HIERARCHICAL IDENTIFIABILITY IN MULTILAYER REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Despite the hierarchical structure exhibited by many real-world generative processes, identifiability theory in representation learning has primarily focused on flat generative models, leaving intermediate representations in hierarchical models largely unexplored. We show that hierarchical structure fundamentally changes the identifiability problem: intermediate representations are governed by admissible inter-layer reparameterizations, which must be controlled across successive layers. We study a class of hierarchical stochastic models with additive nonlinear feature maps, overdetermined linear mixing, and layer-wise innovations. Under explicit nondegeneracy and innovation assumptions, we establish hierarchical identifiability from observational data alone: nonlinear ambiguities at intermediate layers reduce to permutation and component-wise affine transformations, while the root representation is identifiable up to permutation and component-wise local diffeomorphisms. We further show that identifiability can propagate even when only a sufficiently informative subset of intermediate coordinates is affinely identifiable. The proposed structural constraints can be incorporated into simple reconstruction-based estimation. Experiments on synthetic hierarchical data support the qualitative predictions of the theory, while experiments on high-dimensional image data suggest that the resulting identifiability-inspired architectural constraints can provide useful inductive biases beyond the idealized theoretical setting.

1 INTRODUCTION

Representation learning seeks latent representations that capture the underlying factors of variation in high-dimensional observations. A fundamental difficulty, however, is *identifiability*: without additional structure, latent variables cannot in general be uniquely recovered from observational data, and multiple incompatible representations may explain the same observations equally well (Hyvärinen & Pajunen, 1999; Locatello et al., 2019). Most existing identifiability theory focuses on *flat* generative models, where observations are produced by a single nonlinear mapping, and establishes recovery under assumptions such as latent independence, auxiliary variables, interventions, or structural constraints on the generator (Hyvärinen et al., 2023; Schölkopf et al., 2021).

Many real-world generative processes, however, are naturally organized across multiple stages or levels of abstraction. In industrial sensing, low-dimensional operating conditions may give rise to richer physical states, which are then reflected in high-dimensional measurements together with layer-specific disturbances. In financial systems, macroeconomic or firm-level factors may influence sector-level states that subsequently generate asset-level observations. Such considerations motivate hierarchical latent-variable models in which different layers represent different stages of the underlying generative process (Bengio et al., 2013; LeCun et al., 2015; Sønderby et al., 2016; Zhao et al., 2017).

Hierarchical structure introduces a qualitatively different identifiability problem. Even when the overall input–output mapping is well constrained, intermediate representations can remain arbitrarily reparameterized. For a two-layer composition, any invertible map $\boldsymbol{g}$ can be inserted as $\boldsymbol{f}^1 \circ \boldsymbol{f}^0 = (\boldsymbol{f}^1 \circ \boldsymbol{g}^{-1}) \circ (\boldsymbol{g} \circ \boldsymbol{f}^0)$, without changing the observations. In deeper models, such inter-layer reparameterization symmetries can propagate across multiple stages. Consequently, hierarchical identifiability cannot be obtained simply by applying single-layer identifiability results

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.

independently at each layer: the admissible transformations must be controlled consistently across the entire hierarchy.

In this work, we develop a theory of *hierarchically identifiable representation learning (HIRL)* based on explicit architectural control of inter-layer reparameterizations. We study a stochastic hierarchy in which each transition has the form $\mathbf{z}^{\ell+1} = \boldsymbol{A}^{\ell}\boldsymbol{r}^{\ell}(\mathbf{z}^{\ell}) + \boldsymbol{B}^{\ell}\epsilon^{\ell}$, where $\boldsymbol{r}^{\ell}$ is additive across latent coordinates, $\boldsymbol{A}^{\ell}$ is an overdetermined linear mixing matrix, and ϵ^{ℓ} represents a layer-wise innovation. The additive nonlinear features provide coordinate-specific nonlinear variation, while overdetermined mixing creates cross-coordinate constraints that restrict admissible inter-layer reparameterizations.

Our main theoretical result shows that, within the stated admissible model class, all intermediate representations are identifiable from observational data alone up to permutation and component-wise affine transformations. The root representation remains identifiable up to permutation and component-wise local diffeomorphisms, reflecting the absence of a lower transition that could further restrict its coordinate-wise reparameterization. The proof proceeds recursively from the observed layer: a null-space argument forces the next-layer transformation to be affine, innovation-subspace identifiability separates inherited state variation from stochastic innovations, and sufficient nonlinearity of the conditional mean restricts the induced lower-layer transformation to be component-wise up to permutation. Repeating this argument backward yields hierarchical identifiability throughout the model. We further show that full affine identifiability of every intermediate coordinate is stronger than necessary: identifiability can still propagate when only a sufficiently informative subset of next-layer coordinates is affinely identifiable, provided that it retains enough information about the entire lower-layer state after the innovation directions are removed.

Our contributions are threefold. First, we establish an identifiability theory for stochastic hierarchical representation learning that explicitly controls inter-layer reparameterization symmetries. Second, we characterize recursive propagation of identifiability, including a partial-identifiability result in which a sufficiently informative subset of affine-identifiable coordinates can be sufficient to recover lower-layer representations. Third, we show that the proposed architectural constraints are compatible with simple reconstruction-based estimation; synthetic experiments support the qualitative predictions of the theory, while image experiments assess their usefulness as an identifiability-inspired inductive bias beyond the idealized theoretical setting.

2 Background

Flat generative models and identifiability. Representation learning typically considers observations $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^{d_{\mathbf{x}}}$ generated from latent variables $\mathbf{z} \in \mathcal{Z} \subset \mathbb{R}^{d_{\mathbf{z}}}$ through a nonlinear mapping,

$$\mathbf{x} = \boldsymbol{f}(\mathbf{z}) + \epsilon, \qquad \mathbf{z} \sim p_{\mathbf{z}}, \qquad \epsilon \sim p_{\epsilon}, \tag{1}$$

where ϵ denotes observation noise independent of $\mathbf{z}$ (Fig. 1a). Existing identifiability results recover latent variables up to appropriate equivalence classes by imposing assumptions such as independence, auxiliary variables, temporal or causal structure, supervision, interventions, or other restrictions on the latent distribution (Ahuja et al., 2022a;b; Brehmer et al., 2022; Buchholz et al., 2023; Hyvärinen & Morioka, 2016; Hyvarinen & Morioka, 2017; Hyvarinen et al., 2019; Khemakhem et al., 2020; Schölkopf et al., 2021; Shen et al., 2022; Squires et al., 2023; Varici et al., 2023; Yang et al., 2021).

Generator-side constraints. An alternative line of work derives identifiability by restricting the generator $\boldsymbol{f}$ rather than, or in addition to, the latent variables. Examples include orthogonality (Buchholz et al., 2022; Buchholz & Schölkopf, 2024; Gresele et al., 2021; Horan et al., 2021), sparsity (Lachapelle et al., 2023a; Leemann et al., 2023; Moran et al., 2022; Zheng et al., 2022), grouping mechanisms (Daunhawer et al., 2023; Lyu et al., 2022; Morioka & Hyvarinen, 2023; 2024; Sturma et al., 2023; Yao et al., 2023), and piecewise-affine structures (Kivva et al., 2022). Recent works further study restricted nonlinear interactions and compositional generators (Brady et al., 2023; 2025; Lachapelle et al., 2023b). These results primarily concern single-layer representations and do not by themselves resolve reparameterization ambiguities between multiple latent layers.

Hierarchical generative models. Hierarchical latent-variable models represent observations $\mathbf{x} = \mathbf{z}^{L}$ through a cascade of latent states,

$$\mathbf{z}^{\ell+1} = \boldsymbol{f}^{\ell}(\mathbf{z}^{\ell}, \epsilon^{\ell}), \qquad \ell = 0, \dots, L-1,$$

where ϵ^ℓ is innovation, representing layer-wise stochastic variation not inherited from the preceding state (Fig. 1b) (Dinh et al., 2017; Kingma & Dhariwal, 2018; Papamakarios et al., 2021; Ranganath et al., 2016; Sønderby et al., 2016; Vahdat & Kautz, 2020). Such architectures are widely used to represent multiple stages or levels of abstraction (Bengio et al., 2013; LeCun et al., 2015; Sønderby et al., 2016; Zhao et al., 2017). However, intermediate states admit additional inter-layer reparameterizations that leave the observed distribution unchanged. Consequently, assumptions sufficient for identifying a flat representation do not automatically identify intermediate representations in a hierarchy. Our analysis focuses on architectural conditions that restrict these inter-layer ambiguities.

Notation. Scalars are denoted by lowercase letters (e.g., z), vectors by bold lowercase letters (e.g., $\mathbf{z}$), and matrices by bold uppercase letters (e.g., $\boldsymbol{A}$). For a positive integer d, we write $[d] := \{1, \dots, d\}$. For $\mathcal{S} \subset [d]$ and $\boldsymbol{A} \in \mathbb{R}^{d \times n}$, $\boldsymbol{A}_{\mathcal{S}}$ denotes the submatrix consisting of the rows indexed by $\mathcal{S}$. For a vector-valued function $\boldsymbol{\sigma} : \mathbb{R}^d \to \mathbb{R}^m$, we denote its Jacobian by $\mathbf{J}_{\boldsymbol{\sigma}}(\mathbf{z}) := \left(\frac{\partial \sigma_i(\mathbf{z})}{\partial z_j}\right)_{i,j}$, and its matrix of coordinate-wise second derivatives by $\mathbf{M}_{\boldsymbol{\sigma}}(\mathbf{z}) := \left(\frac{\partial^2 \sigma_i(\mathbf{z})}{\partial z_j^2}\right)_{i,j}$. Throughout the paper, superscripts denote layer indices and subscripts denote coordinate indices.

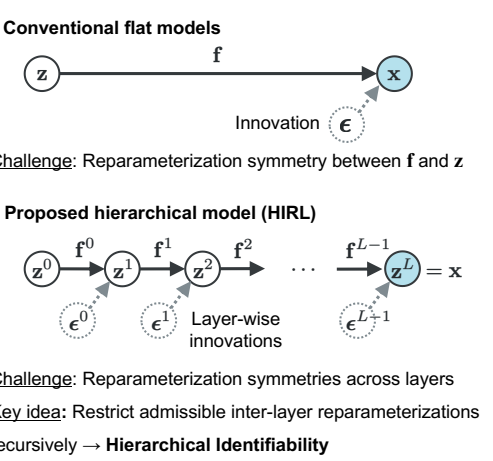


Figure 1: Identifiability in flat and hierarchical generative models. (**a**) In flat models, observations are generated from a single latent layer, and identifiability is governed by assumptions on the latent variables or the generator. (**b**) Hierarchical models contain multiple latent layers and admit additional inter-layer reparameterization symmetries. HIRL restricts the admissible inter-layer reparameterizations through structural conditions on the layer-wise transitions, leading to hierarchical identifiability.

3 Model definition

In this section, we define the hierarchical stochastic model studied in this work and the corresponding learned hierarchical autoencoder model used in the identifiability analysis. The assumptions stated here are the main population-level conditions used in the identifiability theorem in Section 4.

3.1 Stochastic hierarchy with innovations

We formulate the data-generating process as a stochastic latent cascade. Specifically, we consider a hierarchy of latent variables $\mathbf{z}^0, \mathbf{z}^1, \dots, \mathbf{z}^L = \mathbf{x}$ whose inter-layer transitions consist of a nonlinear conditional mean and an additive stochastic innovation.

Assumption 1 (Invertible additive transitions with innovations)**.** *For each layer $\ell = 0, \dots, L-1$, the true transition is given by the following form, with an additive nonlinear feature map:*

$$\mathbf{z}^{\ell+1} = \boldsymbol{f}^\ell(\mathbf{z}^\ell, \epsilon^\ell) = \boldsymbol{\mu}^\ell(\mathbf{z}^\ell) + \boldsymbol{B}^\ell \epsilon^\ell = \boldsymbol{A}^\ell \boldsymbol{r}^\ell(\mathbf{z}^\ell) + \boldsymbol{B}^\ell \epsilon^\ell \tag{2}$$

where $\boldsymbol{r}^\ell : \mathbb{R}^{d_{\mathbf{z}}^\ell} \to \mathbb{R}^{d_{\mathbf{r}}^\ell}$ is a C^2 additive feature map of the form $\boldsymbol{r}^\ell(\mathbf{z}^\ell) = \sum_{i=1}^{d_{\mathbf{z}}^\ell} \boldsymbol{r}_i^\ell(z_i^\ell)$, $\boldsymbol{r}_i^\ell : \mathbb{R} \to \mathbb{R}^{d_{\mathbf{r}}^\ell}$, $\mathbf{z}^\ell \in \mathbb{R}^{d_{\mathbf{z}}^\ell}$, $\epsilon^\ell \in \mathbb{R}^{d_\epsilon^\ell}$, $\boldsymbol{A}^\ell \in \mathbb{R}^{d_{\mathbf{z}}^{\ell+1} \times d_{\mathbf{r}}^\ell}$ and $\boldsymbol{B}^\ell \in \mathbb{R}^{d_{\mathbf{z}}^{\ell+1} \times d_\epsilon^\ell}$ are full column rank matrices. We further assume that each transition is locally invertible with respect to $(\mathbf{z}^\ell, \epsilon^\ell)$; namely, $d_{\mathbf{z}}^{\ell+1} = d_{\mathbf{z}}^\ell + d_\epsilon^\ell$, and $\operatorname{rank} \mathbf{J}_{\boldsymbol{f}^\ell}(\mathbf{z}^\ell, \epsilon^\ell) = d_{\mathbf{z}}^{\ell+1}$ on the relevant support.

Unlike standard flat-decoder formulations, we treat the intermediate variables $\mathbf{z}^\ell$ as latent mediating states generated by successive structural transitions. Each transition consists of variation inherited through the conditional mean and a layer-wise innovation. Although related in spirit to nonlinear additive-noise models (Hoyer et al., 2008), our goal is not causal direction discovery, but identifiability of the intermediate states and the conditional-mean cascade from observations of the final layer. We use *parent* and *child* to refer to adjacent lower and upper layers, respectively.

The restriction on interactions among variables is a key ingredient of the identifiability argument, similarly to existing decoder-based identifiable representation learning frameworks (Brady et al.,

2023; 2025; Lachapelle et al., 2023b). The additive form in Assumption 1 corresponds to the additive model of (Lachapelle et al., 2023b) or at-most-first-order interactions in the terminology of (Brady et al., 2025). Although each layer has this additive form, the hierarchical composition can represent much richer nonlinear functions from $\mathbf{z}^0$ to $\mathbf{x} = \mathbf{z}^L$.

We next impose regularity conditions on the conditional-mean structure.

Assumption 2 (Sufficient nonlinearity)**.** *For every $\ell = 0, \dots, L-1$, the function $\boldsymbol{r}^\ell$ is sufficiently nonlinear in the sense that the $d_{\mathbf{r}}^\ell \times 2d_{\mathbf{z}}^\ell$ matrix*

$$\begin{bmatrix} \mathbf{M}_{\boldsymbol{r}^\ell}(\mathbf{z}^\ell) & \mathbf{J}_{\boldsymbol{r}^\ell}(\mathbf{z}^\ell) \end{bmatrix} \tag{3}$$

has full column rank $2d_{\mathbf{z}}^\ell$ almost everywhere on the support of $\mathbf{z}^\ell$.

