PREDICTIVE SELF-SUPERVISED LEARNING PROVABLY IDENTIFIES STOCHASTIC SIGNALS UNDER NUISANCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Self-supervised learning (SSL) by predicting in latent space, without generating the input data itself, learns highly abstract, useful representations. Intuitively, this success is often attributed to its ability to discard nuisance information that is irrelevant to prediction. However, this poses a conundrum: both stochastic variation in a prediction-relevant latent signal and true nuisance make observations partly unpredictable; how could they be distinguished? Surprisingly, we prove that common SSL methods can achieve exactly this, by implicitly instantiating a latent-variable model with stochastic dynamics and observation-private nuisance. We trace their ability to recover the stochastic signal to two complementary principles: Predictive mutual information maximization ensures that representations retain the information needed for prediction, while latent distribution matching constrains how this information is encoded, thereby making the retained signal identifiable. We confirm this identifiability result in simulations for Gaussian predictors, which recover the true signal up to an affine transformation even in dynamic, nuisance-laden environments.

1 INTRODUCTION

Self-supervised learning (SSL) has become a central paradigm for representation learning, enabling models to extract useful structure from unlabeled data across vision, language, and audio (Noroozi & Favaro, 2016; Oord et al., 2018; Dawid & LeCun, 2024; Gui et al., 2024). Its success is often attributed to a simple intuition: different observations in the same context share a context-specific signal while also containing observation-specific variation that should not be represented. For example, image augmentations alter image details without changing object identity, different modalities can provide different views of the same signal, and temporal observations combine an evolving signal with non-predictive nuisance. By learning what is shared or predictable across observations, non-generative modeling is thought to retain the underlying *signal* while *nuisance* variables are not represented but left implicit in the generative model.

This nuisance-based explanation is pervasive, but it is rarely made explicit in statistical models. For example, non-generative modeling has been derived from the idea that representations should preserve the mutual information (MI) shared between observations, while unpredictable pixel-level nuisance might be ignored (Oord et al., 2018; Tian et al., 2020; Shwartz-Ziv et al., 2023; Rodríguez Gálvez et al., 2023; Shwartz Ziv & LeCun, 2024). Although these information-theoretic approaches provide an intuitive view of how nuisance is treated in non-generative models, they do not formally specify what exactly constitutes nuisance, and how the remaining signal is represented.

One route from the intuitive view to formal insights is given by identifiability theory. Results in nonlinear independent component analysis (ICA) and non-generative modeling show that latent variables can be recovered up to restricted transformations when the learned and true conditional distributions have suitable structure (Hyvarinen et al., 2019). This analysis has been extended to prove that non-generative models can identify the signal even under the presence of nuisance (von Kügelgen et al., 2021). These results require that the model is provided with different views of the same data point, where nuisance is private to each view, and the signal variable is *exactly* the same for each view. However, this does not cover the more general setting where the signal can evolve stochastically itself. This raises the question: is it possible to distinguish noise in otherwise predictable dynamics from completely unpredictable nuisance?

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.

Here we formalize implicit nuisance in stochastic latent variable models and study when non-generative modeling recovers the nuisance-free signal. Our main contributions are:

Disentangling the roles of MI maximization and distribution matching. We show that in non-generative SSL, predictive MI maximization and latent distribution matching (LDM) play distinct roles: MI maximization guarantees that representations retain all predictable information, while LDM alone is responsible for identification of the retained signal. This resolves prior ambiguity about the function of the terms in non-generative SSL objectives (Tschannen et al., 2020; Mikulasch & Zenke, 2026), showing that MI maximization and LDM are not sufficient on their own, but complement each other.

Identifiability under stochastic, private nuisance. We introduce a latent-variable model in which nuisance is private to each observation given a stochastically evolving signal, relaxing the deterministic shared-context assumption of prior identifiability results (von Kügelgen et al., 2021; Daunhawer et al., 2023). For exponential-family predictive models we prove that exact LDM together with predictive-information saturation recovers the true sufficient statistic as an affine readout of the learned statistic.

Affine recovery for Gaussian predictors. We specialize our general result to the practically relevant case of Gaussian predictive models, showing that the signal is recovered as an affine transformation of the learned representation whenever the latent dimensionality is large enough. We confirm this in simulations with stochastic latent dynamics and unpredictable nuisance, and in a stochastic variant of the Causal3DIdent benchmark, demonstrating recovery of causally coupled signal factors from high-dimensional images.

