is an *empirical proxy* motivated by the theoretical score-difference construction, rather than a direct estimate of the interventional score difference in Theorem 1.

We deliberately use non-overlapping temporal halves rather than frame-by-frame contrasts between nearby portions of the same trajectory. At finite sample size, heavily overlapping temporal supports would cause the classifier to observe nearly indistinguishable samples from the two classes over much of the trajectory, thus substantially weakening the density-ratio signal. The half-trajectory construction

instead provides a coarser but more separable contrast while retaining the continuous-video nature of the data.

C.3 Training and Evaluation Details

We follow the three-stage training pipeline described in Section 5. AE1 is first trained to compress each $128 \times 128 \times 1$ camera image to an $8 \times 8 \times 1$ feature representation. The AE1 parameters are then frozen. Features from the two camera views are concatenated to form an $8 \times 8 \times 2$ input to the LDR estimators and AE2.

For the embodied experiments, all trainable stages use Adam optimizer, use learning rate of 10^{-5}, are trained for 10 epochs and using batch size 256. AE1 was trained on 16 H100s whereas LDR and AE2 models were trained on a single H100.

In AE1 and AE2 runs, we process the dataset frame-by-frame as if the data was i.i.d.; consequently, we observe 2.8×10^4 optimization steps in the reported runs. In LDR, trajectories themselves were batched, and for each trajectory one frame from each half is selected uniformly at random to train the model in that epoch. Therefore, the classes were by definition balanced during training.

For AE2, the scale-regularization weight λ_s is fixed at $\lambda_s = 1$. We train bottleneck dimensions

$$d_b \in \{8, 12, 16, 32\}, \tag{23}$$

with the first six coordinates designated as $\hat{\mathcal{I}}$ and the remaining $d_b - 6$ coordinates are treated as nuisance coordinates. The AE2 objective is the block-regularized loss in Eqn. (10). We vary the score-regularization weight λ in the sensitivity experiments below; the bottleneck-dimension sweep uses $\lambda = 50$.

Evaluation protocol. For the embodied setting, we report the three predictability metrics introduced in Eqn. (11): feature predictability $R^2(Z_{\mathcal{I}} \mid Y)$, all-latent predictability $R^2(Z_{\mathcal{I}} \mid \hat{Z})$, and active predictability $R^2(Z_{\mathcal{I}} \mid \hat{Z}_{\hat{\mathcal{I}}})$.

These quantities are estimated using multivariate nonlinear kernel ridge regression with an RBF kernel on held-out labeled states. The kernel bandwidth is selected using the median-distance heuristic, i.e., from the median pairwise distance among the regression inputs. These labels are used only for evaluation and do not enter representation learning. We use two separate sets of 1000 samples each to fit KRR and compute the reported R^2 values.

C.4 Sensitivity and Additional Results

We next examine the two experimental factors for which finalized sweeps are available: the learned bottleneck dimension d_b and the score-regularization weight λ. These experiments are intended to complement the aggregate results of Table 2 by showing the training dynamics underlying the reported final values.

C.4.1 Bottleneck Dimension

Figure 2 shows the evolution of the three predictability quantities as the bottleneck dimension is varied. Increasing d_b substantially increases the capacity available to the complete representation $\hat{Z}$, but active predictability remains comparatively stable. In particular, the information about the true state retained in the designated coordinates $\hat{Z}_{\hat{\mathcal{I}}}$ remains close to that available from the complete learned representation over the tested bottleneck dimensions.

This behavior is consistent with the main result in Table 2: the block regularizer concentrates much of the recoverable state information into the designated six-dimensional subspace even when the learned representation contains substantially more coordinates than the true latent state.

C.4.2 Score-Regularization Weight

Figures 3 to 5 show the training dynamics as the score-regularization weight is varied for $d_b \in \{8, 16, 32\}$. The reconstruction-only baseline, $\lambda = 0$, preserves substantial state information in the

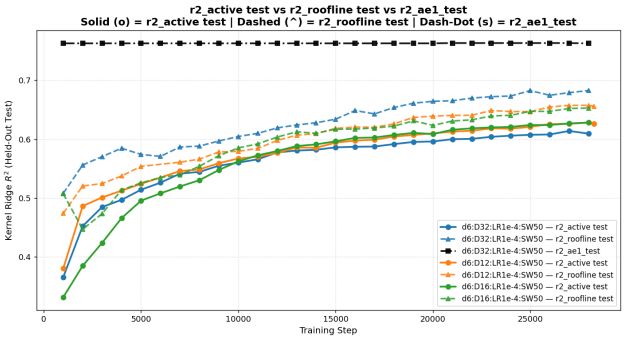


Figure 2: Training dynamics across bottleneck dimensions d_b ($\lambda = 50$).

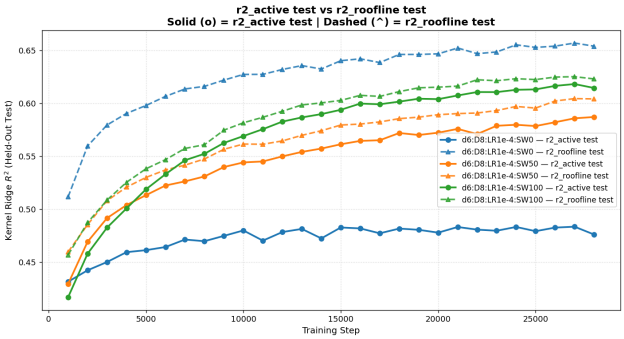


Figure 3: Training dynamics across score-regularization weights λ ($d_b = 8$).

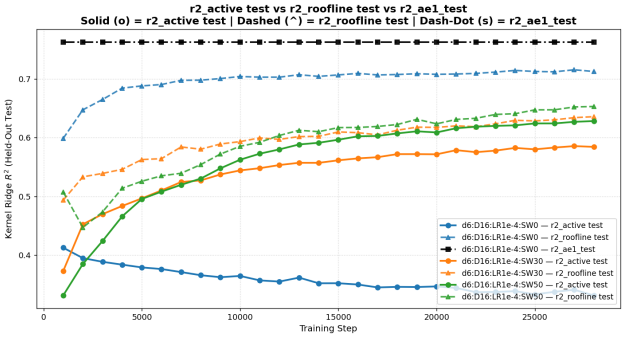


Figure 4: Training dynamics across score-regularization weights λ ($d_b = 16$).

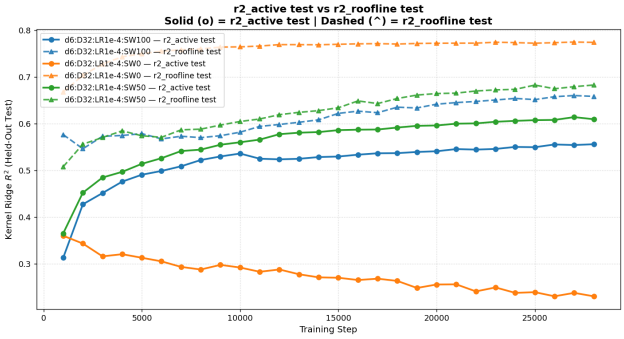


Figure 5: Training dynamics across score-regularization weights λ ($d_b = 32$).

complete representation but does not consistently concentrate that information into $\hat{Z}_{\hat{\mathcal{I}}}$. Introducing the block regularizer substantially closes this gap.

The dependence on λ is not monotone: increasing the regularization weight beyond a useful range does not necessarily improve active predictability. The relevant comparison is therefore between the absence and presence of a substantial block-disentanglement signal, rather than between progressively larger values of λ.

C.4.3 Multi-Dominant-Joint Trajectories

Finally, we consider "multi-dominant-joint trajectories". Specifically, in environment k, we choose *two* joints, k and $k+2 \pmod 6$, as targets. In this regime, we also increase the off-target effects from 0.05 to 0.1.

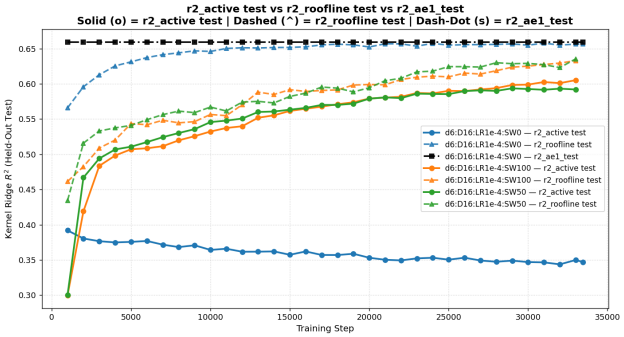


Figure 6: R^2 under multi-dominant-joint data ($d_b = 16$)

In Figure 6, we provide the equivalent of Figure 4 under the multi-dominant-joint data. This experiment increases the extent to which each environment departs from an atomic intervention and therefore provides an additional test of whether the block-based objective remains useful when intervention effects are distributed across multiple coordinates.