Assumption 3 (Overdetermined mixing)**.** *For every $\ell = 0, \dots, L-1$, the conditional-mean mixing matrix has a nontrivial left null space:*

$$\operatorname{rank}(\boldsymbol{A}^\ell) = d_{\mathbf{r}}^\ell < d_{\mathbf{z}}^{\ell+1}.$$

Assumption 2 corresponds to the singleton-block case of the sufficient-nonlinearity condition studied in Lachapelle et al. (2023b). Here, it is used to restrict lower-layer reparameterizations to component-wise transformations. Since $\boldsymbol{\mu}^\ell = \boldsymbol{A}^\ell \boldsymbol{r}^\ell$ and $\boldsymbol{A}^\ell$ has full column rank, it also implies $\operatorname{rank}\begin{bmatrix} \mathbf{M}_{\boldsymbol{\mu}^\ell}(\mathbf{z}^\ell) & \mathbf{J}_{\boldsymbol{\mu}^\ell}(\mathbf{z}^\ell) \end{bmatrix} = 2d_{\mathbf{z}}^\ell$. Assumption 3 provides a nontrivial left null space of the conditional-mean mixing, enabling the constraints on inter-layer reparameterizations. Together, these assumptions imply $2d_{\mathbf{z}}^\ell \le d_{\mathbf{r}}^\ell < d_{\mathbf{z}}^{\ell+1}$, hence $d_{\mathbf{z}}^{\ell+1} > 2d_{\mathbf{z}}^\ell$. Such overcomplete output structure also appears in identifiable decoder-based frameworks (Brady et al., 2025; Lachapelle et al., 2023b).

Assumption 4 (Gaussian innovations and innovation-subspace identifiability)**.** *For every layer $\ell = 0, \dots, L-1$, $\epsilon^\ell \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}^\ell)$, $\boldsymbol{\Sigma}^\ell \succ 0$, and $\epsilon^\ell \perp\!\!\!\perp \mathbf{z}^\ell$. Moreover, we assume that the innovation subspace coincides exactly with the set of Gaussian-flat directions of the next-layer distribution: $\mathcal{G}(p_{\mathbf{z}^{\ell+1}}) = \operatorname{Im}(\boldsymbol{B}^\ell)$, where, for a positive C^3 density p on $\mathbb{R}^m$,*

$$\mathcal{G}(p) := \left\{ \boldsymbol{v} \in \mathbb{R}^m : D^3 \log p(\boldsymbol{y})[\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{v}] = 0 \quad \textit{for all } \boldsymbol{u} \in \mathbb{R}^m \textit{ and all relevant } \boldsymbol{y} \right\},$$

and $D^3 f(\boldsymbol{y})[\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}] := \sum_{i,j,k=1}^m \frac{\partial^3 f(\boldsymbol{y})}{\partial y_i \partial y_j \partial y_k} u_i v_j w_k$.

The innovation ϵ^ℓ represents stochastic variation that is not inherited from the current state $\mathbf{z}^\ell$. We assume independence only between ϵ^ℓ and $\mathbf{z}^\ell$, and make no independence assumption among the coordinates of $\mathbf{z}^\ell$. The condition $\mathcal{G}(p_{\mathbf{z}^{\ell+1}}) = \operatorname{Im}(\boldsymbol{B}^\ell)$ requires the innovation subspace to coincide exactly with the Gaussian-flat directions of the next-layer distribution, ruling out additional such directions in the inherited variation. This can be viewed as a nondegeneracy condition excluding accidental Gaussian-flat structure outside the innovation subspace, rather than requiring a specially constructed innovation geometry. The Gaussian-flat condition provides one sufficient formulation for separating innovation directions from inherited variation.

3.2 Learned stochastic hierarchy

Our analysis considers a hierarchical autoencoder with an unconstrained encoder and structured decoder trained to reconstruct $\mathbf{x}$. Given $\mathbf{x} = \mathbf{z}^L$, the encoder $\hat{\boldsymbol{g}}_\phi : \mathbb{R}^{d_{\mathbf{x}}} \to \mathbb{R}^{d_{\mathbf{z}}^0 + \sum_{\ell=0}^{L-1} d_\epsilon^\ell}$ outputs the bottom-state estimate and layer-wise innovation codes, while the decoder recursively constructs the remaining latent states $\hat{\mathbf{z}}^1, \dots, \hat{\mathbf{z}}^{L-1}$ and the reconstruction $\hat{\mathbf{z}}^L$:

$$\left(\hat{\mathbf{z}}^0, \hat{\epsilon}^0, \dots, \hat{\epsilon}^{L-1}\right) = \hat{\boldsymbol{g}}_\phi(\mathbf{x}), \qquad \hat{\mathbf{x}} = \hat{\boldsymbol{f}}_\theta^{0:L-1}\left(\hat{\mathbf{z}}^0, \hat{\epsilon}^0, \dots, \hat{\epsilon}^{L-1}\right) = \hat{\mathbf{z}}^L. \tag{4}$$

The theorem applies to exact-reconstruction solutions, as in Lachapelle et al. (2023b), within the admissible model class below. As in prior decoder-identifiability analyses, exact reconstruction serves as a population-level anchoring condition rather than a finite-sample training requirement.

Assumption 5 (Learned stochastic hierarchy)**.** *The learned model has the same hierarchical form as the true model:*

$$\hat{\mathbf{z}}^{\ell+1} = \hat{\boldsymbol{f}}_\theta^\ell(\hat{\mathbf{z}}^\ell, \hat{\epsilon}^\ell) = \hat{\boldsymbol{\mu}}^\ell(\hat{\mathbf{z}}^\ell) + \hat{\boldsymbol{B}}^\ell \hat{\epsilon}^\ell = \hat{\boldsymbol{A}}^\ell \hat{\boldsymbol{r}}^\ell(\hat{\mathbf{z}}^\ell) + \hat{\boldsymbol{B}}^\ell \hat{\epsilon}^\ell \tag{5}$$

where $\hat{\boldsymbol{r}}^\ell : \mathbb{R}^{d_{\mathbf{z}}^\ell} \to \mathbb{R}^{d_{\mathbf{r}}^\ell}$ is a C^2 additive feature map of the form $\hat{\boldsymbol{r}}^\ell(\hat{\mathbf{z}}^\ell) = \sum_{i=1}^{d_{\mathbf{z}}^\ell} \hat{\boldsymbol{r}}_i^\ell(\hat{z}_i^\ell)$, $\hat{\boldsymbol{r}}_i^\ell : \mathbb{R} \to \mathbb{R}^{d_{\mathbf{r}}^\ell}$, $\hat{\mathbf{z}}^\ell \in \mathbb{R}^{d_{\mathbf{z}}^\ell}$, $\hat{\epsilon}^\ell \in \mathbb{R}^{d_\epsilon^\ell}$, $\hat{\boldsymbol{A}}^\ell \in \mathbb{R}^{d_{\mathbf{z}}^{\ell+1} \times d_{\mathbf{r}}^\ell}$ is full column rank and overdetermined

as in Assumption 3, $\hat{\boldsymbol{B}}^\ell \in \mathbb{R}^{d_{\mathbf{z}}^{\ell+1} \times d_\epsilon^\ell}$ is full column rank. The learned transitions are also assumed to satisfy the same local invertibility condition as in Assumption 1. The learned hierarchy is further restricted to satisfy the corresponding innovation assumptions (Assumption 4), with the hatted quantities replacing the true ones.

Assumption 6 (Conditional centering of learned innovations)**.** *For every layer ℓ,*

$$\mathbb{E}[\hat{\epsilon}^\ell \mid \mathbf{z}^\ell] = \mathbf{0} \qquad \text{a.s.} \tag{6}$$

This is a conditional first-moment restriction and does not require independence between the learned innovation and the true inherited state, nor does it constrain higher-order conditional moments. Its role is to eliminate the learned innovation after conditioning on the true state, yielding the conditional-mean relation used in the subsequent null-space argument. It does not require access to $\mathbf{z}^\ell$ during training; rather, it restricts the admissible population-level decompositions by ruling out systematic state-dependent variation in the learned innovation.

We also impose admissibility conditions on the learned conditional-mean decoder.

Assumption 7 (Null-space nondegeneracy condition)**.** *For each layer $\ell = 0, \ldots, L-1$, define*

$$\mathbf{W}_{\boldsymbol{\mu}^\ell}(\mathbf{z}^\ell) := \left[\frac{\partial \boldsymbol{\mu}^\ell(\mathbf{z}^\ell)}{\partial z_i^\ell} \odot \frac{\partial \boldsymbol{\mu}^\ell(\mathbf{z}^\ell)}{\partial z_j^\ell} \right]_{1 \le i < j \le d_{\mathbf{z}}^\ell}.$$

Let $\hat{\boldsymbol{Q}}_{\boldsymbol{A}}^\ell \in \mathbb{R}^{d_{\mathbf{z}}^{\ell+1} \times (d_{\mathbf{z}}^{\ell+1} - d_{\mathbf{r}}^\ell)}$ be a basis matrix for the left null space of $\hat{\boldsymbol{A}}^\ell$, so that $(\hat{\boldsymbol{Q}}_{\boldsymbol{A}}^\ell)^\top \hat{\boldsymbol{A}}^\ell = \mathbf{0}$. We assume that

$$\operatorname{rank}\left[(\hat{\boldsymbol{Q}}_{\boldsymbol{A}}^\ell)^\top \odot_{\mathrm{KR}} \mathbf{W}_{\boldsymbol{\mu}^\ell}(\mathbf{z}^\ell)^\top \right] = d_{\mathbf{z}}^{\ell+1} \tag{7}$$

on the relevant support, where $\odot_{\mathrm{KR}}$ denotes the Khatri–Rao product, i.e., the column-wise Kronecker product.

The Khatri–Rao structure arises by stacking the mixed-derivative constraints over coordinate pairs and null-space directions. Assumption 7 requires the resulting cross-coordinate constraints to remain sufficiently rich. After conditional averaging removes the learned innovation, the mixed derivatives yield a homogeneous linear system for the second derivatives of the next-layer reparameterization. Full column rank then forces all these second derivatives to vanish. This is a sufficient rather than necessary condition; Section 4.2 gives admissible constructions, and a closely related rank condition was used in Lyu & Fu (2022).

Further discussion of the roles and interpretation of the assumptions used in the hierarchical identifiability result, including which aspects are structural and which are primarily technical or proof-specific, is provided in Appendix A.

4 Hierarchical identifiability

We now present the main identifiability result. An exact-reconstruction solution only anchors the observation layer, $\hat{\mathbf{z}}^L = \mathbf{z}^L = \mathbf{x}$. It does not by itself identify intermediate representations. The theorem below shows that, within the admissible hierarchical model class, this top-layer correspondence propagates recursively through the hierarchy, yielding identifiable intermediate representations up to well-characterized ambiguities. Importantly, the result does not require independence among the coordinates within each latent state. The proof is given in Appendix B.

Theorem 1 (Hierarchical identifiability)**.** *Consider the true and learned hierarchies defined in equation 2 and equation 5. Suppose that Assumptions 1–7 and the deconvolution regularity condition stated in Assumption 8 (Appendix A) hold at every layer and that the observation layer is anchored, $\hat{\mathbf{z}}^L = \mathbf{z}^L = \mathbf{x}$. Then, on any connected neighborhood on which the stated regularity conditions hold, each intermediate representation satisfies*

$$\hat{\mathbf{z}}^\ell = \mathbf{P}^\ell \mathbf{D}^\ell \mathbf{z}^\ell + \boldsymbol{c}^\ell, \qquad \ell = 1, \ldots, L-1,$$

where $\mathbf{P}^\ell$ is a permutation matrix, $\mathbf{D}^\ell$ is a nonsingular diagonal matrix, and $\mathbf{c}^\ell$ is a constant vector. At the root layer,

$$\hat{\mathbf{z}}^0 = \mathbf{P}^0 \boldsymbol{k}^0(\mathbf{z}^0),$$

where $\boldsymbol{k}^0$ is component-wise and each component is a local diffeomorphism. Furthermore, the layer-wise innovations are identifiable up to invertible linear transformations,

$$\hat{\epsilon}^\ell = \mathbf{R}^\ell \epsilon^\ell, \qquad \ell = 0, \dots, L-1,$$

for some invertible $\mathbf{R}^\ell$.

The theorem establishes hierarchical identifiability recursively from the observed layer toward the root. Importantly, the correspondence between the true and learned latent states is not assumed a priori, but is derived from the structural and stochastic conditions. Intermediate layers are identified up to affine monomial transformations, whereas the root retains a component-wise nonlinear ambiguity because no lower transition is available to further constrain it. It also shows that the layer-wise innovations are identifiable up to invertible linear transformations.

Proof overview. The proof proceeds by backward induction. Suppose inductively that the learned and true representations at layer $\ell+1$ are related by a component-wise transformation up to permutation. Conditional centering (Assumption 6) removes the learned innovation after conditioning on the true current-layer state, yielding a deterministic null-space relation. The null-space nondegeneracy condition (Assumption 7) then forces the smoothed next-layer transformation to be affine, and Gaussian deconvolution transfers this conclusion to the original transformation. Innovation-subspace identifiability (Assumption 4) aligns the true and learned innovation subspaces under the next-layer transformation, allowing the innovation directions to be projected out. The resulting projected conditional means provide local coordinates for the current-layer state. Finally, additivity and sufficient nonlinearity (Assumptions 1, 2, and 5) restrict the induced current-layer correspondence to a component-wise transformation up to permutation. For an intermediate layer, the transition from the layer below provides another instance of the null-space argument, which further reduces this component-wise transformation to an affine monomial transformation.

Additional dimensional requirement. Beyond the expansion condition discussed in Section 3, null-space nondegeneracy requires sufficiently many mixed-derivative constraints. Since $(\hat{\boldsymbol{Q}}^\ell_{\boldsymbol{A}})^\top$ has $d^{\ell+1}_{\mathbf{z}} - d^\ell_{\mathbf{r}}$ rows and $\mathbf{W}_{\boldsymbol{\mu}^\ell}$ contains $\binom{d^\ell_{\mathbf{z}}}{2}$ mixed-coordinate directions, a necessary dimensional condition is $\left(d^{\ell+1}_{\mathbf{z}} - d^\ell_{\mathbf{r}}\right)\binom{d^\ell_{\mathbf{z}}}{2} \geq d^{\ell+1}_{\mathbf{z}}$. Thus, the feasibility of null-space nondegeneracy depends not only on layer width but also on the interaction between the codimension of the mean subspace and the number of available cross-coordinate derivative constraints.

4.1 Partial identifiability

A key step in hierarchical identifiability is the propagation of next-layer affine identifiability to the current-layer representation. Full affine identifiability of the next layer, however, is stronger than necessary: a sufficiently informative affine-identifiable subset can already determine the entire current-layer state.