2 Related work

Theory of non-generative latent models. There has been significant interest in understanding non-generative modeling through theoretical analysis (e.g., Saunshi et al., 2019; Wang & Isola, 2020; Ben-Shaul et al., 2023). A line of work closely related to this article aims to connect contrastive methods (e.g., CPC) to latent variable models (von Kügelgen et al., 2021; Zimmermann et al., 2021; Aitchison & Ganev, 2024; Nakamura et al., 2023; Bizeul et al., 2024; Mikulasch & Zenke, 2026). Others proposed information-theoretic interpretations of regularization-based methods (e.g., VICReg, Shwartz-Ziv et al., 2023) and clustering/contrastive methods (Rodríguez Gálvez et al., 2023).

Identifiable latent variable models without nuisance. Identifiability is a core concept in linear ICA Hyvärinen & Oja (2000), which guarantees the recovery of "true" underlying variables up to trivial transformations. The concept was later extended to nonlinear ICA using assumptions such as temporal structure or auxiliary variables (Sprekeler et al., 2014; Khemakhem et al., 2020; Hyvarinen et al., 2019; Roeder et al., 2021). More recently, these insights also enabled proving identifiability for non-generative models (Zimmermann et al., 2021; Laiz et al., 2025; Mikulasch & Zenke, 2026). While these models study identification under stochastic dynamics, they do not include nuisance in their generative model.

Identifiable latent variable models with nuisance. A related line of work studies identifiability in the presence of view-specific nuisances. Gresele et al. (2020) showed that shared sources can be identified from multiple nonlinear views despite view-specific corruptions. von Kügelgen et al. (2021) and Lyu et al. (2022) extended this approach to show that common SSL methods might be used to extract the shared components between views. Daunhawer et al. (2023) transported these insights to multimodal settings, proving recovery of factors shared across modalities in the presence of modality-specific latent variation. These works establish that identifiability is possible when nuisance variables are present, but require deterministic relationships of the recovered variables between views.

3 Theory

Before we develop our theory of predictive SSL under nuisance we briefly review the concepts of predictive MI maximization and latent distribution matching that lie at the core of our work.

The specific loss function we are going to analyze has been used in several previous works. Initially, Oord et al. (2018) proposed InfoNCE, to maximize predictive MI $I[z; z_c]$ between representations z and predictive variables z_c. Later work has emphasized that the objective that is actually maximized

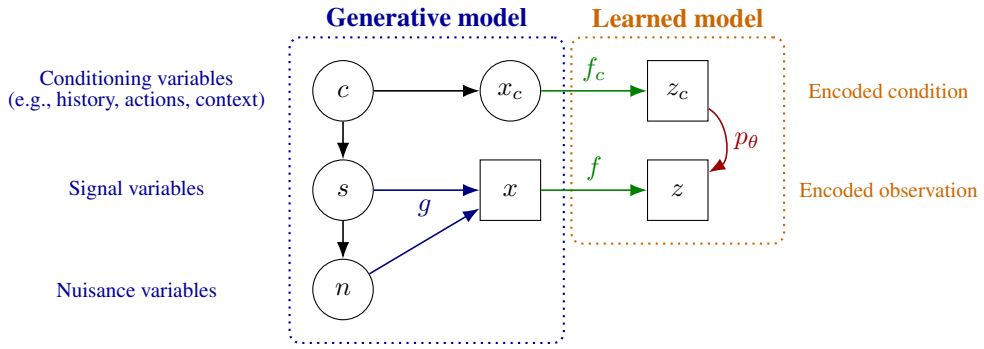


Figure 1: Graphical model for observation-private nuisance. The three layers show (from left to right) unobserved latent variables, observations, and learned representations. Square nodes denote deterministically defined variables.

by many of the common SSL algorithms, including InfoNCE (Oord et al., 2018), SimCLR (Chen et al., 2020), and VICReg (Shwartz-Ziv et al., 2023), has the general form (Aitchison & Ganev, 2024; Mikulasch & Zenke, 2026)

$$\begin{aligned}\mathcal{F}(f, f_c, \theta) =& - D_{\mathrm{KL}}[q_f(z, z_c) \| p_\theta(z, z_c)] + I_{q_f}[z; z_c] \\ =& \langle \log p_\theta(z, z_c) \rangle_{q_f(z, z_c)} + H_{q_f}[z] + H_{q_f}[z_c] \, .\end{aligned} \tag{1}$$