The results show that the multi-target setting is almost indistinguishable from the single-dominant-target regime. This is consistent with the theoretical motivation, which does not require atomic interventions. Furthermore, it demonstrates that our implementation is not dependent on the single-node-like score signatures in data.

Overall, the finalized sweeps support the same qualitative conclusion as the main embodied results: the score-based block regularizer concentrates substantially more recoverable state information into the designated coordinates than reconstruction alone, while the result remains relatively insensitive to substantial overparameterization of the learned representation.

D Assumption-Matched Regime

In this section, we additionally report a standard metric in elementwise CRL: Mean Correlation Coefficient (MCC) (Khemakhem et al., 2020). Furthermore, MSE requires the estimated latent variable estimates to be compared against the ground-truth values. In order to do so, we first calibrate the estimates using a held-out set of 1000 labelled samples using elementwise affine regression.

D.1 Data Generation

Observational and Interventional Distributions. Our inference process is unsupervised, meaning we do not require pose-labeled images. The data needed for inference are generated via interventions, which are applied by manipulating the joints individually and capturing the resulting images. Following Section 4.2, for each joint variable Z_i, we have one observational and two interventional distributions, p, q_i, and $\bar{q}_i$. Our learning algorithm is oblivious to any metadata or statistics associated with these distributions and learns from a random collection of poses generated before and after

interventions. We denote the images drawn from q_i and $\bar{q}_i$ by '0' and '1' respectively. Details of these distributions are described next, and exact parameterizations are given in Appendix D.1.1.

Data Generation. We generate our dataset using a Franka Emika Panda arm in a Panda-Gym simulator (Gallouédec et al., 2021). The arm has six primary joint angles, which are marked in Figure 7. Our data generation process consists of two stages. First, we focus on a setup with a single camera, generating interventions for the joints whose movements are confined to the camera's plane, i.e., the 2D projection plane, perpendicular to the camera's viewing direction. In the second stage, to accommodate the out-of-plane motions from other joints, which cannot be captured from a single viewpoint, we expand the dataset using images generated by two camera angles. The images in the dataset are converted to grayscale, resulting in a shape of $128 \times 128 \times 1$ for each image.

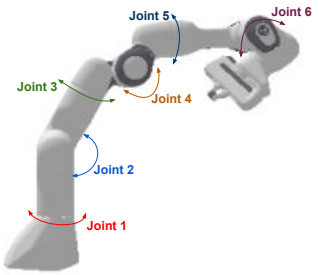


Figure 7: The six primary joint angles of the Franka Emika Panda arm targeted for intervention, together with their respective axes of rotation. Note that the final (7th) joint is kept fixed.

Data on In-Plane Joints using a Single Camera. We first focus on joints 2, 4, and 6, as their movement primarily results in motion within the camera's plane, and a single camera view is sufficient to cover the plane. Each data point in this dataset consists of a set of six interventional images (two for each joint) anchored by a single observational state. This observational state is generated by sampling a configuration for all six joints from a base truncated normal distribution. From this anchor, we create two distinct "hard interventions" for each of the target joints (2, 4, and 6). To perform an intervention on a specific joint, its angle is resampled from a distribution with a mean shifted far from the observational mean, while all other joint angles are held constant. This results in one observational image and six interventional images per data point.

Data on All Joints using Two Cameras. To extend our analysis to all six degrees of freedom, we also need to consider joints 1, 3, and 5, whose actuation causes significant out-of-plane motion. To ensure full observability, we captured every pose from two distinct camera angles ($45°$ and $135°$ yaw). We augment the dataset such that each data point now begins with an observational pose captured from both cameras (2 images). Subsequently, we perform two hard interventions on each of the six joints ($j \in \{1, \ldots, 6\}$). Each of these 12 interventional pose distributions is captured by both camera angles. Thus, a single complete data point in this extended dataset consists of 26 images: 2 observational images plus 24 interventional images (6 joints $\times$ 2 interventions/joint $\times$ 2 cameras/pose).

D.1.1 Independent Interventional Distributions

For the single-camera experiments, a single set of sampling distributions, detailed in Table 3, is used for both training and evaluation. As such, the concepts of in-distribution (ID) and out-of-distribution (OOD) do not apply in this case. For the more comprehensive two-camera experiments, we distinguish between ID and OOD data to evaluate model generalization. The ID dataset, which is used for training the Independent and Occlusion models, is generated using the parameters in Table 4. To test for robustness to distributional shifts, we created a separate OOD test set with modified observational parameters, as shown in Table 5. The experiment with causal model over the joints is also evaluated on this OOD test set. We use a truncated normal distribution, denoted by $\mathcal{TN}_{[a,b]}(\mu, \sigma^2)$, which represents a normal distribution with mean μ and variance σ^2 truncated to the interval $[a, b]$.s

Table 3: Sampling distributions for observational and interventional settings for single camera setup.

Joint	Scenario	Distribution
2	Observational	$\mathcal{TN}_{[-1.5,\,1.5]}(0,\,1)$
	Intervention 1	$\mathcal{TN}_{[-1.5,\,1.5]}(-0.75,\,0.5)$
	Intervention 2	$\mathcal{TN}_{[-1.5,\,1.5]}(\ 0.75,\,0.5)$
4	Observational	$\mathcal{TN}_{[-1.5,\,1.5]}(0,\,1)$
	Intervention 1	$\mathcal{TN}_{[-1.5,\,1.5]}(-0.75,\,0.5)$
	Intervention 2	$\mathcal{TN}_{[-1.5,\,1.5]}(\ 0.75,\,0.5)$
6	Observational	$\mathcal{TN}_{[0,\,3]}(1.5,\,1)$
	Intervention 1	$\mathcal{TN}_{[0,\,3]}(2.25,\,0.5)$
	Intervention 2	$\mathcal{TN}_{[0,\,3]}(\ 0.75,\,0.5)$

Table 4: Sampling distributions for the in-distribution (ID) dataset. These truncated normal distributions define the observational and interventional data used for the two-camera Independent and Occlusion experiments.

Joint	Scenario	Distribution
1	Observational	$\mathcal{TN}_{[0,\,3]}(1.2,\,0.4)$
	Intervention 1	$\mathcal{TN}_{[0,\,3]}(2.0,\,0.4)$
	Intervention 2	$\mathcal{TN}_{[0,\,3]}(\ 0.6,\,0.4)$
2	Observational	$\mathcal{TN}_{[-1.5,\,1.5]}(0,\,0.4)$
	Intervention 1	$\mathcal{TN}_{[-1.5,\,1.5]}(0.7,\,0.4)$
	Intervention 2	$\mathcal{TN}_{[-1.5,\,1.5]}(\ -0.7,\,0.4)$
3	Observational	$\mathcal{TN}_{[-1.5,\,1.5]}(0,\,0.4)$
	Intervention 1	$\mathcal{TN}_{[-1.5,\,1.5]}(0.7,\,0.4)$
	Intervention 2	$\mathcal{TN}_{[-1.5,\,1.5]}(\ -0.7,\,0.4)$
4	Observational	$\mathcal{TN}_{[-1.5,\,1.5]}(0,\,0.4)$
	Intervention 1	$\mathcal{TN}_{[-1.5,\,1.5]}(0.9,\,0.4)$
	Intervention 2	$\mathcal{TN}_{[-1.5,\,1.5]}(\ -0.9,\,0.4)$
5	Observational	$\mathcal{TN}_{[-1.5,\,1.5]}(0,\,0.4)$
	Intervention 1	$\mathcal{TN}_{[-1.5,\,1.5]}(0.9,\,0.4)$
	Intervention 2	$\mathcal{TN}_{[-1.5,\,1.5]}(\ -0.9,\,0.4)$
6	Observational	$\mathcal{TN}_{[0,\,3]}(1.5,\,0.4)$
	Intervention 1	$\mathcal{TN}_{[0,\,3]}(2.4,\,0.4)$
	Intervention 2	$\mathcal{TN}_{[0,\,3]}(\ 0.7,\,0.4)$

D.1.2 Causal Dataset Generation

The observational data is generated from a linear structural equation model (SEM) depicted in Figure 8. Specifically, we first sample the root node joints J_1, J_2, J_4 from the relevant distributions in Table 6. Then, values of $\{J_1, \ldots, J_6\}$ are determined by linear functions of their ancestors using the weights provided in Figure 8 plus noise $(\epsilon_1, \ldots, \epsilon_6)$ sampled from $\mathcal{TN}_{[0,\,1]}(0.0,\,0.1)$ distribution.

$$
\begin{aligned}
J_1 &= J_1 + \epsilon_1 \qquad & J_3 &= 0.88J_2 + \epsilon_3 \\
J_2 &= J_2 + \epsilon_2 \qquad & J_5 &= 0.26J_3 + \epsilon_5 \\
J_4 &= J_4 + \epsilon_4 \qquad & J_6 &= 0.24J_1 + 0.31J_2 + 0.37J_3 + 0.15J_5 + \epsilon_6
\end{aligned}
$$

Table 5: Observational sampling distributions for the out-of-distribution (OOD) dataset. These parameters define the OOD test sets for the Two-Camera Independent, Causal, and Occlusion experiments. The interventional distributions for the OOD dataset remain identical to those of the in-distribution dataset, as defined in Table 4.