Theorem 2 (Propagation of partial affine identifiability)**.** *Consider a single transition at layer ℓ. Let $\mathcal{S} \subseteq [d^{\ell+1}_{\mathbf{z}}]$ denote a subset of next-layer coordinates, and write*

$$\mathbf{z}^{\ell+1}_{\mathcal{S}} = \boldsymbol{\mu}^\ell_{\mathcal{S}}(\mathbf{z}^\ell) + \boldsymbol{B}^\ell_{\mathcal{S}} \epsilon^\ell, \qquad \hat{\mathbf{z}}^{\ell+1}_{\mathcal{S}} = \hat{\boldsymbol{\mu}}^\ell_{\mathcal{S}}(\hat{\mathbf{z}}^\ell) + \hat{\boldsymbol{B}}^\ell_{\mathcal{S}} \hat{\epsilon}^\ell.$$

Suppose that this subset is affinely identifiable in the inductive step of Theorem 1, i.e.,

$$\hat{\mathbf{z}}^{\ell+1}_{\mathcal{S}} = \boldsymbol{H}_{\mathcal{S}} \mathbf{z}^{\ell+1}_{\mathcal{S}} + \boldsymbol{c}_{\mathcal{S}},$$

where $\boldsymbol{H}_{\mathcal{S}} \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{S}|}$ is invertible. Assume the conditional-centering condition for the learned innovation $\mathbb{E}[\hat{\epsilon}^\ell \mid \mathbf{z}^\ell] = \mathbf{0}$ a.s., and suppose that the restricted distributions satisfy

$$\mathcal{G}(p_{\mathbf{z}^{\ell+1}_{\mathcal{S}}}) = \mathrm{Im}(\boldsymbol{B}^\ell_{\mathcal{S}}), \qquad \mathcal{G}(p_{\hat{\mathbf{z}}^{\ell+1}_{\mathcal{S}}}) = \mathrm{Im}(\hat{\boldsymbol{B}}^\ell_{\mathcal{S}}).$$

Suppose further that the restricted mean maps satisfy the restricted analogue of Assumption 2:

$$\mathrm{rank}\Big[\mathbf{M}_{\boldsymbol{\mu}^\ell_{\mathcal{S}}}(\mathbf{z}^\ell) \quad \mathbf{J}_{\boldsymbol{\mu}^\ell_{\mathcal{S}}}(\mathbf{z}^\ell)\Big] = 2d^\ell_{\mathbf{z}}.$$

Let $\hat{\boldsymbol{Q}}_{\boldsymbol{B},\mathcal{S}} \in \mathbb{R}^{|\mathcal{S}| \times (|\mathcal{S}| - \mathrm{rank}(\hat{\boldsymbol{B}}^\ell_{\mathcal{S}}))}$ be a basis matrix for the left null space of $\hat{\boldsymbol{B}}^\ell_{\mathcal{S}}$. Suppose that

$$\mathrm{rank}\Big[\hat{\mathbf{Q}}^\top_{\boldsymbol{B},\mathcal{S}} \boldsymbol{H}_{\mathcal{S}} \mathbf{J}_{\boldsymbol{\mu}^\ell_{\mathcal{S}}}(\mathbf{z}^\ell)\Big] = \mathrm{rank}\Big[\hat{\mathbf{Q}}^\top_{\boldsymbol{B},\mathcal{S}} \mathbf{J}_{\hat{\boldsymbol{\mu}}^\ell_{\mathcal{S}}}(\hat{\mathbf{z}}^\ell)\Big] = d^\ell_{\mathbf{z}}.$$

Then the entire current-layer state is locally identifiable: around every regular point, there exists a local diffeomorphism $\boldsymbol{k}$ such that $\hat{\mathbf{z}}^\ell = \boldsymbol{k}(\mathbf{z}^\ell)$, where $\boldsymbol{k}$ is component-wise up to permutation.

Table 1: Comparison of representation recovery across HIRL and hierarchical baselines. Values are mean ± standard deviation over random seeds. Higher is better.

	Latent recovery ↑			Innovation alignment ↑		
Method	$\mathbf{z}^0$	$\mathbf{z}^1$	$\mathbf{z}^2$	ϵ^0	ϵ^1	ϵ^2
HIRL (ours)	$\mathbf{0.94 \pm 0.03}$	$\mathbf{0.95 \pm 0.03}$	$\mathbf{0.98 \pm 0.01}$	$\mathbf{0.97 \pm 0.01}$	$\mathbf{0.88 \pm 0.01}$	$\mathbf{0.88 \pm 0.00}$
HIRL (reduced feat.)	0.72 ± 0.06	0.83 ± 0.03	0.91 ± 0.02	0.93 ± 0.02	0.86 ± 0.02	0.87 ± 0.00
LVAE	0.46 ± 0.03	0.12 ± 0.02	0.08 ± 0.01	–	–	–
VLAE	0.28 ± 0.02	0.22 ± 0.01	0.15 ± 0.01	–	–	–

The proof is given in Appendix C. The assumptions in Theorem 2 are mostly restricted analogues of those in Theorem 1, and ensure that the selected subset retains enough information about the entire current-layer state after the innovation directions are projected out. The restricted sufficient-nonlinearity condition restricts nonlinear mixing among the current-layer coordinates, while the projected-rank conditions ensure that the projected conditional mean remains locally informative about all current-layer variables. Thus, the subset need not contain all next-layer coordinates, but it must retain both sufficient nonlinear variation and sufficient rank after innovation removal.

In particular, the projected-rank condition requires $|\mathcal{S}| - \operatorname{rank}(\hat{\boldsymbol{B}}^\ell_{\mathcal{S}}) \geq d^\ell_{\mathbf{z}}$. For an intermediate layer, under the corresponding assumptions on the preceding lower transition, the remaining component-wise ambiguity can be further reduced to an affine monomial transformation. Thus, failures of affine identifiability outside an informative subset need not interrupt identifiability propagation through the hierarchy.

4.2 Examples of admissible and degenerate models

The two main nondegeneracy conditions constrain different aspects of the layer-wise transition. A simple construction takes $\boldsymbol{\mu}^\ell(\mathbf{z}^\ell) = \boldsymbol{A}^\ell \boldsymbol{r}^\ell(\mathbf{z}^\ell)$ with $d^\ell_{\mathbf{r}} = 2d^\ell_{\mathbf{z}}$, and assigns two nonlinear features to each latent coordinate:

$$\boldsymbol{r}^\ell(\mathbf{z}^\ell) = \left[r^\ell_{11}(z^\ell_1), r^\ell_{12}(z^\ell_1), \ldots, r^\ell_{d^\ell_{\mathbf{z}}1}(z^\ell_{d^\ell_{\mathbf{z}}}), r^\ell_{d^\ell_{\mathbf{z}}2}(z^\ell_{d^\ell_{\mathbf{z}}})\right]^\top,$$

where $\boldsymbol{A}^\ell$ is a generic dense mixing matrix. If, for each i, the two-dimensional vectors formed by the first and second derivatives of r^ℓ_{i1} and r^ℓ_{i2} are linearly independent, then this construction can satisfy equation 3 with rank $2d^\ell_{\mathbf{z}}$. At the same time, generic dense mixing can cause distinct latent coordinates to jointly influence common output coordinates, which can yield $\partial_i \boldsymbol{\mu}^\ell(\mathbf{z}^\ell) \odot \partial_j \boldsymbol{\mu}^\ell(\mathbf{z}^\ell) \neq \mathbf{0}$, $i \neq j$. Together with the dimensional requirements, sufficiently rich cross-coordinate variations of this type can satisfy the null-space nondegeneracy condition (Assumption 7).

The two conditions can fail for different reasons. In the simplest one-feature-per-coordinate construction, $d^\ell_{\mathbf{r}} = d^\ell_{\mathbf{z}}$. Consequently, $\operatorname{rank}\left[\mathbf{M}_{\boldsymbol{r}^\ell}(\mathbf{z}^\ell) \quad \mathbf{J}_{\boldsymbol{r}^\ell}(\mathbf{z}^\ell)\right] \leq d^\ell_{\mathbf{z}} < 2d^\ell_{\mathbf{z}}$, so sufficient nonlinearity is dimensionally impossible to satisfy, even when the subsequent mixing is dense. Conversely, even with sufficiently rich nonlinear features, null-space nondegeneracy can fail when $\boldsymbol{A}^\ell$ is too sparse or structured. For example, if each output coordinate of $\boldsymbol{\mu}^\ell$ depends on only one latent coordinate, then $\partial_i \boldsymbol{\mu}^\ell(\mathbf{z}^\ell) \odot \partial_j \boldsymbol{\mu}^\ell(\mathbf{z}^\ell) = \mathbf{0}$, $i \neq j$, so $\mathbf{W}_{\boldsymbol{\mu}^\ell}(\mathbf{z}^\ell) = \mathbf{0}$ and the mixed-derivative null-space argument cannot constrain the next-layer nonlinear reparameterization.

Hence, sufficient nonlinearity and null-space nondegeneracy are complementary rather than redundant. The former requires sufficiently rich within-coordinate nonlinear variation to restrict current-layer reparameterizations to component-wise form, whereas the latter requires sufficiently rich cross-coordinate interactions after mixing to eliminate nonlinear ambiguities in the next-layer representation.

5 Experiments

We evaluate HIRL on synthetic hierarchies to test layer-wise latent recovery, controlled violations of the theoretical conditions, partial identifiability, and robustness to model misspecification. We compare HIRL with a reduced-feature ablation, LVAE (Sønderby et al., 2016), and VLAE (Zhao et al.,

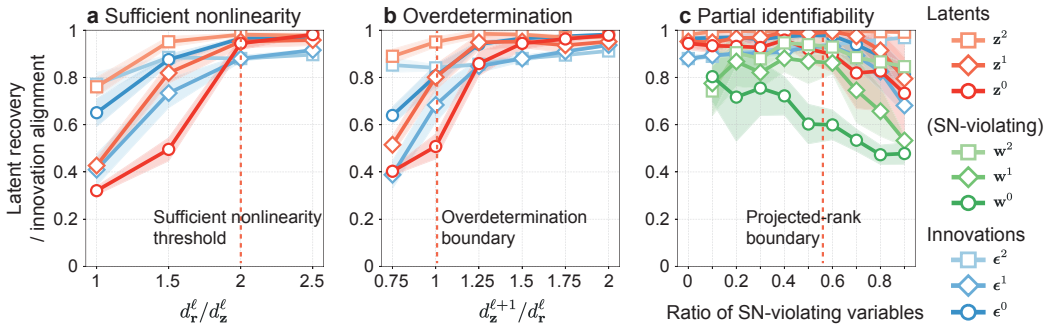


Figure 2: Empirical validation of the identifiability conditions. (**a**) Sufficient nonlinearity: recovery improves as $d_{\mathbf{r}}^{\ell}/d_{\mathbf{z}}^{\ell}$ increases; the vertical boundary at $d_{\mathbf{r}}^{\ell}/d_{\mathbf{z}}^{\ell} = 2$ marks the dimensional requirement implied by the sufficient-nonlinearity condition. (**b**) Overdetermination: recovery changes across the dimensional boundary $d_{\mathbf{z}}^{\ell+1} > d_{\mathbf{r}}^{\ell}$, which is required by the overdetermination condition. (**c**) Partial identifiability: increasing the fraction of variables for which sufficient nonlinearity is intentionally violated primarily degrades recovery of those variables, while variables retaining the sufficient-nonlinearity construction remain comparatively well recovered. The vertical dashed line marks the dimensional boundary implied by the projected-rank condition in Theorem 2. Red curves denote latents retaining the sufficient-nonlinearity construction, green curves denote latents for which this condition is intentionally violated, and blue curves denote innovations. Shaded regions indicate standard deviations over random seeds. Here, SN denotes sufficient nonlinearity.

2017). VLAE and LVAE are chosen as representative hierarchical representation-learning baselines that are amenable to our controlled fully connected setting, whereas more recent methods rely on substantially different convolutional architectures (Vahdat & Kautz, 2020) or temporal structure (Li et al., 2025), and are therefore not included in this controlled comparison.

Latent recovery is measured by mean absolute correlation after optimal permutation matching, while innovations are evaluated after optimal linear alignment to account for their linear indeterminacy (Theorem 1). The experiments do not enforce all assumptions of the theorem and therefore test its qualitative structural predictions rather than the full identifiability guarantee. Complete experimental details are provided in Appendix D.

Synthetic hierarchical data. We first consider synthetic data generated according to the proposed hierarchical additive model, with sufficient nonlinearity and overdetermination satisfied in the default setting. As detailed in Appendix D, the synthetic experiments use a reduced innovation dimension and therefore relax the exact dimensional condition required for transition-level local invertibility.

Table 1 reports layer-wise recovery of the latent variables and innovations. HIRL achieves consistently high recovery across all layers and outperforms the reduced-feature ablation as well as LVAE and VLAE. Notably, strong recovery is still observed despite the reduced innovation dimension, although this setting lies outside the exact transition-level dimensional condition used in the theory. The reduced-feature ablation performs consistently worse than HIRL, particularly at lower layers, while LVAE and VLAE show substantially weaker layer-wise recovery in this setting. Appendix Fig. 3a further examines the effect of sample size on recovery.

Effect of violating theoretical assumptions. We next selectively violate sufficient nonlinearity (Assumption 2) and overdetermination (Assumption 3). Violating either condition substantially degrades recovery, with larger effects at lower layers (Figs. 2a,b). Recovery changes markedly around the corresponding dimensional boundaries, consistent with the qualitative role of both conditions in the theory.

Partial identifiability. We next examine partial identifiability by violating sufficient nonlinearity (Assumption 2) for an increasing fraction of variables while retaining a subset that preserves the sufficient-nonlinearity construction. As this fraction increases, recovery degrades primarily for those variables, whereas variables retaining the sufficient-nonlinearity construction remain comparatively

Table 2: Linear-probe accuracy (%) for CIFAR-100 classification and test reconstruction MSE. Best accuracy at each latent layer and lowest reconstruction error are shown in bold.

Method	$\boldsymbol{z}^0 \uparrow$	$\boldsymbol{z}^1 \uparrow$	$\boldsymbol{z}^2 \uparrow$	$\boldsymbol{z}^3 \uparrow$	Test MSE $\downarrow$
HIRL (ours)	$\mathbf{14.2 \pm 0.2}$	$\mathbf{14.6 \pm 0.3}$	$\mathbf{17.7 \pm 0.3}$	$\mathbf{20.4 \pm 0.3}$	0.267 ± 0.001
HIRL (reduced feat.)	14.1 ± 0.2	14.1 ± 0.2	16.6 ± 0.1	19.2 ± 0.1	0.259 ± 0.001
LVAE	13.6 ± 0.4	9.0 ± 0.4	15.5 ± 0.8	17.3 ± 0.4	$\mathbf{0.043 \pm 0.000}$
VLAE	3.7 ± 0.9	2.8 ± 2.3	12.7 ± 0.6	17.9 ± 0.7	0.173 ± 0.028

well recovered (Fig. 2c). Interestingly, recovery of variables retaining the sufficient-nonlinearity construction begins to degrade more noticeably as the fraction of sufficient-nonlinearity-violating variables approaches roughly 0.6. This is close to the dimensional boundary implied by the projected-rank condition in Theorem 2, $|\mathcal{S}| - \mathrm{rank}(\hat{\boldsymbol{B}}_{\mathcal{S}}^{\ell}) \geq d_{\mathbf{z}}^{\ell}$. For our synthetic dimensions, this gives a generic bound of $\lesssim 0.57$. We emphasize that this is only a necessary dimensional condition; the theorem does not predict a sharp finite-sample phase transition.

Robustness to model misspecification. We further evaluate HIRL under latent-dimension misspecification (Appendix Fig. 3b). Performance remains relatively stable under moderate over-specification, whereas under-specification causes a clearer degradation, particularly for the root representation. At the bottom layer, optimal linear alignment substantially improves recovery of $\mathbf{z}^0$ under misspecification, suggesting that part of the degradation in permutation-based recovery may arise from additional linear mixing rather than complete loss of latent information.

Real-data evaluation on CIFAR-100. We evaluate semantic representation quality on CIFAR-100 (Krizhevsky & Hinton, 2009) using layer-wise linear probes for the 100 classes (Table 2). HIRL's probe accuracy increases from 14.2% at $\boldsymbol{z}^0$ to 20.4% at $\boldsymbol{z}^3$ and exceeds the reduced-feature variant and both hierarchical baselines at every latent layer.

Importantly, both VLAE and LVAE achieve lower test MSE than HIRL (see also Appendix Fig. 4), despite weaker linear-probe performance across the latent hierarchy. This gap may partly reflect architectural differences: LVAE and VLAE allow multiple latent components to contribute to reconstruction, whereas HIRL constructs the hierarchy deterministically from a single root representation. The baselines also have substantially larger parameter counts, so the reconstruction comparison should not be interpreted as a capacity-matched benchmark (Appendix Table 3). Thus, HIRL's stronger semantic representations are not simply explained by reconstruction quality, suggesting that the proposed hierarchical parameterization can help organize semantic information across layers.

6 Discussion and conclusion

Implications. Our results show that hierarchical identifiability requires controlling inter-layer reparameterization symmetries. Sufficient nonlinear variation and overdetermined mixing provide one concrete architectural mechanism for doing so, and are compatible with broad classes of overparameterized nonlinear architectures rather than requiring highly specialized structures. This suggests that identifiability can be treated as an architectural design objective rather than only a property of latent distributions, with these constraints also serving as useful inductive biases when the theoretical assumptions are only approximately satisfied.

Limitations and future directions. Our guarantees rely on additive transitions, dimensional expansion, and nondegeneracy conditions, trading expressive flexibility for identifiability. Real data may violate these assumptions, and higher-order (non-additive) interactions among variables may permit block-wise nonlinear ambiguities, complicating the recursive identifiability argument. Understanding how far these assumptions can be relaxed remains an important direction for future work.