Here $q_f(z, z_c)$ is the pushforward distribution of observations $p(x, x_c)$ through the observation encoder $z = f(x)$ and condition encoder $z_c = f_c(x_c)$, and $p_\theta(z, z_c)$ is a latent model that is learned alongside. Different entropy estimators lead to different SSL algorithms—for example, KDE and log-determinant of covariance estimators can be related to SimCLR (Chen et al., 2020) and VICReg (Shwartz-Ziv et al., 2023), respectively (Mikulasch & Zenke, 2026). By this goal function, representations are not only required to be mutually informative, they are also required to follow a simple learned latent model $p_\theta(z, z_c)$ via LDM. In the following we will therefore refer to Equation 1 as the InfoLDM goal function.

Commonly, but not necessarily, this goal function is simplified by ignoring the marginal of the conditioning variable and setting $p_\theta(z_c) = q_f(z_c)$, leading to

$$\mathcal{F}(f, f_c, \theta) = \langle \log p_\theta(z \mid z_c) \rangle_{q_f(z, z_c)} + H_{q_f}[z] \, . \tag{2}$$

This is a well-known lower bound on the MI between latent variables (Poole et al., 2019), but we here analyze it with the clear framing that maximizing it performs both MI maximization and LDM.

3.1 Exponential family statistical recovery

The main theorem specifies when maximizing the InfoLDM goal function (Equation 1) recovers the signal s in the learned representation z up to simple transformations. For this we make the following assumptions, which are specified more rigorously in Appendix A.1.1.

Generative model. Let $s \in \mathcal{S} \subseteq \mathbb{R}^{d_s}$ be a semantic *signal* variable, $n \in \mathbb{R}^{d_n}$ a noisy *nuisance* variable, and let c denote the true *condition* (or cause, context). Let $x = g(s, n)$ be the observation, and x_c the observed condition. As before, let $z = f(x) \in \mathcal{Z} \subseteq \mathbb{R}^{d_z}$ be the learned representation, and $z_c = f_c(x_c) \in \mathcal{Z}_c \subseteq \mathbb{R}^{d_{z_c}}$ the learned encoded condition. Finally, let $q_f(z, z_c)$ be the pushforward distribution of observations, and $p_\theta(z, z_c)$ the learned model (Figure 1).

Assumption 1. The nuisance is private to the current observation in the sense that

$$n \perp\!\!\!\perp c \mid s. \tag{3}$$

Assumption 2. The true predictive distribution is exponential family

$$p(s \mid c) = h_s(s) \exp\big(\alpha(c)^\top \tau_\star(s) - \psi(\alpha(c))\big) \, , \tag{4}$$

where $\tau_\star : \mathcal{S} \to \mathbb{R}^{r_s}$ is a sufficient statistic, h_s a fixed carrier function, and $\alpha(c)$ encodes the dependency on the condition c. Similarly, the learned latent model is exponential family

$$p_\theta(z \mid z_c) = h_z(z) \exp\big(\beta(z_c)^\top \tau_z(z) - \phi(\beta(z_c))\big), \tag{5}$$

where the predictor $\beta(z_c)$ is learned, and $\tau_z : \mathcal{Z} \to \mathbb{R}^{r_z}$.

Assumption 3. By optimizing the InfoLDM goal function Equation 1 via $q_f(z, z_c)$ and $p_\theta(z, z_c)$ it is possible to reach the global optimum where

$$D_{\mathrm{KL}}[q_f(z, z_c) \| p_\theta(z, z_c)] = 0, \qquad I_{q_f}[z; z_c] = I[(s, n); c] = I[s; c] < \infty. \tag{6}$$

Assumption 4. It is possible to choose $c_0, c_1, \ldots, c_{r_s}$ from the support of $p(c)$ such that

$$D := \begin{bmatrix} (\alpha(c_1) - \alpha(c_0))^\top \\ \vdots \\ (\alpha(c_{r_s}) - \alpha(c_0))^\top \end{bmatrix} \in \mathbb{R}^{r_s \times r_s} \tag{7}$$

is invertible.