Joint	OOD Observational Distribution
1	$\mathcal{TN}_{[0,\,3]}(1.2,\,0.4)$
2	$\mathcal{TN}_{[-1.5,\,1.5]}(0.0,\,0.4)$
3	$\mathcal{TN}_{[-1.5,\,1.5]}(0.0,\,0.4)$
4	$\mathcal{TN}_{[-1.5,\,1.5]}(0.8,\,0.4)$
5	$\mathcal{TN}_{[-1.5,\,1.5]}(0.0,\,0.4)$
6	$\mathcal{TN}_{[0,\,3]}(0.5,\,0.4)$

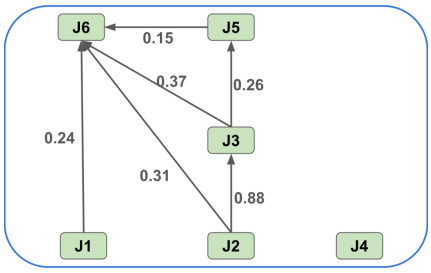


Figure 8: The causal model of the robot joints used to generate the dataset. In this graph, J*i* represents the angle of joint *i*.

An intervention on a variable J_k is denoted by setting it to a new value J'_k sampled from the interventional distribution of choice in Table 6 — this works because we consider stochastic hard interventions that sets the interventional values directly equal to the exogenous noise variables. Next, we provide the equations that describe the system's behavior for interventions on each joint.

Intervention on joint 1 Under $do(J_1 := J'_1)$: The equation for J_6 depends on the new value J'_1 and others being unaffected and J_1 takes the value J'_1.

$$J_6 = 0.24J'_1 + 0.31J_2 + 0.37J_3 + 0.15J_5$$

Intervention on joint 2 Under $do(J_2 := J'_2)$: The equations for the descendants of J_2 are updated as given below.

$$J_3 = 0.88J'_2 + \epsilon_3$$
$$J_5 = 0.26J_3 + \epsilon_5$$
$$J_6 = 0.24J_1 + 0.31J'_2 + 0.37J_3 + 0.15J_5 + \epsilon_6$$

Intervention on joint 3 Under $do(J_3 := J'_3)$: The descendants of J_3 are affected.

$$J_5 = 0.26J'_3 + \epsilon_5$$
$$J_6 = 0.24J_1 + 0.31J_2 + 0.37J'_3 + 0.15J_5 + \epsilon_6$$

Intervention on joint 4 Under $do(J_4 := J'_4)$: No downstream variables are affected by an intervention on J_4, suggesting it is a root node with no children in this system.

Intervention on joint 5 Under $do(J_5 := J'_5)$: The equation for J_6 is updated.

$$J_6 = 0.24J_1 + 0.31J_2 + 0.37J_3 + 0.15J'_5 + \epsilon_6$$

Intervention on joint 6 Under $do(J_6 := J'_6)$: No changes occur in any other variables except J_6, as J_6 is a sink node (it does not cause other variables in the system).

Table 6: Sampling distributions for observational and interventional settings for causal dataset corresponding to the causal graph Figure 8

Joint	Scenario	Distribution
1	Observational	$\mathcal{TN}_{[0,\,3]}(1.2,\,0.4)$
	Intervention 1	$\mathcal{TN}_{[0,\,3]}(2.0,\,0.4)$
	Intervention 2	$\mathcal{TN}_{[0,\,3]}(0.6,\,0.4)$
2	Observational	$\mathcal{TN}_{[-1.5,\,1.5]}(0,\,0.4)$
	Intervention 1	$\mathcal{TN}_{[-1.5,\,1.5]}(0.7,\,0.4)$
	Intervention 2	$\mathcal{TN}_{[-1.5,\,1.5]}(-0.7,\,0.4)$
3	Observational	Not a root node
	Intervention 1	$\mathcal{TN}_{[-1.5,\,1.5]}(0.7,\,0.4)$
	Intervention 2	$\mathcal{TN}_{[-1.5,\,1.5]}(-0.7,\,0.4)$
4	Observational	$\mathcal{TN}_{[-1.5,\,1.5]}(0,\,0.4)$
	Intervention 1	$\mathcal{TN}_{[-1.5,\,1.5]}(0.9,\,0.4)$
	Intervention 2	$\mathcal{TN}_{[-1.5,\,1.5]}(-0.9,\,0.4)$
5	Observational	Not a root node
	Intervention 1	$\mathcal{TN}_{[-1.5,\,1.5]}(0.9,\,0.4)$
	Intervention 2	$\mathcal{TN}_{[-1.5,\,1.5]}(-0.9,\,0.4)$
6	Observational	Not a root node
	Intervention 1	$\mathcal{TN}_{[0,\,3]}(2.4,\,0.4)$
	Intervention 2	$\mathcal{TN}_{[0,\,3]}(0.7,\,0.4)$

D.2 Training Details

All models were trained on TPUs. We performed a hyperparameter search for the optimal learning rate, testing values in the range of 10^{-7} to 10^{-3}. Our dataset consists of 10,000 observational images. We generated corresponding interventional images for two experimental setups: In the single-camera setup, each observational image yields 6 interventional images. In the two-camera setup, each observational image yields 24 interventional images. The training process involved three stages. First, Autoencoder-1 was trained on 70k images (single-camera) and 260k images (two-camera) with a batch size of 256. Second, we trained a separate binary classifier network (i.e., the log-ratio density estimator) for each joint, using 20k samples (single-camera) and 40k samples (two-camera) with a batch size of 64. Finally, for each joint, we trained a corresponding Autoencoder-2 alongside its specific LDR. Autoencoder-2 is trained on the same number of samples as Autoencoder-1 for both single camera and two camera setup.

In the main set of results, we use Adam optimizer with learning rates: AE1 with 1×10^{-4}, LDR with 1×10^{-3}, and AE2 with 7×10^{-5}.

D.3 Additional Quantitative Results

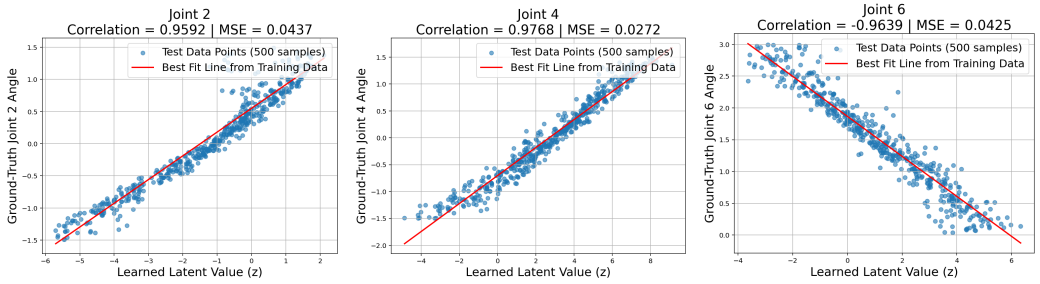
D.3.1 Single-Camera: Independent Joints

We begin by evaluating the model in a single-camera setting, where joint angles are sampled *independently* according to the distributions specified in Table 3. As this setup restricts visibility, we evaluate performance only on the three in-plane joints (2, 4, and 6) discussed in Appendix D.1. Table 7 (first row) shows that MCC is consistently at least 0.94. Remarkably, this disentanglement performance is stronger than that of the prior experiments in CRL studies in much simpler image datasets (e.g., see (Varıcı et al., 2025, Table 14)). A qualitative analysis of the model's reconstruction

Table 7: MCC and MSE in assumption-matched regime. MSE is reported in radians squared.

	Joint 1		Joint 2		Joint 3		Joint 4		Joint 5		Joint 6	
Model Setup	**MCC**	**MSE**	**MCC**	**MSE**	**MCC**	**MSE**	**MCC**	**MSE**	**MCC**	**MSE**	**MCC**	**MSE**
1C, indep.	–	–	0.949	0.053	–	–	0.975	0.029	–	–	0.957	0.049
2C, indep.	0.874	0.083	0.979	0.015	0.634	0.217	0.950	0.035	0.679	0.198	0.884	0.080
2C, causal	0.921	0.058	0.966	0.020	0.788	0.106	0.976	0.019	0.742	0.051	0.756	0.070
2C, indep., occl.	0.844	0.101	0.964	0.025	0.568	0.245	0.884	0.082	0.617	0.225	0.768	0.145

1C = one camera; 2C = two cameras; indep. and causal = joints distribution; occl. = occlusion(32x32).