Conclusion. Overall, we establish that hierarchical identifiability is not a trivial layer-wise extension of flat identifiability: architectural constraints can propagate across layers to restrict inter-layer ambiguities, enabling recovery of intermediate representations and layer-wise innovations from observational data within the proposed model class and suggesting identifiability as a practical architectural design principle.

AI USE STATEMENT

In this work, we used generative AI tools to assist with the development and checking of theoretical arguments, including the formulation and refinement of mathematical claims and proof strategies. We also used these tools to provide feedback on experimental design and interpretation of experimental results, improve the organization, clarity, and wording of the manuscript, assist with literature search and positioning, and support the editing and debugging of experimental code.

All AI-assisted theoretical arguments were independently checked by the authors, including their assumptions, derivations, and consistency with the stated results. AI-assisted suggestions regarding experimental design and interpretation were independently evaluated by the authors, and all AI-assisted code was reviewed and tested. Generative AI tools were used only in an assistive capacity and not as a substitute for verification of mathematical, empirical, or bibliographic claims.

The authors have reviewed all AI-assisted material and take full responsibility for the final content of this work, including its text, claims, proofs, experimental analyses, and code.

ETHICS STATEMENT

This work is methodological and theoretical in nature and does not involve human subjects, personally identifiable information, or sensitive personal data. The empirical evaluation uses synthetic data and publicly available benchmark data, and we do not introduce or release a new dataset containing human-subject information.

The proposed method concerns identifiability and representation learning in hierarchical latent-variable models and is not specifically designed for high-stakes decision-making applications. We are not aware of ethical, privacy, security, or legal concerns specific to the experiments conducted in this work beyond those generally associated with the use of public benchmark data and machine-learning models.

REPRODUCIBILITY STATEMENT

We provide the information needed to verify the theoretical results and reproduce the empirical results of this work. All assumptions used in the identifiability results are stated explicitly in the main text, and complete proofs of the main theorems are provided in the appendix. The experimental setup, synthetic data-generation procedure, model architectures, optimization settings, evaluation metrics, and additional ablation and robustness studies are also described in Appendix D. We report results over multiple random seeds and include standard deviations where applicable. The implementation used for the experiments and scripts for reproducing the reported results are included in the supplementary material.

REFERENCES

K. Ahuja, J. S Hartford, and Y. Bengio. Weakly supervised representation learning with sparse perturbations. In *Advances in Neural Information Processing Systems*, volume 35, pp. 15516–15528, 2022a.

K. Ahuja, Y. Wang, D. Mahajan, and Y. Bengio. Interventional causal representation learning. In *NeurIPS 2022 Workshop on Neuro Causal and Symbolic AI (nCSI)*, 2022b.

Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

J. Brady, R. S. Zimmermann, Y. Sharma, B. Schölkopf, J. Von Kügelgen, and W. Brendel. Provably learning object-centric representations. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 3038–3062, 2023.

Jack Brady, Julius von Kügelgen, Sebastien Lachapelle, Simon Buchholz, Thomas Kipf, and Wieland Brendel. Interaction asymmetry: A general principle for learning composable abstractions. In *The Thirteenth International Conference on Learning Representations*, 2025.

J. Brehmer, P. de Haan, P. Lippe, and T. S Cohen. Weakly supervised causal representation learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 38319–38331, 2022.

S. Buchholz and B. Schölkopf. Robustness of nonlinear representation learning. In *Forty-first International Conference on Machine Learning*, 2024.

S. Buchholz, M. Besserve, and B. Schölkopf. Function classes for identifiable nonlinear independent component analysis. In *Advances in Neural Information Processing Systems*, volume 35, pp. 16946–16961, 2022.

S. Buchholz, G. Rajendran, E. Rosenfeld, B. Aragam, B. Schölkopf, and P. Ravikumar. Learning linear causal representations from interventions under general nonlinear mixing. In *Advances in Neural Information Processing Systems*, volume 36, pp. 45419–45462, 2023.

I. Daunhawer, A. Bizeul, E. Palumbo, A. Marx, and J. E Vogt. Identifiability results for multimodal contrastive learning. In *The Eleventh International Conference on Learning Representations*, 2023.

L. Dinh, J. Sohl-Dickstein, and S. Bengio. Density estimation using real NVP. In *International Conference on Learning Representations*, 2017.

L. Gresele, J. Von Kügelgen, V. Stimper, B. Schölkopf, and M. Besserve. Independent mechanism analysis, a new concept? In *Advances in Neural Information Processing Systems*, 2021.

D. Horan, E. Richardson, and Y. Weiss. When is unsupervised disentanglement possible? In *Advances in Neural Information Processing Systems*, volume 34, pp. 5150–5161, 2021.

P. O. Hoyer, D. Janzing, J. Mooij, J. Peters, and B. Schölkopf. Nonlinear causal discovery with additive noise models. In *Proceedings of the 21st International Conference on Neural Information Processing Systems*, pp. 689–696, 2008.

A. Hyvärinen and H. Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ICA. In *Advances in Neural Information Processing Systems (NIPS) 29*, pp. 3765–3773. 2016.

A. Hyvarinen and H. Morioka. Nonlinear ICA of temporally dependent stationary sources. In *AISTATS*, pp. 460–469, 2017.

A. Hyvärinen and P. Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Netw.*, 12(3):429 – 439, 1999.

A. Hyvarinen, H. Sasaki, and R. Turner. Nonlinear ICA using auxiliary variables and generalized contrastive learning. In *AISTATS*, pp. 859–868, 2019.

Aapo Hyvärinen, Ilyes Khemakhem, and Hiroshi Morioka. Nonlinear independent component analysis for principled disentanglement in unsupervised deep learning. *Patterns*, 4(10), 2023.

I. Khemakhem, D. P. Kingma, R. P. Monti, and A. Hyvärinen. Variational autoencoders and nonlinear ICA: A unifying framework. In *AISTATS*, 2020.

D. P Kingma and P. Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Bohdan Kivva, Goutham Rajendran, Pradeep Ravikumar, and Bryon Aragam. Identifiability of deep generative models without auxiliary information. In *Advances in Neural Information Processing Systems*, volume 35, pp. 15687–15701, 2022.

A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

S. Lachapelle, T. Deleu, D. Mahajan, I. Mitliagkas, Y. Bengio, S. Lacoste-Julien, and Q. Bertrand. Synergies between disentanglement and sparsity: Generalization and identifiability in multi-task learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 18171–18206, 2023a.

Sébastien Lachapelle, Divyat Mahajan, Ioannis Mitliagkas, and Simon Lacoste-Julien. Additive decoders for latent variables identification and cartesian-product extrapolation. In *Advances in Neural Information Processing Systems*, volume 36, pp. 25112–25150, 2023b.

Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

T. Leemann, M. Kirchhof, Y. Rong, E. Kasneci, and G. Kasneci. When are post-hoc conceptual explanations identifiable? In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216, pp. 1207–1218, 2023.

Zijian Li, Minghao Fu, Junxian Huang, Yifan Shen, Ruichu Cai, Yuewen Sun, Guangyi Chen, and Kun Zhang. Towards identifiability of hierarchical temporal causal representation learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

F. Locatello, S. Bauer, M. Lucic, G. Raetsch, S. Gelly, B. Schölkopf, and O. Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 4114–4124, 2019.

Q. Lyu, X. Fu, W. Wang, and S. Lu. Understanding latent correlation-based multiview learning and self-supervision: An identifiability perspective. In *International Conference on Learning Representations*, 2022.

Qi Lyu and Xiao Fu. Provable subspace identification under post-nonlinear mixtures. In *Advances in Neural Information Processing Systems*, volume 35, pp. 1554–1567, 2022.

G. E. Moran, D. Sridhar, Y. Wang, and D. Blei. Identifiable deep generative models via sparse decoding. *Transactions on Machine Learning Research*, 2022.

H. Morioka and A. Hyvarinen. Connectivity-contrastive learning: Combining causal discovery and representation learning for multimodal data. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206, pp. 3399–3426, 2023.

H. Morioka and A. Hyvarinen. Causal representation learning made identifiable by grouping of observational variables. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 36249–36293, 2024.

G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22 (57):1–64, 2021.

R. Ranganath, D. Tran, and D. Blei. Hierarchical variational models. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48, pp. 324–333, 2016.

B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio. Towards causal representation learning. arXiv, 2021.

X. Shen, F. Liu, H. Dong, Q. Lian, Z. Chen, and T. Zhang. Weakly supervised disentangled generative causal representation learning. *Journal of Machine Learning Research*, 23(241):1–55, 2022.

C. K. Sønderby, T. Raiko, L. Maaløe, Søren K. Sønderby, and O. Winther. Ladder variational autoencoders. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

C. Squires, A. Seigal, S. Bhate, and C. Uhler. Linear causal disentanglement via interventions. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

N. Sturma, C. Squires, M. Drton, and C. Uhler. Unpaired multi-domain causal representation learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 34465–34492, 2023.

A. Vahdat and J. Kautz. Nvae: A deep hierarchical variational autoencoder. In *Advances in Neural Information Processing Systems*, volume 33, pp. 19667–19679, 2020.

B. Varici, E. Acarturk, K. Shanmugam, A. Kumar, and A. Tajer. Score-based causal representation learning with interventions. arXiv, 2023.

Julius von Kügelgen, Michel Besserve, Liang Wendong, Luigi Gresele, Armin Kekić, Elias Bareinboim, David Blei, and Bernhard Schölkopf. Nonparametric identifiability of causal representations from unknown interventions. In *Advances in Neural Information Processing Systems*, volume 36, pp. 48603–48638, 2023.

M. Yang, F. Liu, Z. Chen, X. Shen, J. Hao, and J. Wang. CausalVAE: Disentangled representation learning via neural structural causal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9593–9602, 2021.

D. Yao, D. Xu, S. Lachapelle, S. Magliacane, P. Taslakian, G. Martius, J. von Kügelgen, and F. Locatello. Multi-view causal representation learning with partial observability. arXiv, 2023.

S. Zhao, J. Song, and S. Ermon. Learning hierarchical features from deep generative models. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pp. 4091–4099, 2017.

Y. Zheng, I. Ng, and K. Zhang. On the identifiability of nonlinear ICA: Sparsity and beyond. In *Advances in Neural Information Processing Systems*, 2022.

A Interpretation of assumptions

We first state an additional technical regularity condition used in the deconvolution step of the proof.

Assumption 8 (Deconvolution regularity)**.** *For each scalar coordinate transformation h_a arising in the inter-layer correspondence, the Gaussian convolution $h_a * \phi_\sigma$ is well defined for every relevant $\sigma \geq 0$, and h_a'' belongs to a function or distribution class on which convolution with a nondegenerate Gaussian is injective (e.g., h_a'' is a tempered distribution). Moreover, the relevant range of each conditional-mean coordinate μ_a contains a nonempty open interval.*

This assumption is primarily technical. Conditional averaging produces a Gaussian-smoothed coordinate transformation, and the null-space argument establishes that this smoothed transformation is affine. The deconvolution condition is used only to transfer this conclusion back to the original coordinate transformation, analogous to characteristic-function conditions used to deconvolve additive observation noise in identifiable latent-variable models (Khemakhem et al., 2020).

We next summarize the distinct roles of Assumptions 1–7.

Assumption 1 imposes local invertibility of each stochastic transition with respect to the current state and innovation. Invertibility or injectivity of the mixing transformation is a standard regularity condition in identifiable nonlinear representation learning, including nonlinear ICA (Hyvarinen et al., 2019; Hyvärinen et al., 2023) and causal representation learning (von Kügelgen et al., 2023; Morioka & Hyvarinen, 2024). In our setting, local invertibility ensures that, after the innovation directions are identified and projected out, the projected conditional mean retains enough local information to recover the current-layer state. The dimensional equality $d_z^{\ell+1} = d_z^\ell + d_\epsilon^\ell$ is the corresponding dimension-matching requirement for treating the full state–innovation transition as a local diffeomorphism.

Assumption 2 (Sufficient nonlinearity) is a structural richness condition on the within-coordinate nonlinear features. It can be viewed as the singleton-block specialization of the sufficient nonlinearity condition of Lachapelle et al. (2023b): when each latent block contains a single coordinate, their condition reduces to full column rank of the concatenated first- and coordinate-wise second-derivative matrix. Since $\mu^\ell = \boldsymbol{A}^\ell r^\ell$ and $\boldsymbol{A}^\ell$ has full column rank, Assumption 2 is equivalent to

$$\operatorname{rank}\left[M_{\mu^\ell}(\mathbf{z}^\ell)\; J_{\mu^\ell}(\mathbf{z}^\ell)\right] = 2d_z^\ell.$$

Its role here is complementary to Assumption 7: the latter eliminates nonlinear ambiguity in the next-layer transformation, whereas sufficient nonlinearity restricts the induced current-layer reparameterization to act component-wise up to permutation. The full-rank formulation is a sufficient condition ensuring that first- and second-order variations provide enough information for this coordinate separation.

Assumption 3 (Overdetermined mixing) provides a nontrivial left null space used to eliminate the learned conditional mean mixing. It is therefore the geometric mechanism that makes the null-space

constraints available. Once $d_r^\ell < d_z^{\ell+1}$ is imposed, full column rank is generic, since rank-deficient mixing matrices form a measure-zero subset. This dimensional condition alone does not guarantee Assumption 7, which additionally requires sufficiently rich cross-coordinate variation after mixing.

Assumption 4 (Gaussian innovations and innovation-subspace identifiability) combines a structural identifiability requirement with a convenient distributional formulation. Its essential structural role is to ensure that the innovation directions are distinguishable from variation inherited through the current-layer state. The Gaussian-flat characterization $\mathcal{G}(p_{\mathbf{z}^{\ell+1}}) = \mathrm{Im}(\boldsymbol{B}^\ell)$ provides one concrete way to express this distinguishability by ruling out additional Gaussian-flat directions outside the innovation subspace. Gaussianity additionally enables the convolution and deconvolution arguments used in the proof.

Assumption 5 (Learned stochastic hierarchy) restricts the learned hierarchy to the same structured transition class as the true model. This is a model-class restriction rather than a data-distribution assumption, ensuring that identifiability is analyzed within the same structured hierarchical family.

Assumption 6 (Conditional centering of learned innovations) is a population-level moment condition on the learned innovation rather than a directly observable training constraint. Its role in our proof is to ensure that conditioning on the true current-layer state removes the learned innovation contribution and exposes the deterministic conditional-mean relation required by the null-space argument. It does not require access to $\mathbf{z}^\ell$ during training; rather, it restricts the admissible population-level decompositions by ruling out systematic state-dependent variation in the learned innovation.

Assumption 7 (Null-space nondegeneracy condition) is a nondegeneracy condition on the nonlinear variation captured by the left-null-space constraints of the learned conditional-mean mixing. Its role is to ensure that mixed derivatives with respect to distinct current-layer coordinates provide sufficiently many independent constraints to eliminate nonlinear next-layer reparameterizations. The Khatri–Rao structure arises because these constraints are stacked over all pairs of current-layer coordinates and all left-null-space directions. The particular Khatri–Rao full-rank condition used here is a sufficient condition tailored to the mixed-derivative null-space argument and is not intended as a necessary characterization of hierarchical identifiability.

Structural ingredients and sufficient formulations. Thus, the main structural ingredients are local recoverability of the state–innovation decomposition (Assumption 1), distinguishability of innovation directions (Assumption 4), sufficiently rich within-coordinate nonlinear variation (Assumption 2), and sufficiently rich cross-coordinate variation after mixing (Assumption 7), supported by the overdetermined geometry of Assumption 3.

The specific conditions used here—namely, the Gaussian-flat characterization (Assumption 4), mixed-derivative rank condition (Assumption 7), conditional-centering condition (Assumption 6), and deconvolution regularity (Assumption 8)—provide sufficient formulations for our analysis and should not be interpreted as necessary characterizations of hierarchical identifiability; some of their precise forms are tailored to the proof strategy.