Assumption 5. For every condition c the random vector $\tau_\star(s)$ with $s \sim p(s \mid c)$ is not contained in a proper affine hyperplane of $\mathbb{R}^{r_s}$.

Most of the presented assumptions are standard in the context of nonlinear identification theory. Assumption 1 can be understood as a definition of nuisance factors, similar to the definition by von Kügelgen et al. (2021). Assumptions 2, 4 and 5 are standard assumptions in the theory of exponential family identification (e.g., Khemakhem et al., 2020). Chiefly, 4 and 5 ensure that all latent dimensions of the probabilistic model, not only a subset, are well constrained.

Assumption 3 is the strongest. It implicitly replaces common invertibility assumptions on the generative model. To see this relation consider that the generative model (Figure 1) together with the data processing inequality imply $I_{q_f}[z; z_c] \leq I[x; x_c] \leq I[s; c] = I[(s, n); c]$. At the optimum these inequalities become equalities, and in particular information has to be preserved in the observations $I[x; x_c] = I[s; c]$, restricting $g(s, n)$ and $p(x_c \mid c)$ to be information-preserving. Additionally, this restriction implicitly requires that the latent representation is sufficiently high-dimensional to capture the required information, as will be demonstrated by Theorem 1.

With these assumptions in place we are ready to state our main result.

Theorem 1 (Exponential-family statistical recovery through InfoLDM)**.** *Suppose Assumptions 1, 2, 3, 4, and 5 hold, and the InfoLDM goal function (Equation 1) is fully maximized.*

Then the dimension of the model sufficient statistic $r_z \geq r_s$, and there is a full-row-rank matrix $A \in \mathbb{R}^{r_s \times r_z}$ and a vector $b \in \mathbb{R}^{r_s}$ such that

$$\boxed{\tau_\star(s) = A\tau_z(z) + b} \qquad \textit{almost surely}. \tag{8}$$

If $\tau_\star$ is injective, then s is almost surely a measurable function of z. If $r_z = r_s$, then A is invertible and the two sufficient statistics are related by an invertible affine transformation.

The full proof is given in Appendix A.1.2. In broad strokes, the first step is to use successful MI maximization to show that z_c retains all information about s from c, and similarly z about c. This can be leveraged to directly relate the true and learned conditional probability distributions of s and z. From there the result follows via well established identifiability theory approaches for exponential family models (e.g., Khemakhem et al., 2020).

3.2 Specialization to Gaussian prediction

Theorem 1 is fairly general; here we aim to understand the behavior of commonly used non-generative models, which rely on Gaussianity assumptions. The goal function of many of these can be understood as approximately maximizing

$$\mathcal{F} \propto \left\langle -\frac{1}{2\sigma^2} \parallel z - \mu_\theta(z_c) \parallel^2 \right\rangle_{q_f(z, z_c)} + H_{q_f}[z], \tag{9}$$

where μ_θ is a learned prediction function (cf. von Kügelgen et al., 2021; Mikulasch & Zenke, 2026). This expression follows directly from the simplified goal function (Equation 2) by choosing p_θ to be a Gaussian with fixed variance $\Sigma_\theta = \sigma^2 \mathbb{1}$ (Shwartz-Ziv et al., 2023; Mikulasch & Zenke, 2026). Further asserting

$$p(s \mid c) = \mathcal{N}(s; \mu_\star(c), \Sigma_\star), \qquad p_\theta(z \mid z_c) = \mathcal{N}(z; \mu_\theta(z_c), \Sigma_\theta), \qquad \Sigma_\star \succ 0, \quad \Sigma_\theta \succ 0, \tag{10}$$

Theorem 1 can be specialized to give the stronger result of affine recovery of s.

Corollary 1 (Gaussian affine recovery)**.** *Under the assumptions of Theorem 1 and Equation 10, necessarily $d_z \geq d_s$, and there is a full-row-rank matrix $A \in \mathbb{R}^{d_s \times d_z}$ and $b \in \mathbb{R}^{d_s}$ such that*

$$s = Az + b \qquad \textit{almost surely}. \tag{11}$$

If $d_z = d_s$, then A is invertible.

This result follows immediately from Theorem 1 by the fact that the statistic of the fixed variance Gaussian distribution is the identity $\tau_\star(s) = s$ and $\tau_z(z) = z$, and $d_s = r_s$, $d_z = r_z$.