(a) Joint 2 (b) Joint 4 (c) Joint 6

Figure 9: Single Camera: Scatter-plots of ground-truth vs. estimated angles for joints 2, 4, and 6

performance is provided in Figure 9, presenting the scatter plots of the learned latent variables against the ground-truth angles.

D.3.2 Extending to Multiple Views: Two Cameras with Independent Joints

We extend our analysis to a more complex two-camera setup, which provides the multi-view information necessary to disentangle all six robot joints. The joint angles are again sampled *independently* (see Table 4). As shown in Table 7 (second row) and the scatter plots in Figure 10, the model successfully disentangles all six joints. However, we observe a performance drop for joints 3 and 5 compared to the others. We attribute this performance discrepancy to less precise score estimates from the LDR network, an interpretation supported by the fact that these two joints exhibited a relatively larger classification loss during LDR training. This suggests that the images of the hard interventions for joints 3 and 5 are less distinct, making them inherently more challenging to classify. Notably, the MCC scores for joints 2, 4, and 6 remain robust when transitioning from the single-camera to the two-camera setup. The stability of these scores suggests that the score-based method scales effectively, leveraging the multi-view information. Reconstructed images from the AE1 and AE2 are shown in the Figure 21.

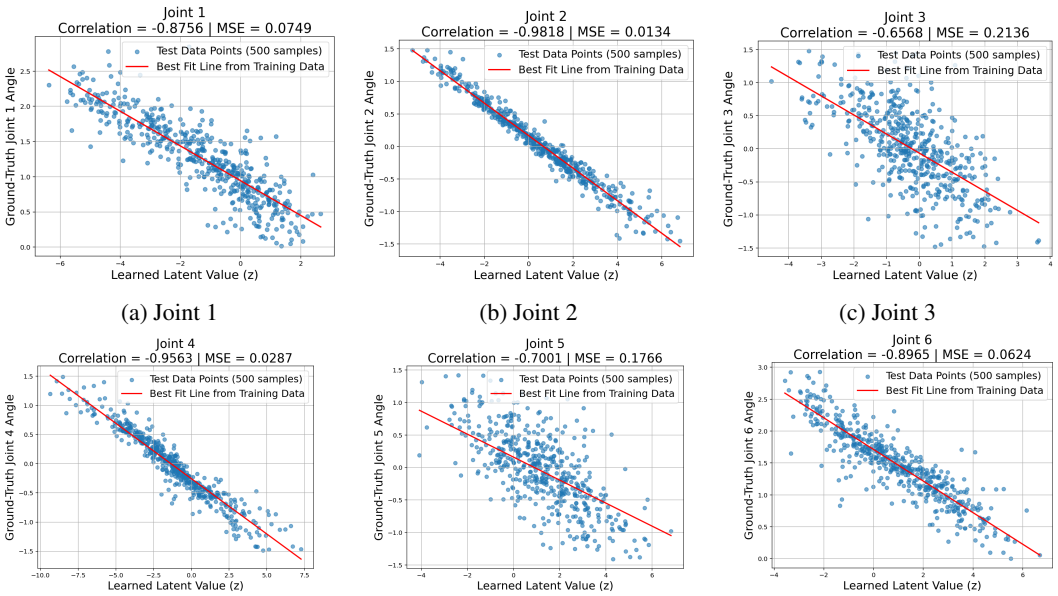
D.3.3 Two-Camera with Causal Model over Joints

To assess whether our method can learn *causal* latents, which is the core premise of CRL, we generate a new dataset where joint angles are sampled according to a linear causal model described in Appendix D.1.2. The structure and edge weights were chosen randomly, with joints 1, 2, and 4 set as root nodes. We report the results in Table 7 (third row). We observe that introducing a causal data generation process generally improves performance, with MSE values decreasing across most joints compared to the independent two-camera model. Overall, this experiment demonstrates the flexibility of the score-based approach to accommodate causal interactions among the joints.

D.3.4 Scatter Plots

To provide a qualitative assessment of our models' disentanglement capabilities, we visualize the relationship between the learned latent variables and their corresponding ground-truth joint angles.

Figures 10 and 11 present these scatter plots for the two-camera Independent and Causal models, respectively. Each plot is generated from the single best trial (out of 15) for that specific joint, determined by the highest MCC score. This visualization of the best-case performance complements the aggregated statistics reported in our main results tables.



(a) Joint 1 (b) Joint 2 (c) Joint 3

(d) Joint 4 (e) Joint 5 (f) Joint 6

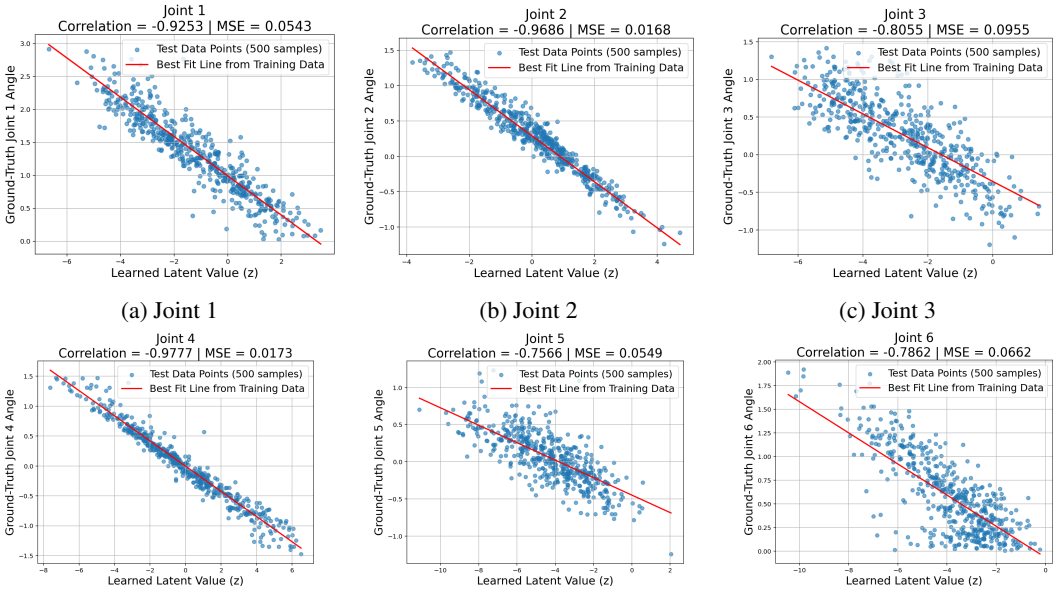
Figure 10: Scatter plots evaluating the two-camera Independent model on the in-distribution (ID) test set (Table4) . Each plot visualizes the relationship between a learned latent variable and its corresponding ground-truth joint angle. The displayed Correlation and MSE values correspond to the single best trial out of 15 runs, while the results presented in Tables 7 and 9 correspond to the mean statistics.

Table 8: Mean Squared Error (MSE) in radians squared (rad^2) for each joint under various experimental conditions for two camera angles. The table compares performance on in-distribution (ID) test set (Table4) and out-of-distribution (OOD) test set (Table5) for independent, causal, and occluded inference models.

Experiment	Distribution	Patch Size	Joint 1	Joint 2	Joint 3	Joint 4	Joint 5	Joint 6
Independent	ID	–	0.083	0.015	0.217	0.035	0.198	0.080
Causal	ID	–	0.058	0.020	0.106	0.019	0.051	0.070
Occlusion	ID	16	0.089	0.018	0.225	0.044	0.200	0.082
Occlusion	ID	32	0.101	0.025	0.245	0.082	0.225	0.145
Occlusion	ID	64	0.186	0.077	0.322	0.258	0.298	0.273
Independent	OOD	–	0.084	0.017	0.239	0.048	0.199	0.120
Causal	OOD	–	0.108	0.044	0.241	0.085	0.219	0.116
Occlusion	OOD	16	0.092	0.019	0.249	0.054	0.205	0.109
Occlusion	OOD	32	0.103	0.024	0.288	0.079	0.226	0.140
Occlusion	OOD	64	0.201	0.087	0.356	0.227	0.331	0.216

D.4 Supervised Comparison and Robustness

RoboPEPP: Label Supervision Baseline. For our experiments, we also compare against a recent supervised baseline. Specifically, we use RoboPEPP (Goswami et al., 2025), which is among the few alternatives that can predict pose with unknown joint angles at test time (along with HPE (Ban et al., 2024) and RoboPOSE (Labbé et al., 2021)), and reports stronger results than the alternatives.