B Proof of Theorem 1

Proof. We proceed by backward induction, restricting throughout to a common connected neighborhood on which all stated regularity conditions and local correspondences hold.

At the observation layer,

$$\hat{\mathbf{z}}^L = \mathbf{z}^L,$$

so the corresponding transformation is

$$\boldsymbol{k}^L = \mathrm{id},$$

which is component-wise and affine.

Consider first the transition at layer $L-1$. Since the child correspondence at layer L is the identity, the premise of Proposition 1 holds with $\boldsymbol{h} = \mathrm{id}$. Therefore,

$$\hat{\mathbf{z}}^{L-1} = \boldsymbol{k}^{L-1}(\mathbf{z}^{L-1}),$$

where $\boldsymbol{k}^{L-1}$ is component-wise up to a permutation, and

$$\hat{\epsilon}^{L-1} = \boldsymbol{R}^{L-1}\epsilon^{L-1}$$

for some invertible $\boldsymbol{R}^{L-1}$.

Now suppose that for some $\ell \in \{0, \dots, L-2\}$, the correspondence at layer $\ell+1$ has already been established as

$$\hat{\mathbf{z}}^{\ell+1} = \boldsymbol{k}^{\ell+1}(\mathbf{z}^{\ell+1}),$$

where $\boldsymbol{k}^{\ell+1}$ is component-wise up to a permutation. Thus the premise of Proposition 1 holds for layer ℓ with

$$\boldsymbol{h} = \boldsymbol{k}^{\ell+1}.$$

The proposition sharpens the child correspondence to

$$\boldsymbol{k}^{\ell+1}(\mathbf{z}^{\ell+1}) = \boldsymbol{H}^{\ell+1}\mathbf{z}^{\ell+1} + \boldsymbol{c}^{\ell+1},$$

where $\boldsymbol{H}^{\ell+1}$ is nonsingular and monomial. It also yields the parent correspondence

$$\hat{\mathbf{z}}^{\ell} = \boldsymbol{k}^{\ell}(\mathbf{z}^{\ell}),$$

where $\boldsymbol{k}^{\ell}$ is component-wise up to a permutation, together with

$$\hat{\epsilon}^{\ell} = \boldsymbol{R}^{\ell}\epsilon^{\ell}$$

for some invertible $\boldsymbol{R}^{\ell}$. This establishes the induction step.

Consequently, for every $\ell = 1, \dots, L-1$, the map $\boldsymbol{k}^{\ell}$, initially obtained as a component-wise correspondence by applying the proposition to layer ℓ, subsequently appears as the child transformation in the proposition applied to layer $\ell-1$. It is therefore further restricted to an affine monomial map:

$$\hat{\mathbf{z}}^{\ell} = \boldsymbol{H}^{\ell}\mathbf{z}^{\ell} + \boldsymbol{c}^{\ell}.$$

Writing

$$\boldsymbol{H}^{\ell} = \boldsymbol{P}^{\ell}\boldsymbol{D}^{\ell},$$

with $\boldsymbol{P}^{\ell}$ a permutation matrix and $\boldsymbol{D}^{\ell}$ a nonsingular diagonal matrix, gives the stated intermediate-layer ambiguity.

At the root layer, no lower transition exists to impose this additional restriction. The proposition gives only a component-wise correspondence up to permutation. Writing the root correspondence as

$$\hat{\mathbf{z}}^{0} = \boldsymbol{P}^{0}\tilde{\boldsymbol{k}}^{0}(\mathbf{z}^{0}),$$

where $\tilde{\boldsymbol{k}}^{0}$ is component-wise and each component is a scalar local diffeomorphism, gives the stated root-layer ambiguity.

Finally, repeated application of Proposition 1 yields

$$\hat{\epsilon}^{\ell} = \boldsymbol{R}^{\ell}\epsilon^{\ell}, \qquad \ell = 0, \dots, L-1,$$

with each $\boldsymbol{R}^{\ell}$ invertible. □

Lemma 1 (Coordinate separation under additive sufficient nonlinearity)**.** *Let*

$$\boldsymbol{\mu}, \hat{\boldsymbol{\mu}} : \mathbb{R}^d \to \mathbb{R}^n$$

be C^2 additive maps. Suppose that a C^2 local diffeomorphism $\boldsymbol{k} : \mathbb{R}^d \to \mathbb{R}^d$ satisfies

$$\hat{\boldsymbol{\mu}}(\boldsymbol{k}(\mathbf{z})) = \boldsymbol{H}\boldsymbol{\mu}(\mathbf{z}) + \boldsymbol{c}, \tag{8}$$

where $\boldsymbol{H} \in \mathbb{R}^{n\times n}$ is nonsingular.

If

$$\operatorname{rank}\left[\mathbf{M}_{\boldsymbol{\mu}}(\mathbf{z}) \quad \mathbf{J}_{\boldsymbol{\mu}}(\mathbf{z})\right] = 2d, \tag{9}$$

then, locally,

$$\boldsymbol{k}(\mathbf{z}) = \boldsymbol{P}\begin{bmatrix} k_1(z_1) \\ \vdots \\ k_d(z_d) \end{bmatrix},$$

where $\boldsymbol{P}$ is a permutation matrix and each k_i is a scalar local diffeomorphism.

Proof. Collecting the first derivatives and coordinate-wise second derivatives of equation 8 gives

$$[\mathbf{M}_{\hat{\boldsymbol{\mu}}}(\boldsymbol{k}(\mathbf{z})) \quad \mathbf{J}_{\hat{\boldsymbol{\mu}}}(\boldsymbol{k}(\mathbf{z}))] \begin{bmatrix} \mathbf{J}_{\boldsymbol{k}}(\mathbf{z}) \odot \mathbf{J}_{\boldsymbol{k}}(\mathbf{z}) & \mathbf{0} \\ \mathbf{M}_{\boldsymbol{k}}(\mathbf{z}) & \mathbf{J}_{\boldsymbol{k}}(\mathbf{z}) \end{bmatrix} = \boldsymbol{H} [\mathbf{M}_{\boldsymbol{\mu}}(\mathbf{z}) \quad \mathbf{J}_{\boldsymbol{\mu}}(\mathbf{z})], \tag{10}$$

where $\mathbf{J}_{\boldsymbol{k}} \odot \mathbf{J}_{\boldsymbol{k}}$ denotes the column-wise Hadamard square, whose jth column is $\partial_j \boldsymbol{k} \odot \partial_j \boldsymbol{k}$. By equation 9 and the nonsingularity of $\boldsymbol{H}$, the right-hand side has column rank $2d$. Hence the product on the left-hand side also has rank $2d$. Since

$$[\mathbf{M}_{\hat{\boldsymbol{\mu}}}(\boldsymbol{k}(\mathbf{z})) \quad \mathbf{J}_{\hat{\boldsymbol{\mu}}}(\boldsymbol{k}(\mathbf{z}))]$$

has exactly $2d$ columns, it must have full column rank.

For $i \neq j$, additivity of $\boldsymbol{\mu}$ gives

$$\partial^2_{ij} \boldsymbol{\mu}(\mathbf{z}) = \mathbf{0}.$$

Taking the mixed derivative of equation 8, and using additivity of $\hat{\boldsymbol{\mu}}$, therefore yields

$$\mathbf{M}_{\hat{\boldsymbol{\mu}}}(\boldsymbol{k}(\mathbf{z})) \left(\partial_i \boldsymbol{k}(\mathbf{z}) \odot \partial_j \boldsymbol{k}(\mathbf{z})\right) + \mathbf{J}_{\hat{\boldsymbol{\mu}}}(\boldsymbol{k}(\mathbf{z})) \partial^2_{ij} \boldsymbol{k}(\mathbf{z}) = \mathbf{0}. \tag{11}$$

Since $[\mathbf{M}_{\hat{\boldsymbol{\mu}}} \; \mathbf{J}_{\hat{\boldsymbol{\mu}}}]$ has full column rank, equation 11 implies

$$\partial_i \boldsymbol{k}(\mathbf{z}) \odot \partial_j \boldsymbol{k}(\mathbf{z}) = \mathbf{0}, \qquad \partial^2_{ij} \boldsymbol{k}(\mathbf{z}) = \mathbf{0}, \qquad i \neq j.$$

Thus distinct columns of $\mathbf{J}_{\boldsymbol{k}}(\mathbf{z})$ have disjoint row supports. Since $\boldsymbol{k}$ is a local diffeomorphism, $\mathbf{J}_{\boldsymbol{k}}(\mathbf{z})$ is nonsingular. Hence its d nonzero columns have pairwise disjoint supports within d rows, so each column has exactly one nonzero entry, and each row likewise contains exactly one nonzero entry. The resulting permutation pattern is locally constant and therefore constant on each connected component of the regular region. After applying this permutation, $\mathbf{J}_{\boldsymbol{k}}(\mathbf{z})$ is diagonal, so each output coordinate depends locally on only one input coordinate. Therefore,

$$\boldsymbol{k}(\mathbf{z}) = \boldsymbol{P} \begin{bmatrix} k_1(z_1) \\ \vdots \\ k_d(z_d) \end{bmatrix}.$$

Since $\mathbf{J}_{\boldsymbol{k}}$ is nonsingular, each k_i' is nonzero locally, and hence each k_i is a scalar local diffeomorphism. □

Proposition 1 (Stochastic null-space rigidity)**.** *Consider one layer of the true and learned hierarchies:*

$$\mathbf{y} = \boldsymbol{\mu}(\mathbf{z}) + \boldsymbol{B}\epsilon, \qquad \mathbf{z} \in \mathbb{R}^d, \qquad \epsilon \in \mathbb{R}^q, \qquad \mathbf{y} \in \mathbb{R}^m, \tag{12}$$

and

$$\hat{\mathbf{y}} = \hat{\boldsymbol{\mu}}(\hat{\mathbf{z}}) + \hat{\boldsymbol{B}}\hat{\epsilon}, \qquad \hat{\mathbf{z}} \in \mathbb{R}^d, \qquad \hat{\epsilon} \in \mathbb{R}^q, \qquad \hat{\mathbf{y}} \in \mathbb{R}^m, \tag{13}$$

where $\boldsymbol{\mu}, \hat{\boldsymbol{\mu}} : \mathbb{R}^d \to \mathbb{R}^m$, $\boldsymbol{B}, \hat{\boldsymbol{B}} \in \mathbb{R}^{m \times q}$. Suppose that the next-layer representations are related by a component-wise transformation up to permutation,

$$\hat{\mathbf{y}} = \boldsymbol{h}(\mathbf{y}) = \boldsymbol{P} \begin{bmatrix} h_1(y_1) \\ \vdots \\ h_m(y_m) \end{bmatrix}, \tag{14}$$

where every h_a is a C^2 scalar diffeomorphism.

Suppose that Assumptions 1–7 hold, together with Assumption 8. Then, on every connected neighborhood on which the stated regularity conditions hold, there exist

1. *a nonsingular monomial matrix $\boldsymbol{H}$ and a vector $\mathbf{c}$,*
2. *an invertible matrix $\boldsymbol{R} \in \mathrm{GL}(q)$, and*
3. *a local diffeomorphism $\boldsymbol{k} : \mathbb{R}^d \to \mathbb{R}^d$ that is component-wise up to permutation,*

such that

$$\boldsymbol{h}(\mathbf{y}) = \boldsymbol{H}\mathbf{y} + \boldsymbol{c}, \tag{15}$$
$$\hat{\mathbf{z}} = \boldsymbol{k}(\mathbf{z}), \tag{16}$$
$$\hat{\boldsymbol{\mu}}(\boldsymbol{k}(\mathbf{z})) = \boldsymbol{H}\boldsymbol{\mu}(\mathbf{z}) + \boldsymbol{c}, \tag{17}$$
$$\boldsymbol{H}\boldsymbol{B} = \hat{\boldsymbol{B}}\boldsymbol{R}, \tag{18}$$
$$\hat{\epsilon} = \boldsymbol{R}\epsilon. \tag{19}$$

Proof. Without loss of generality, absorb the permutation in equation 14 into the learned coordinates, so that

$$\boldsymbol{h}(\mathbf{y}) = \begin{bmatrix} h_1(y_1) \\ \vdots \\ h_m(y_m) \end{bmatrix}.$$

Step 1: Conditional averaging. For each coordinate a, define

$$\bar{h}_a(t) := \mathbb{E}_\epsilon \left[h_a \left(t + \boldsymbol{e}_a^\top \boldsymbol{B}\epsilon \right) \right], \tag{20}$$

where $\boldsymbol{e}_a$ is the ath canonical basis vector. Since $\epsilon \perp\!\!\!\perp \mathbf{z}$,

$$\mathbb{E}[\boldsymbol{h}(\mathbf{y}) \mid \mathbf{z}] = \bar{\boldsymbol{h}}(\boldsymbol{\mu}(\mathbf{z})). \tag{21}$$

Let $\hat{\boldsymbol{Q}}_{\boldsymbol{A}}$ span the left null space of $\hat{\boldsymbol{A}}$, so that

$$\hat{\boldsymbol{Q}}_{\boldsymbol{A}}^\top \hat{\boldsymbol{A}} = \mathbf{0}.$$

Since

$$\hat{\boldsymbol{\mu}}(\hat{\mathbf{z}}) = \hat{\boldsymbol{A}}\hat{\boldsymbol{r}}(\hat{\mathbf{z}}),$$

we have

$$\hat{\boldsymbol{Q}}_{\boldsymbol{A}}^\top \hat{\boldsymbol{\mu}}(\hat{\mathbf{z}}) = \mathbf{0}$$

pointwise. Together with

$$\mathbb{E}[\hat{\epsilon} \mid \mathbf{z}] = \mathbf{0},$$

the learned transition therefore gives

$$\hat{\boldsymbol{Q}}_{\boldsymbol{A}}^\top \mathbb{E}[\boldsymbol{h}(\mathbf{y}) \mid \mathbf{z}] = \hat{\boldsymbol{Q}}_{\boldsymbol{A}}^\top \mathbb{E}[\hat{\boldsymbol{\mu}}(\hat{\mathbf{z}}) + \hat{\boldsymbol{B}}\hat{\epsilon} \mid \mathbf{z}] = \mathbf{0}.$$

Therefore,

$$\hat{\boldsymbol{Q}}_{\boldsymbol{A}}^\top \bar{\boldsymbol{h}}(\boldsymbol{\mu}(\mathbf{z})) = \mathbf{0}. \tag{22}$$

Step 2: Stochastic null-space argument. For $i \neq j$, additivity of $\boldsymbol{\mu}$ implies

$$\partial_{ij}^2 \boldsymbol{\mu}(\mathbf{z}) = \mathbf{0}.$$

Taking mixed derivatives of equation 22 yields

$$\hat{\boldsymbol{Q}}_{\boldsymbol{A}}^\top \operatorname{diag}\left(\bar{\boldsymbol{h}}''(\boldsymbol{\mu}(\mathbf{z}))\right) \left(\partial_i \boldsymbol{\mu}(\mathbf{z}) \odot \partial_j \boldsymbol{\mu}(\mathbf{z})\right) = \mathbf{0}. \tag{23}$$

Collecting these equations over all $1 \leq i < j \leq d$ gives

$$\left[\hat{\boldsymbol{Q}}_{\boldsymbol{A}}^\top \odot_{\mathrm{KR}} \mathbf{W}_{\boldsymbol{\mu}}(\mathbf{z})^\top\right] \bar{\boldsymbol{h}}''(\boldsymbol{\mu}(\mathbf{z})) = \mathbf{0}. \tag{24}$$

By Assumption 7,

$$\bar{\boldsymbol{h}}''(\boldsymbol{\mu}(\mathbf{z})) = \mathbf{0}$$

on the relevant range of $\boldsymbol{\mu}$.