Intuitively, affine recovery fully relies on noise in the dynamics, as described in previous analyses (Zimmermann et al., 2021; Laiz et al., 2025; Mikulasch & Zenke, 2026): Noise in the true latent variables has Gaussian structure, and forcing learned variables to be Gaussian locally straightens latent coordinates around the predicted mean, which globally leads to an affine relation between true and recovered variables. Information sufficiency on the other hand supplies the model with a guarantee that all predictable signal variables are recovered. Note that nothing forces the model to discard nuisance variables entirely, except if the latent dimensionality of the model is restricted to match the dimensionality of the true signal. Similar results might be obtained for other exponential family distributions, such as von Mises-Fisher for variables on the sphere.

4 Simulation experiments

To numerically verify our proof, we performed a range of control experiments. First, we followed a common recovery experimental setup (e.g., Zimmermann et al., 2021). Specifically, we generated length-five sequences with a signal following

$$s_{t+1} = \rho R s_t + \sqrt{1-\rho^2}\,\epsilon_t, \qquad \epsilon_t \sim \mathcal{N}(0, I),$$

where $\rho = 0.9$ and R is a fixed random orthogonal matrix. At every step, we independently sampled a 15-dimensional Gaussian nuisance variable and map the combined signal and nuisance through a fixed injective 3-layer multi-layer perceptron (MLP) with leaky ReLU activation to obtain a 100-dimensional observation. We varied the signal dimension d_S from 5 to 20.

To recover the signal variables we trained a 5-layer MLP mapping the input to a latent space with same dimension as the signal, $d_Z = d_S$. Predictions were generated through an long short-term memory (LSTM) taking all previous timesteps and a linear prediction head. Training maximized the InfoLDM loss for Gaussian predictors (Equation 9), where the entropy was estimated through either kNN, KDE, or log-determinant of covariance (logdet) entropy estimators, as specified in Mikulasch & Zenke (2026). Simulations showed good affine signal recovery up to ten-dimensional latents, except for KDE entropy estimation, which showed slightly worse performance in this scenario (Table 1).

Table 1: R^2 of the learned representations with the true factors of variation for the numerical test, averaged over 5 runs plus minus standard deviations. We highlight well learned factors ($R^2 > 0.9$).

Entropy estimator	$d_s = 5$	$d_s = 10$	$d_s = 15$	$d_s = 20$
KDE	$\mathbf{0.92 \pm 0.16}$	0.81 ± 0.04	0.60 ± 0.08	0.48 ± 0.08
kNN	$\mathbf{0.98 \pm 0.00}$	$\mathbf{0.95 \pm 0.04}$	0.76 ± 0.01	0.64 ± 0.02
logdet	$\mathbf{0.99 \pm 0.00}$	$\mathbf{0.95 \pm 0.04}$	0.78 ± 0.02	0.68 ± 0.01

4.1 Identifying stochastic causal relations between images

As second example we tested signal recovery in high-dimensional images with strong, structured signal stochasticity, and structured, but independent nuisance. For this purpose we constructed image

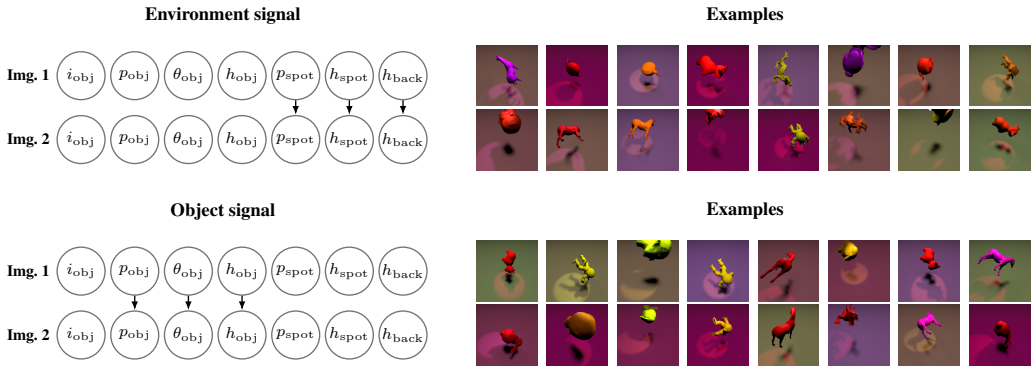


Figure 2: Causal3DIdent image-pairs. The left diagrams show the true latent factors which are: object identity (i_{obj}), position ($p_{\text{obj-}x}$, etc., here displayed grouped), rotation ($\theta_{\text{obj-}\alpha}$, etc.) and hue ($h_{\text{obj}}$), spotlight position ($p_{\text{spot}}$), hue ($h_{\text{spot}}$), and background hue ($h_{\text{back}}$). Arrows mark the factors causally coupled across image pairs. Right panels show example rendered image pairs. Even though factors between images are causally connected, there still exists significant variation due to noisy coupling.