(a) Joint 1 (b) Joint 2 (c) Joint 3

(d) Joint 4 (e) Joint 5 (f) Joint 6

Figure 11: Scatter plots evaluating the two-camera causal model on the causally-generated test set. Each plot visualizes the relationship between a learned latent variable and its corresponding ground-truth joint angle. The displayed Correlation and MSE values correspond to the single best trial out of 15 runs, while the results presented in Tables 7 and 9 correspond to the mean statistics.

Table 9: Comparison of MCC and MSE (rad^2) for each joint across two model settings with error bars. Mean and Std Dev are calculated across the 15 runs. Values are reported as Mean ± Std Dev.

	2C, indep.		**2C, causal**	
Joint Angle	**MCC**	**MSE**	**MCC**	**MSE**
Joint 1	0.874 ± 0.004	0.083 ± 0.006	0.921 ± 0.003	0.058 ± 0.003
Joint 2	0.979 ± 0.001	0.015 ± 0.001	0.966 ± 0.002	0.020 ± 0.001
Joint 3	0.634 ± 0.010	0.217 ± 0.011	0.788 ± 0.003	0.106 ± 0.005
Joint 4	0.950 ± 0.002	0.035 ± 0.002	0.976 ± 0.001	0.019 ± 0.001
Joint 5	0.679 ± 0.010	0.198 ± 0.011	0.742 ± 0.006	0.051 ± 0.003
Joint 6	0.884 ± 0.005	0.080 ± 0.007	0.756 ± 0.010	0.070 ± 0.003

2C = two cameras; indep. and causal refer to the joint distribution model.

RoboPEPP involves a two-stage training process. First, a self-supervised I-JEPA backbone is pre-trained on the entire independent ID training dataset in Table 4. In the second stage, a "JointNet" is trained to predict six joint angles. Its input is formed by concatenating the I-JEPA embeddings from two camera angles, and it is optimized using an L_2 loss with the ground-truth joint angle labels. We observe that the MSE of the RoboPEPP baseline decreases with increasing size of labeled data, with our model achieving a comparable performance on joints 1, 2, 4, and 6 to RoboPEPP when using labels of 5% (approx. 6.5K labels and 2 camera orientation per pose = 13K samples) of the entire dataset for training.

We pre-train a ViT-Huge model using the I-JEPA objective on 128x128x3 images. The core masking strategy involves one large context block (85-100% scale) and four smaller, non-overlapping target blocks (15-20% scale). The model is trained for 10 epochs with a batch size of 32, using a cosine learning rate schedule with a peak learning rate of $1e-3$. The target encoder's weights are updated with an exponential moving average (EMA). Notably, we only use random resized cropping for data augmentation, disabling strong augmentations like color jitter and horizontal flipping. Training is accelerated with bfloat16 mixed-precision.Our training is a two-stage process. In the first stage, we pre-train a ViT-Huge backbone using the I-JEPA self-supervised objective on the full dataset,

comprising 260k images, as described previously. For the second stage, we fine-tune our JointNet model for a robotic perception task on the Panda robot. This supervised training uses the smaller RoboPEPP dataset (1%, 5%, 10%, 100%) with a batch size of 8. The model is trained for (100, 75, 25, 15) epochs respectively using a one-cycle learning rate schedule with a maximum learning rate of $1e-4$ and a weight decay of $1e-5$. Tables 8 to 12 show a detailed comparison of the our and RoboPEPP performance.

Table 10: Per-joint Mean Squared Error (MSE) for the RoboPEPP model in radians squared (rad^2). The table presents an ablation study on the effect of varying training data labels, evaluated on both in-distribution (ID) test set (Table4) and out-of-distribution (OOD) test set (Table5).

Dataset	Distribution	Epochs	Patch Size	Joint 1	Joint 2	Joint 3	Joint 4	Joint 5	Joint 6
2.6k	ID	100	-	0.136	0.039	0.237	0.053	0.253	0.200
2.6k	ID	100	16	0.166	0.030	0.287	0.072	0.309	0.218
2.6k	ID	100	32	0.186	0.060	0.305	0.125	0.331	0.263
2.6k	ID	100	64	0.412	0.358	0.462	1.328	0.515	0.573
2.6k	OOD	100	-	0.202	0.054	0.374	0.101	0.378	0.454
2.6k	OOD	100	16	0.220	0.058	0.413	0.117	0.412	0.432
2.6k	OOD	100	32	0.235	0.085	0.480	0.164	0.346	0.461
2.6k	OOD	100	64	0.496	0.334	0.538	1.156	0.529	0.570
13k	ID	75	-	0.075	0.017	0.134	0.024	0.153	0.081
13k	ID	75	16	0.081	0.021	0.155	0.040	0.189	0.111
13k	ID	75	32	0.140	0.050	0.304	0.198	0.299	0.232
13k	ID	75	64	0.446	0.474	0.648	2.088	0.976	0.501
13k	OOD	75	-	0.083	0.023	0.202	0.063	0.275	0.225
13k	OOD	75	16	0.127	0.040	0.304	0.088	0.321	0.388
13k	OOD	75	32	0.162	0.058	0.445	0.299	0.582	0.718
13k	OOD	75	64	0.489	0.412	0.489	1.926	1.348	1.255
26k	ID	25	-	0.030	0.010	0.072	0.022	0.091	0.063
26k	ID	25	16	0.034	0.011	0.071	0.037	0.093	0.073
26k	ID	25	32	0.077	0.036	0.194	0.180	0.181	0.136
26k	ID	25	64	0.375	0.380	0.529	2.808	0.524	1.474
26k	OOD	25	-	0.064	0.020	0.125	0.037	0.173	0.152
26k	OOD	25	16	0.050	0.019	0.126	0.055	0.197	0.146
26k	OOD	25	32	0.101	0.066	0.230	0.262	0.308	0.202
26k	OOD	25	64	0.394	0.297	0.497	2.475	0.582	1.028
260k	ID	15	-	0.003	0.001	0.007	0.003	0.010	0.011
260k	ID	15	16	0.016	0.004	0.026	0.007	0.030	0.011
260k	ID	15	32	0.066	0.036	0.097	0.053	0.090	0.045
260k	ID	15	64	0.268	0.517	0.593	1.891	0.478	0.430
260k	OOD	15	-	0.007	0.003	0.013	0.004	0.018	0.019
260k	OOD	15	16	0.029	0.008	0.059	0.014	0.081	0.028
260k	OOD	15	32	0.130	0.060	0.210	0.088	0.201	0.125
260k	OOD	15	64	0.262	0.447	0.649	1.437	0.465	0.616

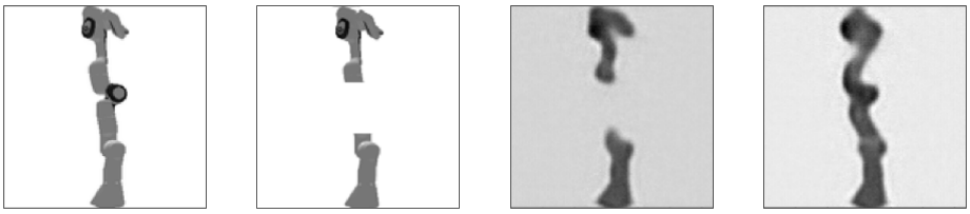
Robustness to Occlusion. Finally, we evaluate the score-based method's robustness to corruptions using the trained two-camera model with independent joints. At test time, we introduce artificial occlusions in the form of 32x32 white pixel squares into the input images, and report results for our method and RoboPEPP in Table 10. Our method was not trained on data containing occlusions or any infilling task. Despite this, our method demonstrates superior robustness to occlusions at test time. This is evident from the MSE of joints 2 and 4, where the performance degradation for our method is significantly less than for RoboPEPP when occlusions are introduced. Furthermore, our method maintains a lower overall MSE on these joints at both 10% and 100% training data labels, highlighting its robustness to this unseen perturbation. Figure 12 visualizes the process. While the initial reconstruction from AE1 clearly shows the occluded patch, the final reconstruction from AE2 successfully inpaints the missing region by leveraging the learned disentangled representation. Table 7 (fourth row) presents the quantitative results for this condition, demonstrating strong performance

Table 11: Comparison of In-Distribution (ID) Mean Squared Error (MSE) in radians squared (rad^2) for the our score-based method and RoboPEPP models. Results are shown per joint under various experimental conditions.