Step 3: Gaussian deconvolution. For each coordinate,

$$\bar{h}_a = h_a * \phi_a, \qquad \bar{h}_a'' = h_a'' * \phi_a,$$

where ϕ_a is the centered Gaussian measure induced by $\boldsymbol{e}_a^\top \boldsymbol{B}\epsilon$, with variance

$$\sigma_a^2 := \mathrm{Var}\big(\boldsymbol{e}_a^\top \boldsymbol{B}\epsilon\big).$$

By Assumption 8, the relevant range of μ_a contains a nonempty open interval. If $\sigma_a^2 > 0$, Gaussian smoothing makes $\bar{h}_a''$ real analytic. Since Step 2 gives $\bar{h}_a'' = 0$ on this interval, the identity theorem implies $\bar{h}_a'' \equiv 0$. Injectivity of Gaussian convolution then gives $h_a'' = 0$. If $\sigma_a^2 = 0$, then $\bar{h}_a = h_a$, so Step 2 directly gives $h_a'' = 0$ on the relevant range.

Thus, in either case, h_a is affine on the relevant range. Since each h_a is a scalar diffeomorphism, its affine slope is nonzero. Restoring the permutation absorbed at the beginning of the proof therefore gives

$$\boldsymbol{h}(\mathbf{y}) = \boldsymbol{H}\mathbf{y} + \boldsymbol{c},$$

where $\boldsymbol{H}$ is a nonsingular monomial matrix.

Step 4: Innovation-subspace identification and state recovery. For any invertible affine transformation $\tilde{\mathbf{y}} = \boldsymbol{H}\mathbf{y} + \boldsymbol{c}$, Gaussian-flat directions transform equivariantly:

$$\mathcal{G}(p_{\tilde{\mathbf{y}}}) = \boldsymbol{H}\,\mathcal{G}(p_{\mathbf{y}}).$$

Indeed, writing $\boldsymbol{y} = \boldsymbol{H}^{-1}(\tilde{\boldsymbol{y}} - \boldsymbol{c})$, the change-of-variables formula gives

$$\log p_{\tilde{\mathbf{y}}}(\tilde{\boldsymbol{y}}) = \log p_{\mathbf{y}}(\boldsymbol{y}) + \mathrm{const}.$$

Hence, by the chain rule,

$$D^3 \log p_{\tilde{\mathbf{y}}}(\tilde{\boldsymbol{y}})[\tilde{\boldsymbol{u}}, \tilde{\boldsymbol{v}}, \tilde{\boldsymbol{v}}] = D^3 \log p_{\mathbf{y}}(\boldsymbol{y})[\boldsymbol{H}^{-1}\tilde{\boldsymbol{u}}, \boldsymbol{H}^{-1}\tilde{\boldsymbol{v}}, \boldsymbol{H}^{-1}\tilde{\boldsymbol{v}}].$$

Since $\boldsymbol{H}$ is invertible, $\tilde{\boldsymbol{u}}$ ranges over all directions if and only if $\boldsymbol{H}^{-1}\tilde{\boldsymbol{u}}$ does. Therefore

$$\tilde{\boldsymbol{v}} \in \mathcal{G}(p_{\tilde{\mathbf{y}}}) \quad \Longleftrightarrow \quad \boldsymbol{H}^{-1}\tilde{\boldsymbol{v}} \in \mathcal{G}(p_{\mathbf{y}}),$$

which proves the transformation rule. Since $\hat{\mathbf{y}} = \boldsymbol{H}\mathbf{y} + \boldsymbol{c}$, Assumption 4 therefore implies

$$\mathrm{Im}(\hat{\boldsymbol{B}}) = \mathcal{G}(p_{\hat{\mathbf{y}}}) = \boldsymbol{H}\mathcal{G}(p_{\mathbf{y}}) = \mathrm{Im}(\boldsymbol{H}\boldsymbol{B}). \tag{25}$$

Let $\hat{\boldsymbol{Q}}_{\boldsymbol{B}}$ span the left null space of $\hat{\boldsymbol{B}}$:

$$\hat{\boldsymbol{Q}}_{\boldsymbol{B}}^\top \hat{\boldsymbol{B}} = \mathbf{0}.$$

Then

$$\hat{\boldsymbol{Q}}_{\boldsymbol{B}}^\top \boldsymbol{H}\boldsymbol{B} = \mathbf{0}.$$

Applying $\hat{\boldsymbol{Q}}_{\boldsymbol{B}}^\top$ to the sample-wise equality

$$\boldsymbol{H}\left[\boldsymbol{\mu}(\mathbf{z}) + \boldsymbol{B}\epsilon\right] + \boldsymbol{c} = \hat{\boldsymbol{\mu}}(\hat{\mathbf{z}}) + \hat{\boldsymbol{B}}\hat{\epsilon}$$

gives

$$\hat{\boldsymbol{Q}}_{\boldsymbol{B}}^\top \left[\boldsymbol{H}\boldsymbol{\mu}(\mathbf{z}) + \boldsymbol{c}\right] = \hat{\boldsymbol{Q}}_{\boldsymbol{B}}^\top \hat{\boldsymbol{\mu}}(\hat{\mathbf{z}}). \tag{26}$$

Define

$$\boldsymbol{g}(\mathbf{z}) := \hat{\boldsymbol{Q}}_{\boldsymbol{B}}^\top \left[\boldsymbol{H}\boldsymbol{\mu}(\mathbf{z}) + \boldsymbol{c}\right], \qquad \hat{\boldsymbol{g}}(\hat{\mathbf{z}}) := \hat{\boldsymbol{Q}}_{\boldsymbol{B}}^\top \hat{\boldsymbol{\mu}}(\hat{\mathbf{z}}).$$

By local invertibility of the transition,

$$m = d + q.$$

Since $\hat{\boldsymbol{B}} \in \mathbb{R}^{m \times q}$ has full column rank, its left null space has dimension

$$m - q = d.$$

Hence

$$\hat{\boldsymbol{Q}}_{\boldsymbol{B}} \in \mathbb{R}^{m \times d},$$

and therefore

$$\boldsymbol{g}, \hat{\boldsymbol{g}} : \mathbb{R}^d \to \mathbb{R}^d.$$

We next show that both projected Jacobians have full rank d. By local invertibility of the true transition,

$$[\mathbf{J}_{\boldsymbol{\mu}}(\mathbf{z}) \quad \boldsymbol{B}]$$

is nonsingular. Suppose that

$$\hat{\boldsymbol{Q}}_{\boldsymbol{B}}^{\top} \boldsymbol{H} \mathbf{J}_{\boldsymbol{\mu}}(\mathbf{z}) \boldsymbol{a} = \mathbf{0}.$$

Then

$$\boldsymbol{H} \mathbf{J}_{\boldsymbol{\mu}}(\mathbf{z}) \boldsymbol{a} \in \ker(\hat{\boldsymbol{Q}}_{\boldsymbol{B}}^{\top}) = \mathrm{Im}(\hat{\boldsymbol{B}}) = \mathrm{Im}(\boldsymbol{H}\boldsymbol{B}).$$

Hence there exists $\boldsymbol{b}$ such that

$$\boldsymbol{H} \mathbf{J}_{\boldsymbol{\mu}}(\mathbf{z}) \boldsymbol{a} = \boldsymbol{H}\boldsymbol{B}\boldsymbol{b}.$$

Since $\boldsymbol{H}$ is invertible,

$$\mathbf{J}_{\boldsymbol{\mu}}(\mathbf{z}) \boldsymbol{a} = \boldsymbol{B}\boldsymbol{b}.$$

Therefore

$$[\mathbf{J}_{\boldsymbol{\mu}}(\mathbf{z}) \quad \boldsymbol{B}] \begin{bmatrix} \boldsymbol{a} \\ -\boldsymbol{b} \end{bmatrix} = \mathbf{0}.$$

The nonsingularity of the full transition Jacobian implies $\boldsymbol{a} = \mathbf{0}$. Thus,

$$\mathrm{rank} \left[\hat{\boldsymbol{Q}}_{\boldsymbol{B}}^{\top} \boldsymbol{H} \mathbf{J}_{\boldsymbol{\mu}}(\mathbf{z}) \right] = d. \tag{27}$$

The same argument applied to the learned transition gives

$$\mathrm{rank} \left[\hat{\boldsymbol{Q}}_{\boldsymbol{B}}^{\top} \mathbf{J}_{\hat{\boldsymbol{\mu}}}(\hat{\mathbf{z}}) \right] = d. \tag{28}$$

Hence both $\boldsymbol{g}$ and $\hat{\boldsymbol{g}}$ are local diffeomorphisms. By the inverse function theorem, on a sufficiently small neighborhood,

$$\hat{\mathbf{z}} = \boldsymbol{k}(\mathbf{z}) := \hat{\boldsymbol{g}}^{-1}(\boldsymbol{g}(\mathbf{z})), \tag{29}$$

where $\boldsymbol{k}$ is a local diffeomorphism.

Step 5: Conditional-mean alignment. Now that $\hat{\mathbf{z}} = \boldsymbol{k}(\mathbf{z})$ has been established, the sample-wise equality becomes

$$\boldsymbol{H} \left[\boldsymbol{\mu}(\mathbf{z}) + \boldsymbol{B}\epsilon \right] + \boldsymbol{c} = \hat{\boldsymbol{\mu}}(\boldsymbol{k}(\mathbf{z})) + \hat{\boldsymbol{B}}\hat{\epsilon}.$$

Taking conditional expectations given $\mathbf{z}$ yields

$$\hat{\boldsymbol{\mu}}(\boldsymbol{k}(\mathbf{z})) = \boldsymbol{H}\boldsymbol{\mu}(\mathbf{z}) + \boldsymbol{c}, \tag{30}$$

because

$$\mathbb{E}[\epsilon \mid \mathbf{z}] = \mathbf{0}, \qquad \mathbb{E}[\hat{\epsilon} \mid \mathbf{z}] = \mathbf{0}.$$

Since $\boldsymbol{\mu} = \boldsymbol{A}\boldsymbol{r}$ and $\boldsymbol{A}$ has full column rank, Assumption 2 implies

$$\mathrm{rank} \left[\mathbf{M}_{\boldsymbol{\mu}}(\mathbf{z}) \quad \mathbf{J}_{\boldsymbol{\mu}}(\mathbf{z}) \right] = 2d.$$

Moreover, both $\boldsymbol{\mu}$ and $\hat{\boldsymbol{\mu}}$ are additive. Lemma 1 therefore implies that $\boldsymbol{k}$ is component-wise up to permutation.

Step 6: Innovation coordinates. By equation 25, $\hat{\boldsymbol{B}}$ and $\boldsymbol{H}\boldsymbol{B}$ have the same column space. Since both are full column rank with q columns, there exists a unique matrix

$$\boldsymbol{R} \in \mathrm{GL}(q)$$

such that

$$\boldsymbol{H}\boldsymbol{B} = \hat{\boldsymbol{B}}\boldsymbol{R}. \tag{31}$$

Subtracting equation 30 from the sample-wise transition equality gives

$$\hat{\boldsymbol{B}}\hat{\epsilon} = \boldsymbol{H}\boldsymbol{B}\epsilon = \hat{\boldsymbol{B}}\boldsymbol{R}\epsilon.$$

Since $\hat{\boldsymbol{B}}$ has full column rank,

$$\hat{\epsilon} = \boldsymbol{R}\epsilon.$$

This proves the proposition. □

C Proof of Theorem 2

Proof. Since

$$\hat{\mathbf{z}}_{\mathcal{S}}^{\ell+1} = \boldsymbol{H}_{\mathcal{S}}\mathbf{z}_{\mathcal{S}}^{\ell+1} + \boldsymbol{c}_{\mathcal{S}},$$

affine equivariance of $\mathcal{G}$ gives

$$\mathcal{G}(p_{\hat{\mathbf{z}}_{\mathcal{S}}^{\ell+1}}) = \boldsymbol{H}_{\mathcal{S}}\mathcal{G}(p_{\mathbf{z}_{\mathcal{S}}^{\ell+1}}).$$

By the restricted innovation-subspace assumptions,

$$\mathcal{G}(p_{\mathbf{z}_{\mathcal{S}}^{\ell+1}}) = \mathrm{Im}(\boldsymbol{B}_{\mathcal{S}}^{\ell}), \qquad \mathcal{G}(p_{\hat{\mathbf{z}}_{\mathcal{S}}^{\ell+1}}) = \mathrm{Im}(\hat{\boldsymbol{B}}_{\mathcal{S}}^{\ell}),$$

and therefore

$$\mathrm{Im}(\hat{\boldsymbol{B}}_{\mathcal{S}}^{\ell}) = \mathrm{Im}(\boldsymbol{H}_{\mathcal{S}}\boldsymbol{B}_{\mathcal{S}}^{\ell}).$$

Hence

$$\hat{\boldsymbol{Q}}_{\boldsymbol{B},\mathcal{S}}^{\top}\boldsymbol{H}_{\mathcal{S}}\boldsymbol{B}_{\mathcal{S}}^{\ell} = \mathbf{0}.$$

Substituting the two restricted transition models into the affine relation and multiplying by $\hat{\boldsymbol{Q}}_{\boldsymbol{B},\mathcal{S}}^{\top}$ gives

$$\hat{\boldsymbol{Q}}_{\boldsymbol{B},\mathcal{S}}^{\top}\hat{\boldsymbol{\mu}}_{\mathcal{S}}^{\ell}(\hat{\mathbf{z}}^{\ell}) = \hat{\boldsymbol{Q}}_{\boldsymbol{B},\mathcal{S}}^{\top}\left[\boldsymbol{H}_{\mathcal{S}}\boldsymbol{\mu}_{\mathcal{S}}^{\ell}(\mathbf{z}^{\ell}) + \boldsymbol{c}_{\mathcal{S}}\right].$$

Define

$$\boldsymbol{g}(\mathbf{z}^{\ell}) := \hat{\boldsymbol{Q}}_{\boldsymbol{B},\mathcal{S}}^{\top}\left[\boldsymbol{H}_{\mathcal{S}}\boldsymbol{\mu}_{\mathcal{S}}^{\ell}(\mathbf{z}^{\ell}) + \boldsymbol{c}_{\mathcal{S}}\right],$$

and

$$\hat{\boldsymbol{g}}(\hat{\mathbf{z}}^{\ell}) := \hat{\boldsymbol{Q}}_{\boldsymbol{B},\mathcal{S}}^{\top}\hat{\boldsymbol{\mu}}_{\mathcal{S}}^{\ell}(\hat{\mathbf{z}}^{\ell}).$$

Then

$$\hat{\boldsymbol{g}}(\hat{\mathbf{z}}^{\ell}) = \boldsymbol{g}(\mathbf{z}^{\ell}).$$

By the projected-rank assumptions, the Jacobians of both $\boldsymbol{g}$ and $\hat{\boldsymbol{g}}$ have full column rank $d_{\mathbf{z}}^{\ell}$. Hence, by the constant-rank theorem, both maps are local embeddings. Restricting to sufficiently small neighborhoods, $\hat{\boldsymbol{g}}$ is locally invertible on its image. Using $\hat{\boldsymbol{g}}(\hat{\mathbf{z}}^{\ell}) = \boldsymbol{g}(\mathbf{z}^{\ell})$, we may define

$$\hat{\mathbf{z}}^{\ell} = \boldsymbol{k}(\mathbf{z}^{\ell}) := \hat{\boldsymbol{g}}^{-1}(\boldsymbol{g}(\mathbf{z}^{\ell})).$$

Differentiating

$$\hat{\boldsymbol{g}}(\boldsymbol{k}(\mathbf{z}^{\ell})) = \boldsymbol{g}(\mathbf{z}^{\ell})$$

gives

$$\mathbf{J}_{\boldsymbol{g}}(\mathbf{z}^{\ell}) = \mathbf{J}_{\hat{\boldsymbol{g}}}(\boldsymbol{k}(\mathbf{z}^{\ell}))\mathbf{J}_{\boldsymbol{k}}(\mathbf{z}^{\ell}).$$

Since $\mathbf{J}_{\boldsymbol{g}}$ has rank $d_{\mathbf{z}}^{\ell}$,

$$d_{\mathbf{z}}^{\ell} = \mathrm{rank}\,\mathbf{J}_{\boldsymbol{g}}(\mathbf{z}^{\ell}) \le \mathrm{rank}\,\mathbf{J}_{\boldsymbol{k}}(\mathbf{z}^{\ell}) \le d_{\mathbf{z}}^{\ell},$$

and hence $\mathbf{J}_{\boldsymbol{k}}$ is nonsingular. Thus $\boldsymbol{k}$ is a local diffeomorphism.