pairs in which a designated subset of generative factors was causally related. We based this on the Causal3DIdent dataset (von Kügelgen et al., 2021), which provides 11 image factors (Figure 2). After mapping these factors to standard-normal coordinates, we sampled $s_2 = \rho s_1 + \sqrt{1-\rho^2}\,\epsilon$ with $\epsilon \sim \mathcal{N}(0, I)$ and $\rho = 0.9$. Since Causal3DIdent provides a finite set of rendered images, we selected the second image uniformly from the $k = 16$ nearest neighbors of s_2. All remaining factors, including object class, are thus random and determined by the selected image. We considered two settings: an *object* signal comprising object position, rotation, and hue, and an *environment* signal comprising spotlight position, spotlight hue, and background hue (Figure 2).

For signal recovery we used the same architecture and losses as in the numerical examples (Section 4), except for switching to a ResNet-18 as the image encoder. To give models the option to encode also nuisance we chose the latent dimension to be 10, similar to the total number of signal and nuisance factors. The simulations showed that learning consistent resulted in affine recovery of most of the causal factors, while performance depended on the entropy estimator, with KDE performing best in this scenario (Table 2). We further found that the entropy estimator also influenced whether or not nuisance variables were retained in the latent representation. Especially object position tended to be partly retained, which is sensible given the convolutional encoder; but also other variables were partly retained, an effect that became even more pronounced when computing nonlinear R^2 (Appendix, Table 3).

Table 2: R^2 of the learned representations with the true factors of variation for the stochastic Causal3DIdent dataset, averaged over 5 runs. Standard deviations are not displayed here but are mostly small (< 0.02) except for unidentified signal / identified nuisance variables, indicating intermittent optimization failures. We highlight well learned factors ($R^2 > 0.9$) in bold and signal factors with grey background.

Scenario	Entr. est.	$p_{\text{obj-}x}$	$p_{\text{obj-}y}$	$p_{\text{obj-}z}$	$\theta_{\text{obj-}\alpha}$	$\theta_{\text{obj-}\beta}$	$\theta_{\text{obj-}\gamma}$	h_{obj}	p_{spot}	h_{spot}	h_{back}
Env. signal	KDE	0.20	0.47	0.37	0.05	0.08	0.12	0.41	**0.97**	**0.95**	**0.97**
Env. signal	kNN	0.05	0.55	0.39	0.08	0.05	0.07	0.11	**0.97**	**0.95**	**0.97**
Env. signal	logdet	−0.01	−0.02	−0.00	−0.01	−0.00	−0.01	−0.01	**0.98**	0.76	**0.99**
Object signal	KDE	**0.98**	**0.98**	**0.97**	**0.95**	**0.96**	**0.94**	**0.92**	−0.01	0.01	0.13
Object signal	kNN	**0.97**	**0.96**	**0.96**	0.87	**0.92**	0.26	**0.91**	−0.01	−0.00	0.56
Object signal	logdet	**0.98**	**0.97**	**0.97**	**0.92**	**0.95**	**0.91**	0.87	0.15	−0.00	−0.00

4.2 Latent world model with actions and nuisance

In a last experiment we tested recovery and nuisance removal in a physical system with controlled actions. To that end we extended the MuJoCo suite (Todorov et al., 2012). We chose the hopper environment, where actions were defined by a pretrained controller (Nikulin et al., 2025), and for