Model	Experiment	Patch Size	Joint 1	Joint 2	Joint 3	Joint 4	Joint 5	Joint 6
Our method	Independent	–	0.083	0.015	0.217	0.035	0.198	0.080
	Causal	–	0.058	0.020	0.106	0.019	0.051	0.070
	Occlusion	16	0.089	0.018	0.225	0.044	0.200	0.082
	Occlusion	32	0.101	0.025	0.245	0.082	0.225	0.145
	Occlusion	64	0.186	0.077	0.322	0.258	0.298	0.273
RoboPEPP	2.6k Dataset	–	0.136	0.039	0.237	0.053	0.253	0.200
	2.6k Dataset	16	0.166	0.030	0.287	0.072	0.309	0.218
	2.6k Dataset	32	0.186	0.060	0.305	0.125	0.331	0.263
	2.6k Dataset	64	0.412	0.358	0.462	1.328	0.515	0.573
	13k Dataset	–	0.075	0.017	0.134	0.024	0.153	0.081
	13k Dataset	16	0.081	0.021	0.155	0.040	0.189	0.111
	13k Dataset	32	0.140	0.050	0.304	0.198	0.299	0.232
	13k Dataset	64	0.446	0.474	0.648	2.088	0.976	0.501
	26k Dataset	–	0.030	0.010	0.072	0.022	0.091	0.063
	26k Dataset	16	0.034	0.011	0.071	0.037	0.093	0.073
	26k Dataset	32	0.077	0.036	0.194	0.180	0.181	0.136
	26k Dataset	64	0.375	0.380	0.529	2.808	0.524	1.474
	260k Dataset	–	0.003	0.001	0.007	0.003	0.010	0.011
	260k Dataset	16	0.016	0.004	0.026	0.007	0.030	0.011
	260k Dataset	32	0.066	0.036	0.097	0.053	0.090	0.045
	260k Dataset	64	0.268	0.517	0.593	1.891	0.478	0.430

Table 12: Comparison of Out-of-Distribution (OOD) Mean Squared Error (MSE) in radians squared (rad^2) for our method and RoboPEPP. Results are shown per joint under various experimental conditions.

Model	Experiment	Patch Size	Joint 1	Joint 2	Joint 3	Joint 4	Joint 5	Joint 6
Our method	Independent	–	0.084	0.017	0.239	0.048	0.199	0.120
	Causal	–	0.108	0.044	0.241	0.085	0.219	0.116
	Occlusion	16	0.092	0.019	0.249	0.054	0.205	0.109
	Occlusion	32	0.103	0.024	0.288	0.079	0.226	0.140
	Occlusion	64	0.201	0.087	0.356	0.227	0.331	0.216
RoboPEPP	2.6k Dataset	–	0.202	0.054	0.374	0.101	0.378	0.454
	2.6k Dataset	16	0.220	0.058	0.413	0.117	0.412	0.432
	2.6k Dataset	32	0.235	0.085	0.480	0.164	0.346	0.461
	2.6k Dataset	64	0.496	0.334	0.538	1.156	0.529	0.570
	13k Dataset	–	0.083	0.023	0.202	0.063	0.275	0.225
	13k Dataset	16	0.127	0.040	0.304	0.088	0.321	0.388
	13k Dataset	32	0.162	0.058	0.445	0.299	0.582	0.718
	13k Dataset	64	0.489	0.412	0.489	1.926	1.348	1.255
	26k Dataset	–	0.064	0.020	0.125	0.037	0.173	0.152
	26k Dataset	16	0.050	0.019	0.126	0.055	0.197	0.146
	26k Dataset	32	0.101	0.066	0.230	0.262	0.308	0.202
	26k Dataset	64	0.394	0.297	0.497	2.475	0.582	1.028
	260k Dataset	–	0.007	0.003	0.013	0.004	0.018	0.019
	260k Dataset	16	0.029	0.008	0.059	0.014	0.081	0.028
	260k Dataset	32	0.130	0.060	0.210	0.088	0.201	0.125
	260k Dataset	64	0.262	0.447	0.649	1.437	0.465	0.616



(a) The original input image to the system.

(b) Input with 32×32 pixel white occlusion.

(c) Reconstruction from autoencoder-1.

(d) Reconstruction from autoencoder-2

Figure 12: A step-by-step visualization of the reconstruction process for an occluded input. The final reconstruction from autoencoder-2, generated by passing its output through the decoder of autoencoder-1, successfully inpaints the occluded region.

despite the corruption. A detailed analysis of model performance across varying occlusion sizes is provided in Table 8.

We conclude that AE1 is tasked with high-quality reconstruction, while AE2's loss function sparsifies the score difference. To capture out-of-plane rotations missed by a single camera, an additional viewpoint was essential. This lowered the LDR loss cross-entropy, allowing the model to successfully differentiate all six joint movements.

D.5 Ablations

We conducted a series of experiments to find the proper training setting for our score-based method. Specifically, we investigated the effect of the learning rate, batch size, and weights of the loss components ($\lambda = 3$ in Eqn. (9)) on the model's ability to disentangle the latent features, as measured by the MCC score. All experiments in this section were performed using Autoencoder-2 with a 2-camera setup on the independent dataset for joint 4.

Learning Rate. First, we analyzed the impact of the learning rate. Figure 13 shows the training progress for different learning rates while keeping the batch size fixed at 128 and the weight of the sparsity loss constant ($\lambda = 3$ in Eqn. (9)). We observe that smaller learning rates, such as $5e-5$ and $8e-5$, lead to better performance.

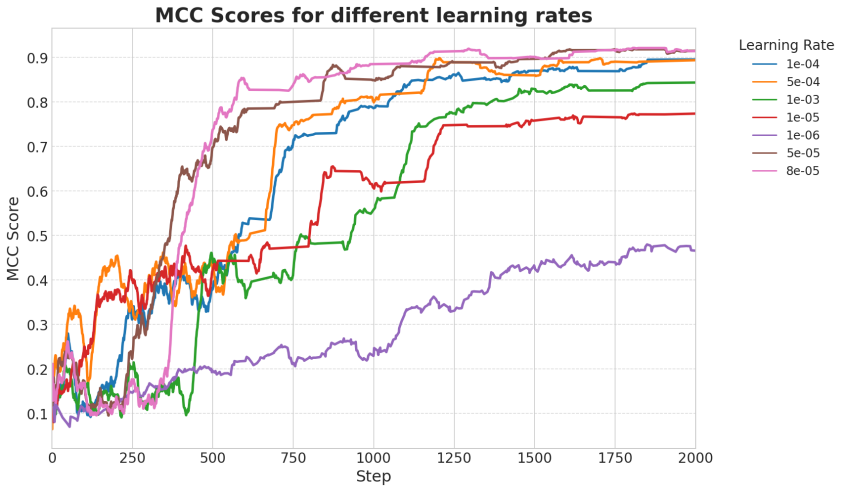


Figure 13: Evolution of MCC scores during training for AE2 using a 2-camera setup on the independent dataset. The plot compares performance across various learning rates. Training is performed with a batch size of 128, and sparsity loss weight of $\lambda = 3$.

Batch Size. Next, we performed an ablation study on the batch size (Figure 14), using the optimal learning rate $5e-5$ found previously. We observed that models with larger batch sizes converged faster; however, their final asymptotic performances were nearly identical, irrespective of the batch size.

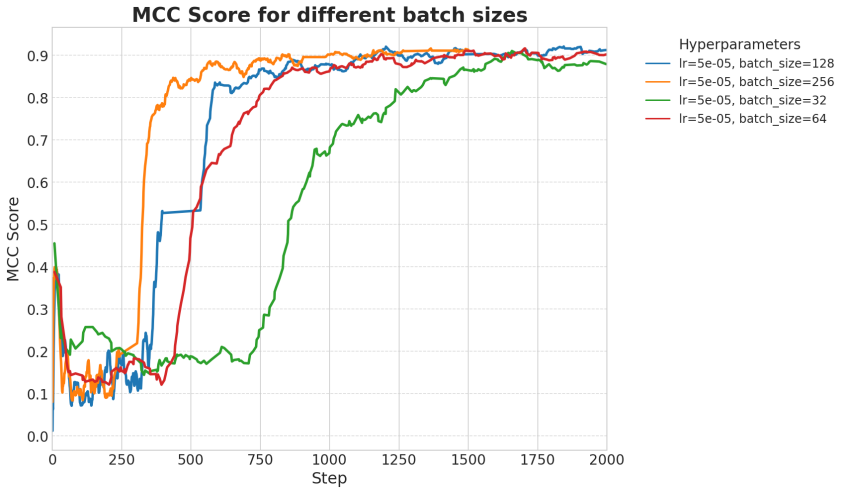


Figure 14: Evolution of MCC scores during training for AE2 using a 2-camera setup on the independent dataset. The plot compares performance across various batch sizes. Training is performed with a learning rate of $5e-5$ and sparsity loss weight of $\lambda = 3$.

Loss Reweighting. We also looked at how to balance the two training objectives: reconstructing the image and the sparsity loss. Figure 15 compares the MCC scores when we vary the weights given to the reconstruction loss versus the sparsity loss. For these experiments, we fixed the learning rate at $5e{-}5$ and batch size at 128. The figure shows that giving a significant weight to the score-based loss (i.e., sparsity loss) is crucial for disentanglement. This aligns with our theory that a reconstruction loss alone cannot perform disentanglement (e.g., MCC of less than 0.15 in the figure), but the score-based loss is the main driver of the disentanglement (e.g., over 0.9 MCC in the figure).