Conditioning the partial affine relation on $\mathbf{z}^{\ell}$ and using $\hat{\mathbf{z}}^{\ell} = \boldsymbol{k}(\mathbf{z}^{\ell})$ gives

$$\hat{\boldsymbol{\mu}}_{\mathcal{S}}^{\ell}(\boldsymbol{k}(\mathbf{z}^{\ell})) = \boldsymbol{H}_{\mathcal{S}}\boldsymbol{\mu}_{\mathcal{S}}^{\ell}(\mathbf{z}^{\ell}) + \boldsymbol{c}_{\mathcal{S}},$$

where we used

$$\mathbb{E}[\epsilon^{\ell} \mid \mathbf{z}^{\ell}] = \mathbf{0}, \qquad \mathbb{E}[\hat{\epsilon}^{\ell} \mid \mathbf{z}^{\ell}] = \mathbf{0}.$$

The first follows from $\epsilon^{\ell} \perp\!\!\!\perp \mathbf{z}^{\ell}$ and $\mathbb{E}[\epsilon^{\ell}] = \mathbf{0}$, and the second from Assumption 6.

Finally, by the restricted sufficient-nonlinearity assumption,

$$\mathrm{rank}\begin{bmatrix}\mathbf{M}_{\boldsymbol{\mu}_{\mathcal{S}}^{\ell}}(\mathbf{z}^{\ell}) & \mathbf{J}_{\boldsymbol{\mu}_{\mathcal{S}}^{\ell}}(\mathbf{z}^{\ell})\end{bmatrix} = 2d_{\mathbf{z}}^{\ell}.$$

Since both restricted mean maps are additive, Lemma 1 implies that $\boldsymbol{k}$ is component-wise up to permutation. □

D EXPERIMENTAL DETAILS

This appendix provides detailed descriptions of the experimental setup used in Section 5. The synthetic experiments are designed to examine the main structural predictions of hierarchical identifiability, including sufficient nonlinearity, overdetermination, and partial identifiability. Unless otherwise specified, results are averaged over ten random seeds. The implementation and scripts for reproducing the experiments are included in the supplementary material.

D.1 SYNTHETIC HIERARCHICAL DATA

The goal is to recover $\{\mathbf{z}^\ell\}_{\ell=0}^{L-1}$ and $\{\epsilon^\ell\}_{\ell=0}^{L-1}$ from $\mathbf{x} = \mathbf{z}^L$ alone. Unless otherwise specified, we generate $n = 2^{16}$ observational samples.

Default configuration. Unless otherwise specified, we use $L = 3$ mixing layers $(\mathbf{z}^0, \dots, \mathbf{z}^3 = \mathbf{x})$, with bottom-layer width $d_{\mathbf{z}}^0 = 10$, nonlinear expansion ratio

$$K_r = \frac{d_{\mathbf{r}}^\ell}{d_{\mathbf{z}}^\ell} = 2,$$

and overdetermination ratio

$$K_d = \frac{d_{\mathbf{z}}^{\ell+1}}{d_{\mathbf{r}}^\ell} = 1.5.$$

The latent widths therefore grow as

$$d_{\mathbf{z}}^\ell = d_{\mathbf{z}}^0 (K_r K_d)^\ell, \tag{32}$$

giving

$$(d_{\mathbf{z}}^0, d_{\mathbf{z}}^1, d_{\mathbf{z}}^2, d_{\mathbf{z}}^3) = (10, 30, 90, 270).$$

Throughout the experiments, whenever a latent, feature, or innovation dimension computed from a ratio is non-integer, we round it to the nearest integer, with a minimum value of one.

The local-invertibility condition in Theorem 1 would additionally require

$$d_\epsilon^\ell = d_{\mathbf{z}}^{\ell+1} - d_{\mathbf{z}}^\ell.$$

In finite-sample experiments, however, we found that the resulting high-dimensional innovation spaces substantially degraded optimization and representation recovery. We therefore use a smaller innovation dimension,

$$d_\epsilon^\ell = K_\epsilon d_{\mathbf{z}}^\ell, \qquad K_\epsilon = 0.3,$$

yielding

$$(d_\epsilon^0, d_\epsilon^1, d_\epsilon^2) = (3, 9, 27).$$

Thus, the synthetic experiments retain the hierarchical additive structure, use a construction designed to satisfy sufficient nonlinearity, and satisfy the overdetermination condition, but do not enforce the dimensional condition required for transition-level local invertibility. We use this relaxed setting to obtain stable finite-sample estimation while retaining the structural conditions examined in the main experiments.

Hierarchical architecture. Synthetic data follow the hierarchical generative structure introduced in Section 3. Each layer is generated as

$$\mathbf{z}^{\ell+1} = \boldsymbol{A}^\ell \boldsymbol{r}^\ell(\mathbf{z}^\ell) + \boldsymbol{B}^\ell \epsilon^\ell.$$

For numerical stability, we additionally standardize each generated next-layer coordinate using the empirical mean and standard deviation of the generated data before it is used as input to the subsequent layer. This normalization prevents subsequent nonlinearities from operating only in a near-linear range. This normalization is an experimental preprocessing step and is not part of the theoretical transition model.

Nonlinearities (Assumption 2). The default generator uses the additive feature map $\boldsymbol{r}^\ell$ defined in Section 3. Equivalently, each output feature is a sum of coordinate-wise nonlinear functions:

$$[\boldsymbol{r}^\ell(\mathbf{z}^\ell)]_i = \sum_{j=1}^{d_{\mathbf{z}}^\ell} u_{ij}^\ell \tanh\left(s_{ij}^\ell z_j^\ell + t_{ij}^\ell\right),$$

where

$$s_{ij}^\ell \sim \mathrm{Unif}[-1.5, 1.5], \qquad t_{ij}^\ell, u_{ij}^\ell \sim \mathrm{Unif}[-1, 1].$$

Linear mixing (Assumption 3). The joint matrix $[\boldsymbol{A}^\ell, \boldsymbol{B}^\ell]$ is drawn with i.i.d. $\mathrm{Unif}[-1, 1]$ entries. We accept only draws whose condition number lies below the threshold. We set the rejection threshold to the 1st percentile of the condition-number distribution obtained from 1000 independently drawn matrices of the same shape. Thus, only particularly well-conditioned draws are retained. The default $K_d = 1.5$ satisfies $d_{\mathbf{z}}^{\ell+1} > d_{\mathbf{r}}^\ell$, whereas $K_d \leq 1$ violates the overdetermination condition; the effect of this violation is examined in Fig. 2b.

Gaussian innovations. The root latent variable and innovations are sampled independently:

$$\mathbf{z}^0 \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}), \qquad \epsilon^\ell \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}),$$

with independence across all layers. We do not explicitly verify the exact Gaussian-flat subspace condition $\mathcal{G}(p_{\mathbf{z}^{\ell+1}}) = \mathrm{Im}(\boldsymbol{B}^\ell)$ in the finite-sample simulations.

Partially identifiable setting. To construct partially identifiable settings, we designate a trailing fraction of the latent coordinates as violating sufficient nonlinearity:

$$\mathbf{z}^\ell = [\mathbf{z}_{\mathrm{id}}^\ell, \mathbf{w}^\ell],$$

where $\mathbf{w}^\ell$ denotes coordinates for which sufficient nonlinearity is intentionally violated. For these coordinates, we use a single-feature element-wise map

$$r_{\mathbf{w}}(\mathbf{w}) = \tanh(\mathbf{w}),$$

for which the sufficient-nonlinearity rank condition cannot hold. The block $\mathbf{z}_{\mathrm{id}}^\ell$ retains the default general additive construction with output dimension $K_r \dim(\mathbf{z}_{\mathrm{id}}^\ell)$, whereas the $\mathbf{w}^\ell$ block is mapped element-wise using one feature per coordinate:

$$\boldsymbol{r}^\ell(\mathbf{z}^\ell) = \begin{bmatrix} \boldsymbol{r}_{\mathrm{id}}^\ell(\mathbf{z}_{\mathrm{id}}^\ell) \\ \tanh(\mathbf{w}^\ell) \end{bmatrix}.$$

In this experiment, the latent widths $d_{\mathbf{z}}^\ell$ are kept fixed at their default values; only the feature construction is modified as w_r varies:

$$d_{\mathbf{r}}^\ell = K_r \dim(\mathbf{z}_{\mathrm{id}}^\ell) + \dim(\mathbf{w}^\ell) = 2d_{\mathbf{z}}^\ell - \dim(\mathbf{w}^\ell),$$

where the final equality uses $K_r = 2$. Because the latent widths are fixed while $d_{\mathbf{r}}^\ell$ decreases with w_r, the overdetermination condition remains satisfied throughout this experiment. For $w_r > 0$, we set

$$\dim(\mathbf{w}^\ell) = \max\left(1, \lfloor w_r d_{\mathbf{z}}^\ell \rfloor\right),$$

while $w_r = 0$ denotes the setting with no coordinates intentionally violating sufficient nonlinearity, i.e., $\dim(\mathbf{w}^\ell) = 0$.

D.2 Models and optimization

Unless otherwise specified, the learned model uses the same dimensional hyperparameters as the data-generating model, i.e.,

$$\hat{d}_{\mathbf{z}}^0 = d_{\mathbf{z}}^0, \qquad \hat{K}_r = K_r, \qquad \hat{K}_d = K_d, \qquad \hat{K}_\epsilon = K_\epsilon.$$

Thus, the latent-state and innovation dimensions are matched between the true and learned models.

HIRL decoder. The decoder $\hat{\boldsymbol{f}}_\theta^{0:L-1}$ uses the same hierarchical structure:

$$\hat{\mathbf{z}}^{\ell+1} = \hat{\boldsymbol{A}}^\ell \hat{\boldsymbol{r}}^\ell(\hat{\mathbf{z}}^\ell) + \hat{\boldsymbol{B}}^\ell \hat{\epsilon}^\ell.$$

For each $i = 1, \dots, \hat{d}^{\ell}_{\mathbf{r}}$, the learned feature map is

$$[\hat{\boldsymbol{r}}^{\ell}(\hat{\mathbf{z}}^{\ell})]_i = \sum_{j=1}^{\hat{d}^{\ell}_{\mathbf{z}}} \sum_{k=1}^{K} \hat{u}^{\ell}_{ijk} \tanh\big(\hat{s}^{\ell}_{ijk} \hat{z}^{\ell}_{j} + \hat{t}^{\ell}_{ijk}\big) ,$$

where $\hat{u}^{\ell}_{ijk}$, $\hat{s}^{\ell}_{ijk}$, and $\hat{t}^{\ell}_{ijk}$ are trainable, and we use $K = 3$. The data-generating feature map with a single inner unit ($K = 1$) is therefore included as a special case.

HIRL encoder. The encoder $\hat{\boldsymbol{g}}_{\phi}$ is an unconstrained MLP with L hidden layers followed by a linear output layer. Each hidden layer has width $d^{L}_{\mathbf{z}}$ and uses maxout units with pool size 2. The output layer is linear with dimension

$$\hat{d}^{0}_{\mathbf{z}} + \sum_{\ell=0}^{L-1} \hat{d}^{\ell}_{\epsilon}.$$

The output is partitioned into

$$(\hat{\mathbf{z}}^{0}, \hat{\epsilon}^{0}, \dots, \hat{\epsilon}^{L-1})$$

and passed through the hierarchical decoder.

Reduced-feature ablation. To isolate the effect of sufficient nonlinearity, we use an ablation with the same encoder, hierarchical decoder structure, latent dimensions, and optimization settings as HIRL, but restrict the learned feature dimension to

$$\hat{d}^{\ell}_{\mathbf{r}} = \hat{d}^{\ell}_{\mathbf{z}}.$$

Consequently,

$$\operatorname{rank}\left[\mathbf{M}_{\hat{\boldsymbol{r}}^{\ell}}(\hat{\mathbf{z}}^{\ell}) \quad \mathbf{J}_{\hat{\boldsymbol{r}}^{\ell}}(\hat{\mathbf{z}}^{\ell})\right] \leq \hat{d}^{\ell}_{\mathbf{z}} < 2\hat{d}^{\ell}_{\mathbf{z}},$$

so the sufficient-nonlinearity condition is dimensionally impossible. This provides a controlled comparison in which the hierarchical architecture is retained while the sufficient-nonlinearity rank condition is made dimensionally impossible on the learned side.

Optimization. HIRL and the reduced-feature ablation are trained using mean squared reconstruction error. We use SGD with learning rate 0.1, momentum 0.9, and weight decay 10^{-5}. Mini-batches have size 512. Training runs for 10^{6} steps, with the learning rate reduced by a factor of 0.1 at step 8×10^{5}.

Baseline methods. We compare HIRL with one controlled ablation and two hierarchical representation-learning baselines.

- **HIRL (reduced feat.).** This is the reduced-feature ablation of HIRL described above.
- **Ladder VAE (LVAE) (Sønderby et al., 2016).** LVAE uses a hierarchical inference model combining bottom-up evidence with top-down generative information.
- **Variational Ladder Autoencoder (VLAE) (Zhao et al., 2017).** VLAE partitions the latent representation into multiple groups that are injected at different depths of the generative network.

See Section E for details.

LVAE and VLAE are chosen as representative hierarchical representation-learning baselines that are amenable to our controlled fully connected setting. More recent methods rely on substantially different convolutional architectures (Vahdat & Kautz, 2020) or temporal structure (Li et al., 2025), and are therefore not included in this controlled comparison.

Unlike HIRL and the reduced-feature ablation, LVAE and VLAE are not designed to match our identifiability assumptions or innovation decomposition. For all methods, we match the number of latent layers and corresponding decoder widths where applicable, while retaining each baseline's native variational objective and inference structure.

VLAE uses Adam with learning rate 2×10^{-4} and $\beta = (0.9, 0.999)$, together with KL warm-up over 2000 steps, following the original implementation. LVAE uses Adamax with learning rate 3×10^{-4} and free bits 0.5. For numerical stability, we additionally clip gradients to a global norm of 1.0. Both baselines use the same batch size and total number of optimization steps as HIRL.

Ablations. Unless a quantity is the sweep variable, all other settings are fixed at the default values above. We vary:

1. the sample size,
$$n = 2^t, \qquad t \in \{8, 10, 12, 14, 16, 18\};$$
2. the fraction of variables violating sufficient nonlinearity,
$$w_r \in \{0, 0.1, \ldots, 0.9\};$$
3. the overdetermination ratio,
$$K_d \in \{0.75, 1, 1.25, 1.5, 1.75, 2\};$$
4. the nonlinear expansion ratio,
$$K_r \in \{1, 1.5, 2, 2.5\};$$
5. the latent-dimension misspecification ratio,
$$K_{\text{est}} \in \{0.5, 0.75, 1, 1.25, 1.5\},$$
where the estimated latent and innovation dimensions are scaled by K_{est} relative to the generator dimensions, with the learned feature dimensions adjusted accordingly through the fixed expansion ratio; $\dim(\mathbf{x})$ is kept unchanged.

Computational resources. The experiments were run on a multi-node compute cluster. A typical HIRL run under the default synthetic configuration required approximately one to two hours on a node equipped with a 3.0 GHz 32-core CPU, 756 GB RAM, and one NVIDIA RTX PRO 6000 GPU.

D.3 Evaluation metrics

For intermediate latent variables $\mathbf{z}^\ell$, $\ell = 1, \ldots, L-1$, we compute the absolute Pearson correlation matrix between estimated and true coordinates, determine the optimal one-to-one matching using the Hungarian algorithm, and report the mean correlation over matched coordinates.