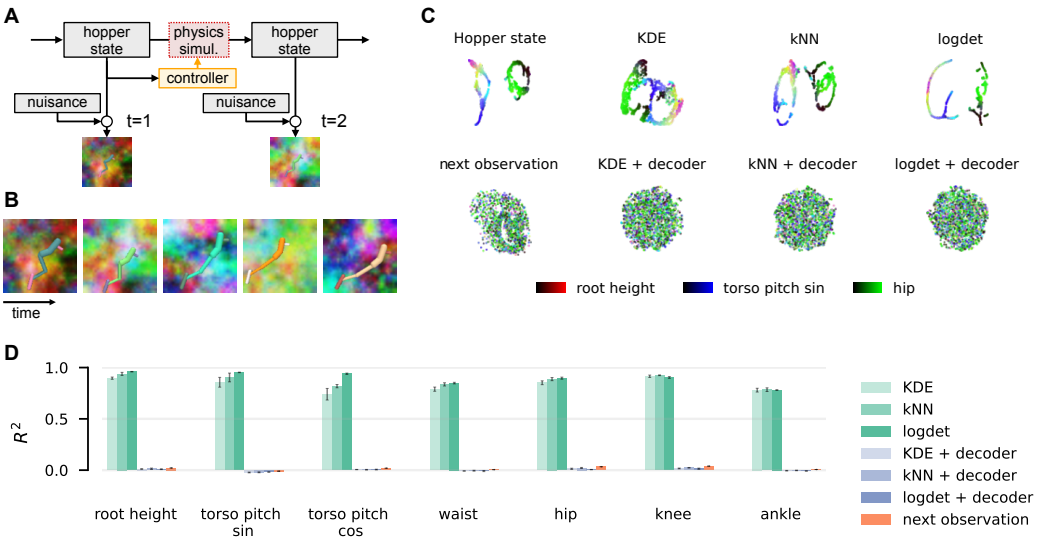


Figure 3: Learning a world model for the controlled MuJoCo hopper. **(A)** Dataset generation. **(B)** Example timeseries. **(C)** UMAP of true signal (hopper state) and learned representations. Colors denote variable values. **(D)** R^2 of latent representations with hopper variables from linear readout.

simulation diversity we added small Gaussian noise on both environment dynamics and controller actions. To introduce nuisance variables we sampled 12-dimensional independent Gaussian nuisance vectors per timestep that were then translated into nuisances such as color changes of the physical body and background contrast, as well as additional high-dimensional background colored noise (Figure 3A,B). We expected that latent representations would capture hopper state variables that are causally related through the simulation.

We used the same setup for learning as for the image dataset (Section 4.1) with a 16-dimensional latent state, while replacing the prediction head with an MLP that received the LSTM output and the encoded action. All models learned representations that allowed to linearly decode the hopper state (Figure 3D), while nonlinear decoding further improved R^2, depending on the entropy estimator (Appendix, Figure 4). The nonlinear recovery probe is reasonable given that the environment does not explicitly follow the Gaussian dynamics of Corollary 1. We also trained models with an image encoder with attached gradient, resulting in representations that did not allow to decode hopper state, but consistently encoded parts of the nuisance variables (Appendix, Figure 4). As additional control we trained a model with the same predictor setup and an image decoder to predict the next observation, which did not result in better performance than direct decoding. This discrepancy can also be seen in the different topological structure of the latent representations (Figure 3C).

5 Conclusion and Limitations

We proved that common SSL learning methods recover signal variables even in the presence of high-dimensional nuisance and stochastic latent dynamics, closing the gap between prior identifiability results, which considered either but not both. This answers the question posed at the outset: representations can formally distinguish unpredictable nuisance from mere stochastic variation in the predictable signal, because predictive MI maximization and LDM jointly constrain what is retained and how. Our result requires only few assumptions, the strongest of which are exponential family conditional distributions in the true latent variables, and an algorithm that successfully maximizes the InfoLDM loss; otherwise, the latent dynamics can be almost arbitrarily complicated, and nuisance can affect observations nontrivially. Future work should explore the viability and real-world relevance of predictive exponential family distributions beyond the Gaussian scenario tested here.

Our theory does not preclude that nuisance is encoded in the learned representation (orthogonally to the signal) if there is capacity to accommodate it, and entropy maximization in principle should encourage this (von Kügelgen et al., 2021; Mikulasch & Zenke, 2026). Empirically we found that

nuisance variables are often not decodable even if the latent dimensionality would permit it, either because they collapse, or because they are represented in a highly nonlinear manner that is not easily inverted. We also found that both signal recovery performance and nuisance leakage depended on the choice of entropy estimator, and the data it was applied to. This suggest that improved entropy estimation is one of the most pressing open problems towards robust and generally applicable SSL methods.

