On Causal Representation Learning with Internal Auxiliaries

Kwonho Kim[1] Heejeong Nam[2] Inwoo Hwang[3] Sanghack Lee[1]

[1]Graduate School of Data Science, Seoul National University [2]Brown University

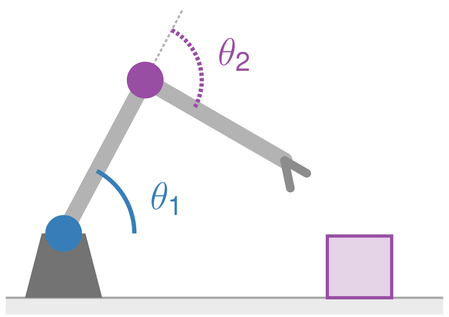
[3]Causal AI Lab, Columbia University

Seoul National University
Graduate School of Data Science

Overview

- **Diagnose.** An observed source coordinate also enters g, so the standard auxiliary-variable proof fails.
- **Identify.** When every observed source is conditioned on ($O = C$), volume preservation (VP), variability, and alignment yield ISA-style subspace identifiability.
- **Select.** A known graph targets finer partitions; a graph-structured VP-flow handles observed-but-unselected sources empirically.

The auxiliary is *inside* the picture



- θ_1 shoulder angle — **observed**
- θ_2 elbow angle + object pose — **targets**

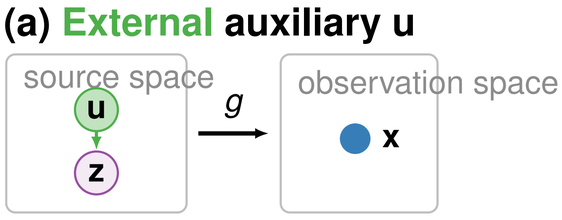
$$\mathbf{x} = g(\underbrace{\theta_1}_{\text{known}}, \underbrace{\theta_2, \text{ object pose}}_{\text{recover}})$$

θ_1 is measured at inference, yet it is also **one of the coordinates the renderer consumed**. Thus the internal auxiliary $\mathbf{o}$ is both known and an input to $\mathbf{x} = g(\mathbf{o}, \mathbf{z})$; the goal is to recover $\mathbf{z}$.

Why the standard proof breaks

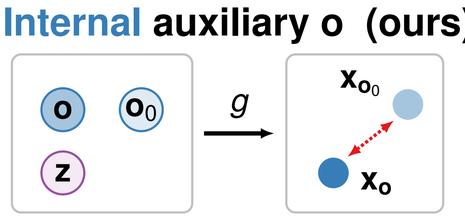
Write $\ell_g(\mathbf{x}) := \log|\det \mathbf{J}_{g^{-1}}(\mathbf{x})|$. There is **one fixed** g; changing context changes where this term is evaluated.

(a) External auxiliary u



Same $\mathbf{x} \Longrightarrow \ell_g$ **cancels**.

(b) Internal auxiliary o (ours)



$\mathbf{x}_{\mathbf{o}} \neq \mathbf{x}_{\mathbf{o}_0} \Longrightarrow$ **need not cancel**.

Obstruction. For $d = \ell_{\hat{g}} - \ell_g$,

$$\Delta_{\mathbf{o}} d = d(\mathbf{x}_{\mathbf{o}}) - d(\mathbf{x}_{\mathbf{o}_0}) \not\equiv 0.$$

Sufficient fix: VP on both maps.

$$|\det \mathbf{J}_g| = |\det \mathbf{J}_{\hat{g}}| = 1 \Longrightarrow \ell_g = \ell_{\hat{g}} = 0$$

Identifiability theorem (Prop. 1)

THEOREM: $O = C$

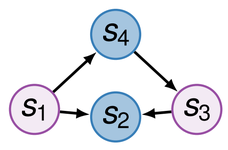
Assume observed-coordinate preservation, conditional block factorization, VP on both maps, full score–Hessian variability, and the paper's alignment/regularity conditions (including a conditioning-independent reparameterization and block matching).

Then the unobserved subspaces are identifiable up to **permutation across subspaces** + **an invertible transform within each** (ISA-style class).

Scope. Square invertible source-to-observation map. The image panels are empirical tests, not theorem instances.

Choosing what to condition on

Conditioning on *more* observed sources can make the target partition **coarser**: conditioning on a collider opens a path.



$O = \{s_2, s_4\}$; targets $\{s_1, s_3\}$.
The observed s_2 is a collider.

Condition on C	Target blocks	
$\{s_2, s_4\}$	$\{s_1, s_3\}$	coarse
$\{s_4\}$	$\{s_1\}, \{s_3\}$	**finest ✓**

Selection rule. Among nonempty $C \subseteq O$, choose the **smallest** set that maximizes the number of d-separated target blocks.

OUTSIDE PROP. 1: $C \subsetneq O$

Here $C = \{s_4\}$ leaves s_2 **observed but unselected**. The graph term below models this case empirically; **Prop. 1 does not cover it**.

Model

VP encoder: GIN, a volume-preserving flow.

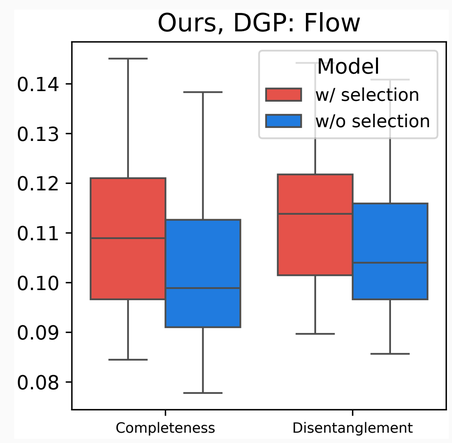
$$\mathcal{L}(\theta) = \mathbb{E}_{\mathcal{D}}\Big[\log p_{\text{align}}(\hat{\mathbf{o}}_C \mid \mathbf{o}_C) + \sum_j \log p(\hat{\mathbf{z}}_{C_j} \mid \mathbf{o}_C) + \sum_{i \in O \setminus C} \log p\Big(o_i \mid \hat{f}_i(\hat{\mathbf{s}}_i^{\text{slot}})\Big)\Big].$$

- **Align** selected source coordinates with $\mathbf{o}_C$.
- **Separate** target blocks conditionally on $\mathbf{o}_C$.
- **Predict** each unselected source from its graph-indexed slots.

Evidence

1. Source selection

Flow DGP. Both median scores improve; the distributions overlap.



2. Less cross-factor leakage

Pendulum MCC. SL/SP: shadow length/position.
Worst off-diagonal ↓:
ours .096 | GIN .23 | iVAE .38.

	Ours		GIN		iVAE	
True SL	0.90	0.096	0.63	0.064	0.19	0.056
True SP	0.082	0.88	0.23	0.99	0.38	0.88
	SL	SP	SL	SP	SL	SP
			Learned			

snu-causality-lab/crl-internal-auxiliaries

arXiv:2509.19058