For partially identifiable settings, we separately report recovery of $\mathbf{z}^\ell_{\text{id}}$ and $\mathbf{w}^\ell$ after matching.

At the bottom layer, where information may be redistributed between the inferred state and innovation slots, we evaluate both state and innovation recovery from the joint inferred code $[\hat{\mathbf{z}}^0; \hat{\epsilon}^0]$. For $\mathbf{z}^0$, we compute absolute Pearson correlations between the coordinates of the joint code and the true root coordinates, and apply Hungarian matching. Because the theory allows a component-wise nonlinear ambiguity at the root, this Pearson-correlation metric is stricter than the theoretical identifiability criterion. For ϵ^0, we fit an optimal linear map from the joint code. Note that this joint evaluation is only for the bottom layer. For layers $\ell = 1, \ldots, L-1$, latent states and innovations are evaluated from their corresponding inferred slots. We additionally use optimal linear alignment as a diagnostic in the model-misspecification experiments.

Innovations are identifiable only up to an invertible linear transformation. For layers $\ell = 1, \ldots, L-1$, we fit
$$\epsilon^\ell \approx \boldsymbol{W}^\ell \hat{\epsilon}^\ell + \boldsymbol{b}^\ell$$
using ordinary least squares. At the bottom layer, the same linear alignment is instead fitted from the joint code $[\hat{\mathbf{z}}^0; \hat{\epsilon}^0]$, as described above. In both cases, we report the mean absolute correlation between the fitted and true innovation coordinates.

Overparameterized and underparameterized cases. When the estimated latent width is scaled by $K_{\text{est}} \neq 1$, we apply the Hungarian algorithm to the rectangular absolute-correlation matrix and report the mean absolute correlation over matched coordinate pairs. Unmatched estimated coordinates are ignored. For underparameterized models, the reported score therefore measures recovery among matched ground-truth coordinates rather than coverage of all true coordinates.

Table 3: Number of trainable parameters for the CIFAR-100 experiments. Encoder and decoder parameter counts are reported separately.

Method	Encoder	Decoder	Total
HIRL (ours)	10.5M	6.58M	17.1M
HIRL (reduced feat.)	10.5M	3.29M	13.8M
LVAE	31.5M	72.6M	104.1M
VLAE	23.4M	16.4M	39.8M

D.4 CIFAR-100 EXPERIMENTAL DETAILS

We evaluate unsupervised hierarchical representations on the CIFAR-100 dataset (Krizhevsky & Hinton, 2009). Training uses only the $50{,}000$ training images, and class labels are not used during representation learning. The $10{,}000$ test images are held out for evaluation with linear probes. Each image belongs to one of 100 classes.

Observation and preprocessing. The observation is the flattened 32×32 RGB image, $\mathbf{x} \in \mathbb{R}^{3072}$. Pixel values are first scaled to $[0, 1]$ and then standardized using the per-channel mean and standard deviation estimated from the training set. The same training-set statistics are applied to both the training and test splits. No random cropping, flipping, or other data augmentation is used. All models operate directly on the 3072-dimensional vectorized image and use fully connected architectures rather than convolutional networks.

HIRL Decoder. The decoder constructs the deterministic hierarchy

$$\hat{\mathbf{z}}^0 \longmapsto \hat{\mathbf{z}}^1 \longmapsto \hat{\mathbf{z}}^2 \longmapsto \hat{\mathbf{z}}^3 \longmapsto \hat{\mathbf{z}}^4 = \hat{\mathbf{x}},$$

using the same additive nonlinear mean-transition parameterization as in the synthetic experiments, but without explicit innovation variables:

$$\hat{\mathbf{z}}^{\ell+1} = \hat{\boldsymbol{A}}^{\ell} \hat{\boldsymbol{r}}^{\ell}(\hat{\mathbf{z}}^{\ell}).$$

For the latent transitions, we use the same dimensional rule as in equation 32, with $\hat{d}^0_{\mathbf{z}} = 16$, $K_r = 2$, and $K_d = 1.5$, yielding $\hat{d}^1_{\mathbf{z}} = 48$, $\hat{d}^2_{\mathbf{z}} = 144$, and $\hat{d}^3_{\mathbf{z}} = 432$. For the final transition to the observation layer, we instead fix the output dimension to the CIFAR-100 input dimension, $\hat{d}^4_{\mathbf{z}} = \dim(\mathbf{x}) = 3072$.

$\hat{\mathbf{z}}^0$	$\hat{\mathbf{z}}^1$	$\hat{\mathbf{z}}^2$	$\hat{\mathbf{z}}^3$	$\hat{\mathbf{z}}^4 = \hat{\mathbf{x}}$
16	48	144	432	3072

HIRL encoder. Unlike in the synthetic experiments, HIRL uses a deterministic encoder that outputs only the root representation $\hat{\mathbf{z}}^0$; no explicit innovation variables are inferred. The encoder node widths are

$$[3072,\ g_{\mathrm{h}},\ g_{\mathrm{h}},\ g_{\mathrm{h}},\ 16]\,,$$

where $g_{\mathrm{h}} = 1024$.

Baselines. We use LVAE and VLAE as baselines, as in the synthetic experiments. The corresponding latent representation widths are matched across HIRL, VLAE, and LVAE where applicable. See Table 3 for a comparison of the parameter counts across methods.

We do not include NVAE (Vahdat & Kautz, 2020) in the controlled CIFAR-100 comparison because its architecture is specifically designed around deep convolutional image modeling, whereas our comparison uses fully connected models with matched latent widths. Including the original NVAE would therefore confound differences in the hierarchical representation-learning structure with substantially different architectural inductive biases.

Optimization. HIRL uses the same optimization scheme as in the synthetic experiments, except that the mini-batch size is set to 256. VLAE and LVAE retain their baseline-specific optimization settings described in Section E, while using the same total number of training steps and mini-batch size. Each random seed determines both the parameter initialization and the training mini-batch sequence.

Linear probes. Class labels from the training split are used only for fitting the linear probes. For each trained model, we encode the complete CIFAR-100 training and test sets. Linear probes are independently fitted to the representations $\hat{\mathbf{z}}^0$, $\hat{\mathbf{z}}^1$, $\hat{\mathbf{z}}^2$, and $\hat{\mathbf{z}}^3$ for a 100-way image-classification task.

Each probe is an ℓ_2-regularized multinomial logistic-regression model implemented with `scikit-learn`, using $C = 1$, the `lbfgs` solver, and at most 1000 iterations. Before fitting each probe, a `StandardScaler` is fitted to the training representations and then applied unchanged to the corresponding test representations. The chance accuracy is 1%.

E BASELINES

E.1 VLAE

We use the Variational Ladder Autoencoder (VLAE) of Zhao et al. (2017) as a hierarchical representation-learning baseline. VLAE differs from a conventional hierarchical VAE in that its latent variables do not form a stochastic Markov hierarchy. Instead, the latent code is partitioned into ladder components, each of which is inferred from deterministic features at a different depth of the encoder.

Let $\boldsymbol{h}^\ell$ denote the deterministic encoder representation at depth ℓ. Each ladder component is inferred as

$$q(\boldsymbol{z}^\ell \mid \boldsymbol{h}^\ell) = \mathcal{N}\big(\boldsymbol{\mu}^\ell, \operatorname{diag}\big((\boldsymbol{\sigma}^\ell)^2\big)\big)\,,$$

with a factorized standard Gaussian prior over the full latent code.

The encoder propagates a deterministic hidden representation upward while reading out a separate latent component at each depth. The latent component closest to the observation is inferred from the shallowest ladder representation, whereas the most abstract latent component is inferred from the deepest deterministic features.

The decoder follows the VLAE top-down construction. Starting from the most abstract latent component, it propagates a single hidden state toward the observation. At each subsequent level, the corresponding latent component is projected into the decoder hidden space, combined with the incoming top-down hidden state, and passed through the next generative MLP. Using our HIRL-consistent layer ordering, which reverses the original VLAE ladder ordering, we write this construction as

$$\tilde{\mathbf{z}}^0 = \boldsymbol{f}^0(\hat{\mathbf{z}}^0),$$

and recursively as

$$\tilde{\mathbf{z}}^{\ell+1} = \boldsymbol{f}^{\ell+1}\big(\tilde{\mathbf{z}}^\ell, \hat{\mathbf{z}}^{\ell+1}\big)\,, \qquad \ell = 0, \ldots, L-2,$$

where $\tilde{\mathbf{z}}^\ell$ denotes the deterministic decoder hidden state. Following the construction of Zhao et al. (2017), each recursive stage is implemented as

$$\tilde{\mathbf{z}}^{\ell+1} = \boldsymbol{u}^{\ell+1}\big(\big[\tilde{\mathbf{z}}^\ell; \boldsymbol{v}^{\ell+1}(\hat{\mathbf{z}}^{\ell+1})\big]\big)\,,$$

where $[\cdot;\cdot]$ denotes concatenation. Thus, the decoder does not contain separate parallel branches; instead, latent components are sequentially injected into a single top-down generative stream.

We use a fixed-variance Gaussian observation model, so that the reconstruction term is proportional to mean squared error. This is an adaptation to the controlled fully connected comparison used in our experiments.

For the controlled comparison, we adapt VLAE to the fully connected setting and use the same hidden width as HIRL for both the encoder and decoder MLPs.

We initialize the VLAE optimization settings from the original implementation and adapt them to our controlled training protocol. Optimization uses Adam with learning rate 2×10^{-4}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and weight decay 10^{-5}. For the controlled comparison, we use the same 10^6 optimization steps and batch size as HIRL in each experiment, with the learning rate reduced by a factor of 10 at step 8×10^5. The KL coefficient is gradually increased according to $\beta_t = 1 - \exp\big(-\frac{t+1}{2000}\big)$.

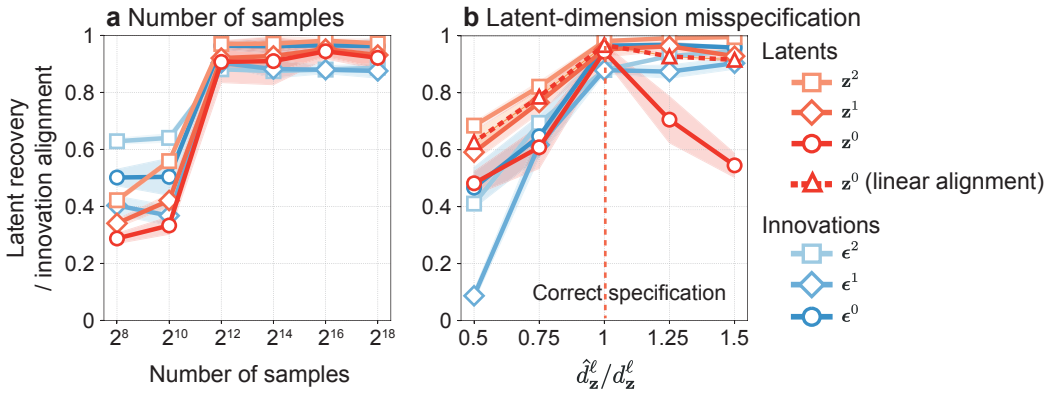


Figure 3: Finite-sample behavior and robustness to latent-dimension misspecification. **(a)** Latent recovery and innovation alignment as a function of the number of training samples. **(b)** Performance under misspecification of the latent dimensionality, where $\hat{d}^{\ell}_{\mathbf{z}}/d^{\ell}_{\mathbf{z}} = 1$ denotes the correctly specified model. Solid latent curves report mean absolute correlation after optimal permutation matching using the Hungarian algorithm, while the dashed $\mathbf{z}^0$ curve reports the corresponding score after optimal linear alignment. Innovation variables are evaluated using linear alignment. Shaded regions indicate variability across random seeds.

E.2 LVAE

We use a fully connected adaptation of the Ladder Variational Autoencoder (LVAE) of Sønderby et al. (2016), based primarily on the public implementation. The LVAE introduces a hierarchical inference model that recursively combines bottom-up approximate-likelihood information with top-down generative information.

The inference network consists of a deterministic bottom-up path with ResidualMLP blocks. At each lower stochastic layer, the bottom-up deterministic features are combined with the corresponding top-down prior information and transformed to parameterize the approximate posterior. At the top stochastic layer, the posterior is parameterized from the deepest bottom-up features. In the public implementation used in our experiments, the top-level prior is learned separately, and the deterministic path uses batch normalization, ELU activations, and gated residual connections, with two residual blocks per stochastic layer. We also retain the concatenation-based residual merging and generative skip connections of the public implementation.

For our controlled comparison, we replace the original two-dimensional convolutional operations with fully connected linear layers and match the stochastic latent widths to those used by HIRL. In our fully connected adaptation, the latent distributions are diagonal Gaussians, and we use a diagonal Gaussian observation model for the controlled comparison. The latent Gaussian parameters are represented by a mean and log variance.

In our implementation, following the public codebase, we use free bits rather than the deterministic KL warm-up used in the original LVAE paper, with 0.5 nats per stochastic layer. Optimization uses Adamax with learning rate 3×10^{-4}. For the controlled comparison, we additionally use weight decay 10^{-5}, the same batch size, total number of optimization steps, and learning-rate schedule as in the VLAE experiments. We clip gradients to an ℓ_2 norm of 1 for numerical stability.

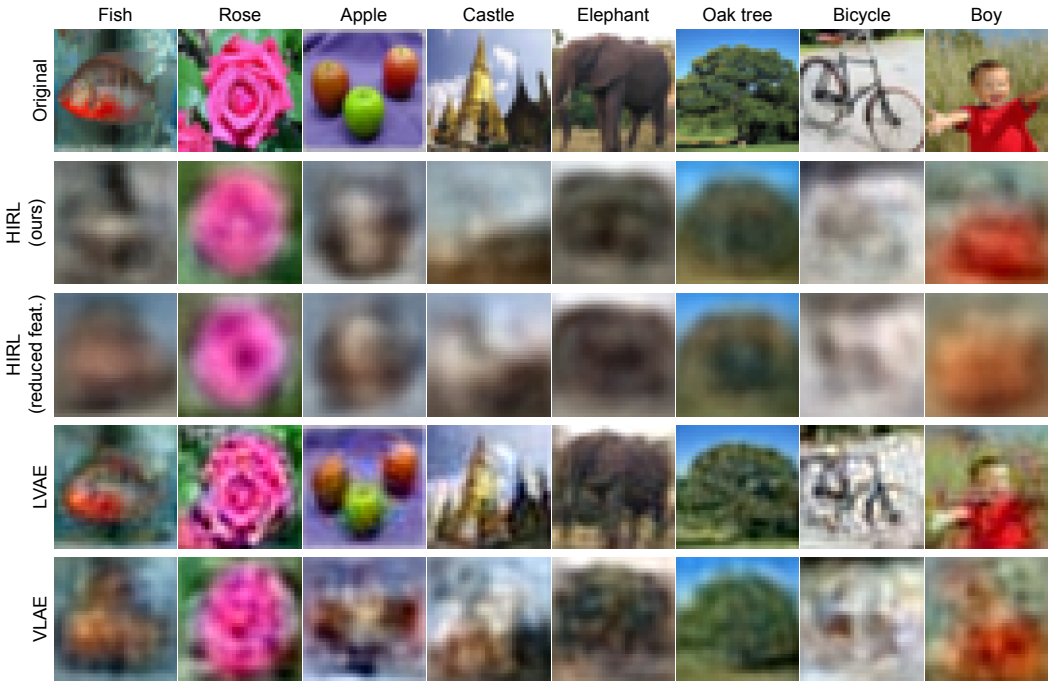


Figure 4: Qualitative reconstruction comparison on CIFAR-100. Each column shows the same test image reconstructed by HIRL, the reduced-feature HIRL variant, LVAE, and VLAE. Test images are randomly selected using a fixed random seed. The qualitative differences are consistent with the test reconstruction errors reported in Table 2. Together with the linear-probe results, they illustrate that lower pixel-level reconstruction error does not necessarily imply stronger semantic representations at intermediate latent layers.