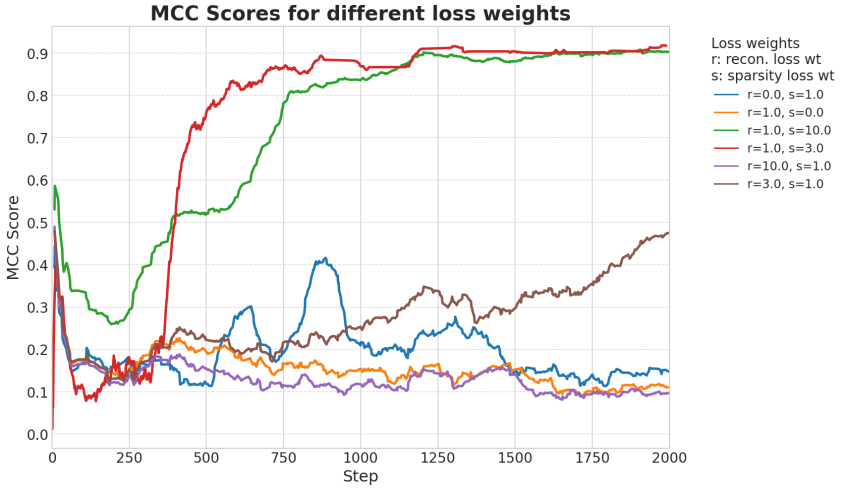


Figure 15: Evolution of MCC scores during training for AE2 using a 2-camera setup on the independent dataset. The plot compares performance across various combinations of reconstruction loss weights and sparsity loss weights. Training is performed with a batch size of 128 and a fixed learning rate of $5e-5$.

Architectural Changes - Residual Connection. The inclusion of intra-module residual connections—applied independently within the encoder and decoder architectures—is a key factor for optimal performance. This architectural choice as seen in the Figure 16 improves the reconstruction loss and hence the final MCC score.

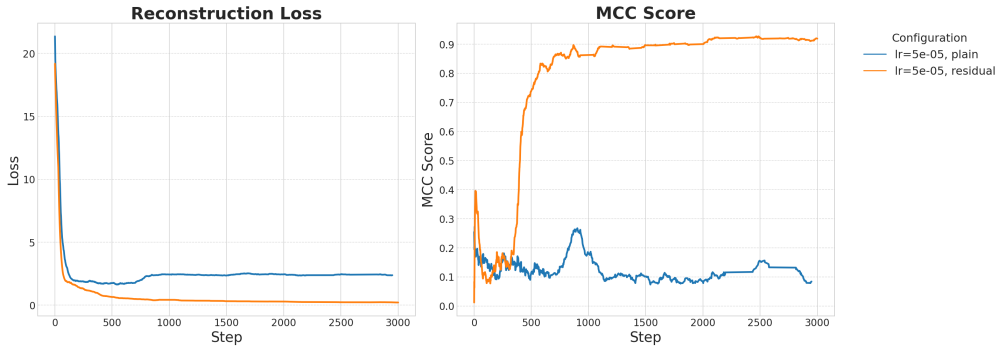


Figure 16: Ablation showing the advantage of having residual connections for better reconstruction loss and MCC score

Calibration Samples. As demonstrated in Table 13, the model's Mean Squared Error (MSE) remains stable even when the number of calibration samples is reduced from 1000 to 100.

Table 13: MCC and MSE of our method across different Calibration settings. MSE is reported in radians squared.

	Joint 1		Joint 2		Joint 3		Joint 4		Joint 5		Joint 6	
Calibration Samples	**MCC**	**MSE**	**MCC**	**MSE**	**MCC**	**MSE**	**MCC**	**MSE**	**MCC**	**MSE**	**MCC**	**MSE**
1000 samples	0.874	0.083	0.979	0.015	0.634	0.217	0.950	0.035	0.679	0.198	0.884	0.080
500 samples	0.863	0.094	0.979	0.015	0.618	0.234	0.945	0.037	0.677	0.191	0.895	0.075
100 samples	0.881	0.079	0.975	0.019	0.612	0.227	0.942	0.039	0.661	0.199	0.879	0.084

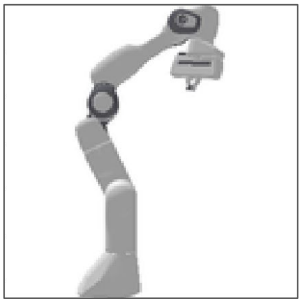
Lighting Condition. Table 14 evaluates the model's performance under distinct lighting scenarios, simulated by scaling pixel brightness to 0.8x (dark) and 1.25x (light) of the original. Performance degradation is most pronounced in the dark, particularly for Joint 5 and Joint 6, which exhibit a significant drop in MCC. In contrast, the remaining joints maintain strong performance with only minimal degradation under both conditions.

Table 14: MCC and MSE of our method across different lighting condition. MSE is reported in radians squared.

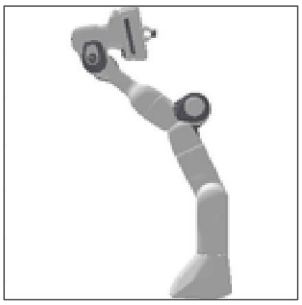
	Joint 1		Joint 2		Joint 3		Joint 4		Joint 5		Joint 6	
Lighting Condition	**MCC**	**MSE**	**MCC**	**MSE**	**MCC**	**MSE**	**MCC**	**MSE**	**MCC**	**MSE**	**MCC**	**MSE**
original	0.874	0.083	0.979	0.015	0.634	0.217	0.950	0.035	0.679	0.198	0.884	0.080
bright	0.874	0.072	0.978	0.015	0.640	0.207	0.950	0.033	0.677	0.180	0.874	0.094
dark	0.811	0.107	0.940	0.046	0.520	0.211	0.860	0.093	0.199	0.303	0.526	0.265

D.6 Qualitative Examples

To provide a qualitative understanding of our dataset, Figure 18 and Figure 19 visualize the hard intervention images that form the basis of our training data. Figure 18 shows intervention images for Joint 4 from the single-camera dataset. Figure 19 shows intervention images for Joint 3 from the two-camera dataset. Since Joint 3 moves out of the plane, these images highlight the necessity of our two-camera setup.

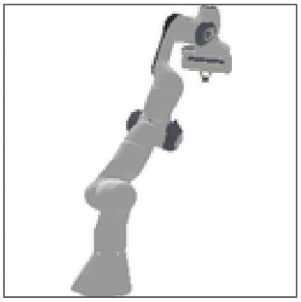


(a) Intervention 1 (label 0)

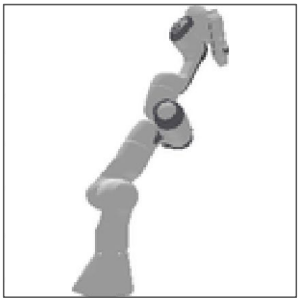


(b) Intervention 2 (label 1)

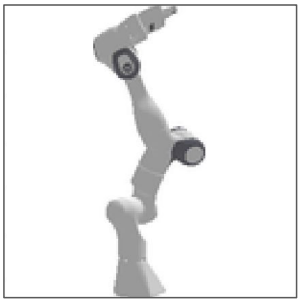
Figure 18: Two different hard interventions on Joint 4.



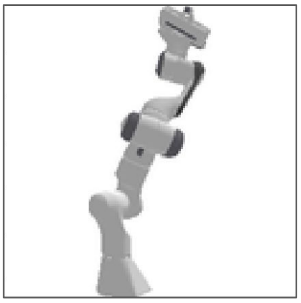
(a) Intervention 1 (45°)



(b) Intervention 2 (45°)



(c) Intervention 1 (135°)



(d) Intervention 2 (135°)

Figure 19: Two different hard interventions on Joint 3, each shown from two camera angles.

Figures 20 and 21 provide a qualitative analysis of our pipeline's reconstruction performance. Specifically, Figures 20b and 20c compare an original image (Figure 20a) to its reconstructions from AE1 and AE2 respectively for the single-camera setup, while Figure 21 shows the equivalent comparison for the two-camera setup. We observe a slight degradation in the reconstruction quality of AE1 when trained on the two-camera dataset compared to the single-camera setup. As the final autoencoder, AE2, is trained on the feature maps from AE1, this reduction in quality of AE1 consequently affects the quality of the final AE2 reconstructions as well. We hypothesize that this performance difference is attributable to the increased data complexity of the multi-view dataset. The inclusion of multiple viewpoints introduces greater visual variance, presenting a more challenging reconstruction task for the autoencoder compared to the more constrained single-view data.

E Architecture Details

The architectural details of the first autoencoder (AE1), the log-density ratio network (LDR) and the second autoencoder (AE2) are presented in Tables 15 to 17, respectively.

Table 15: Autoencoder1 architecture ResNet-style with GroupNorm

Component	Layer-wise Details
Block Def.	**ResBlockGN(f)**: GroupNorm → ReLU → Conv(features=f, ks=3, pad='SAME') → GroupNorm → ReLU → Conv(features=f, ks=3, pad='SAME') → Add residual input *(Note: 'ks'=kernel size, 's'=stride, 'pad'='SAME')*
Encoder	**Input: Image** $X \in \mathbb{R}^{128\times128\times1}$ Conv(features=64, ks=3, pad='SAME') ResBlockGN(64) x 2 Conv(features=64, ks=3, s=2, pad='SAME'), ReLU *// Downsample 128 → 64* Conv(features=128, ks=3, pad='SAME') ResBlockGN(128) x 2 Conv(features=128, ks=3, s=2, pad='SAME'), ReLU *// Downsample 64 → 32* Conv(features=256, ks=3, pad='SAME') ResBlockGN(256) x 2 Conv(features=256, ks=3, s=2, pad='SAME'), ReLU *// Downsample 32 → 16* Conv(features=512, ks=3, pad='SAME') ResBlockGN(512) x 2 Conv(features=1, ks=3, s=2, pad='SAME'), ReLU *// Downsample 16 → 8* **Output: Feature map** $Y \in \mathbb{R}^{8\times8\times1}$
Decoder	**Input: Feature map** $Y \in \mathbb{R}^{8\times8\times1}$ Conv(features=512, ks=3, pad='SAME') ResBlockGN(512) x 2 ConvTranspose(features=512, ks=4, s=2, pad='SAME'), ReLU *// Upsample 8 → 16* Conv(features=256, ks=3, pad='SAME') ResBlockGN(256) x 2 ConvTranspose(features=256, ks=4, s=2, pad='SAME'), ReLU *// Upsample 16 → 32* Conv(features=128, ks=3, pad='SAME') ResBlockGN(128) x 2 ConvTranspose(features=128, ks=4, s=2, pad='SAME'), ReLU *// Upsample 32 → 64* Conv(features=64, ks=3, pad='SAME') ResBlockGN(64) x 2 ConvTranspose(features=64, ks=4, s=2, pad='SAME'), ReLU *// Upsample 64 → 128* **Conv(features=1, ks=3, pad='SAME'), ReLU** *// Final convolution to 1 channel* Reshape to (batch, $128 \times 128 \times 1$) **Output: Reconstructed Image** $\hat{X} \in \mathbb{R}^{128\times128\times1}$

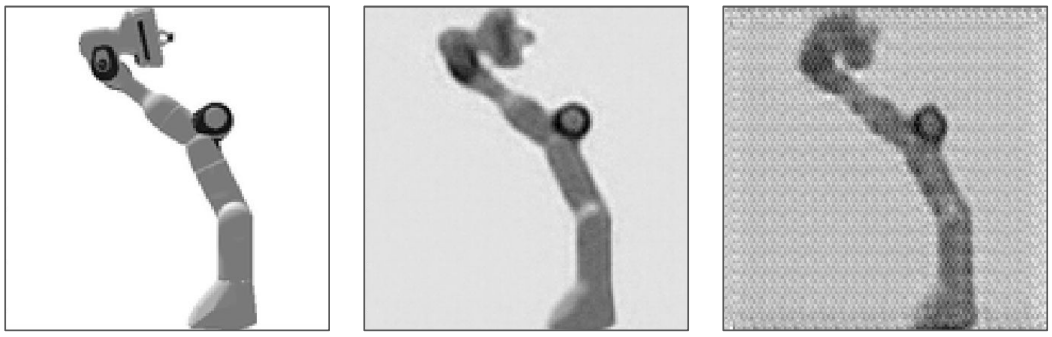
Table 16: LDR Network Architectures for Single- and Two-Camera Setups.

Component	Layer-wise Details
Input Processing	The input shape depends on the camera setup: **Single-Camera:** Input $y \in \mathbb{R}^{8\times8\times1}$ is used directly. **Two-Camera:** Two feature maps y are concatenated to form an input $\in \mathbb{R}^{8\times8\times2}$.
Core Architecture	The following layers are applied to the processed input: Conv(features=32, ks=3), ReLU *// Spatial dim: 8x8 → 6x6* Conv(features=64, ks=3), ReLU *// Spatial dim: 6x6 → 4x4* Conv(features=128, ks=3), ReLU *// Spatial dim: 4x4 → 2x2* Flatten Dense(features=128), ReLU Dense(features=1) **Output: Logit** $\in \mathbb{R}^1$
Training	Minimize binary cross-entropy with logits loss on the output.

Table 17: Autoencoder 2 (AE2) Architecture.

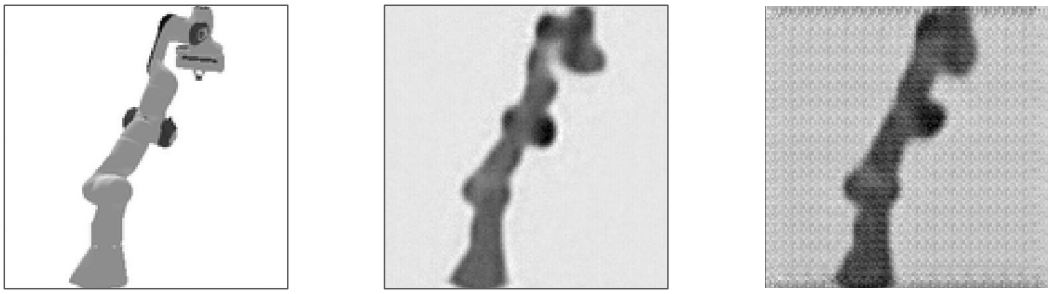
Component	Layer-wise Details
Block Def.	**ResBlockGN(f)**: GroupNorm → ReLU → Conv(features=f, ks=3, pad='SAME') → GroupNorm → ReLU → Conv(features=f, ks=3, pad='SAME') → Add residual input *(Note: 'ks'=kernel size, 's'=stride, 'pad'='SAME')*
Encoder	**Input:** $y \in \mathbb{R}^{8\times8\times C_{in}}$, where C_{in} is 1 (single-cam) or 2 (two-cam). Conv(features=64, ks=3, pad='SAME') ResBlockGN(64) x 2 Conv(features=64, ks=3, s=2, pad='SAME'), ReLU *// Downsample 8 → 4* Conv(features=128, ks=3, pad='SAME') ResBlockGN(128) x 2 Conv(features=128, ks=3, s=2, pad='SAME'), ReLU *// Downsample 4 → 2* Conv(features=256, ks=3, pad='SAME') ResBlockGN(256) x 2 Conv(features=256, ks=3, s=2, pad='SAME'), ReLU *// Downsample 2 → 1* Flatten to (batch, 256) Dense(features=d_b). **Output: Latent** $\hat{z} \in \mathbb{R}^{d_b}$

Decoder	**Input: Latent** $\hat{z} \in \mathbb{R}^{d_b}$
	Dense(features=256), ReLU
	Reshape to (batch, $1 \times 1 \times 256$)
	Conv(features=512, ks=3, pad='SAME')
	ResBlockGN(512) x 2
	ConvTranspose(features=512, ks=4, s=2, pad='SAME'), ReLU *// Upsample 1 → 2*
	Conv(features=256, ks=3, pad='SAME')
	ResBlockGN(256) x 2
	ConvTranspose(features=256, ks=4, s=2, pad='SAME'), ReLU *// Upsample 2 → 4*
	Conv(features=128, ks=3, pad='SAME')
	ResBlockGN(128) x 2
	ConvTranspose(features=128, ks=4, s=2, pad='SAME'), ReLU // Upsample 4 → 8
	Conv(features=C_{in}, ks=3, pad='SAME'), ReLU
	Reshape to (batch, $8 \times 8 \times C_{in}$)
	Output: Reconstructed $\hat{y} \in \mathbb{R}^{8\times8\times C_{in}}$



(a) Observational Image (b) AE1 Reconstruction (c) AE2 Reconstruction

Figure 20: Visual comparison of the reconstruction quality at each stage of our pipeline for the single camera setup. (a) The original input image. (b) The reconstruction from the first autoencoder (AE1). (c) The final reconstruction from the second autoencoder (AE2)



(a) Observational Image (b) AE1 Reconstruction (c) AE2 Reconstruction

Figure 21: Visual comparison of the reconstruction quality at each stage of our pipeline using the two camera independent model. (a) The original input image. (b) The reconstruction from the first autoencoder (AE1). (c) The final reconstruction from the second autoencoder (AE2)