AI USE STATEMENT

In this work, we used generative AI tools to assist in the writing of proofs, and to provide critical ingredients for proving mathematical claims, especially correcting the proof. We have not used generative AI tools to help develop theoretical models or conceptual frameworks, formulate mathematical claims, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods and generate synthetic data sets, or to interpret results. Cleaning and reformatting datasets and the support qualitative and thematic data analysis does not apply to our work. Additionally, we used generative AI tools to create or edit software code. We have reviewed all AI-assisted work. LLM-generated code was verified and tested for correctness, proofs were checked and formalized in Lean and the Lean statements critically reviewed. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This paper presents work whose goal is to advance the field of machine learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

REPRODUCIBILITY STATEMENT

For reproducibility we will provide all simulation and evaluation code, as well as Lean proof formalizations at github.com.

REFERENCES

Laurence Aitchison and Stoil Krasimirov Ganev. InfoNCE is variational inference in a recognition parameterised model. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=chbRsWwjax.

Ido Ben-Shaul, Ravid Shwartz-Ziv, Tomer Galanti, Shai Dekel, and Yann LeCun. Reverse engineering self-supervised learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 58324–58345, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/b63ad8c24354b0e5bcb7aea16490beab-Paper-Conference.pdf.

Alice Bizeul, Bernhard Schölkopf, and Carl Allen. A probabilistic model behind self- supervised learning. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=QEwz7447tR.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 1597–1607. PMLR, 2020. URL https://proceedings.mlr.press/v119/chen20j.html.

Imant Daunhawer, Alice Bizeul, Emanuele Palumbo, Alexander Marx, and Julia E Vogt. Identifiability results for multimodal contrastive learning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=U_2kuqoTcB.

Anna Dawid and Yann LeCun. Introduction to latent variable energy-based models: a path toward autonomous machine intelligence. *Journal of Statistical Mechanics: Theory and Experiment*, 2024 (10):104011, 2024. URL https://doi.org/10.1088/1742-5468/ad292b.

Luigi Gresele, Paul K. Rubenstein, Arash Mehrjou, Francesco Locatello, and Bernhard Schölkopf. The incomplete rosetta stone problem: Identifiability results for multi-view nonlinear ica. In *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115, pp. 217–227. PMLR, 2020. URL https://proceedings.mlr.press/v115/gresele20a.html.

Jie Gui, Tuo Chen, Jing Zhang, Qiong Cao, Zhenan Sun, Hao Luo, and Dacheng Tao. A survey on self-supervised learning: Algorithms, applications, and future trends. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9052–9071, 2024. URL https://doi.org/10.1109/TPAMI.2024.3415112.

Aapo Hyvarinen, Hiroaki Sasaki, and Richard Turner. Nonlinear ica using auxiliary variables and generalized contrastive learning. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89, pp. 859–868. PMLR, 2019. URL https://proceedings.mlr.press/v89/hyvarinen19a.html.

A. Hyvärinen and E. Oja. Independent component analysis: algorithms and applications. *Neural Networks*, 13(4):411–430, 2000. ISSN 0893-6080. URL https://www.sciencedirect.com/science/article/pii/S0893608000000265.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108, pp. 2207–2217. PMLR, 2020. URL https://proceedings.mlr.press/v108/khemakhem20a.html.

Rodrigo González Laiz, Tobias Schmidt, and Steffen Schneider. Self-supervised contrastive learning performs non-linear system identification. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=ONfWFluZBI.

Qi Lyu, Xiao Fu, Weiran Wang, and Songtao Lu. Understanding latent correlation-based multiview learning and self-supervision: An identifiability perspective. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=5FUq05QRc5b.

Fabian A Mikulasch and Friedemann Zenke. Understanding self-supervised learning via latent distribution matching. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=schXkGSTus.

Hiroki Nakamura, Masashi Okada, and Tadahiro Taniguchi. Representation uncertainty in self-supervised learning as variational inference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16484–16493, 2023. URL https://doi.org/10.1109/ICCV51070.2023.01511.

Alexander Nikulin, Ilya Zisman, Denis Tarasov, Lyubaykin Nikita, Andrei Polubarov, Igor Kiselev, and Vladislav Kurenkov. Latent action learning requires supervision in the presence of distractors. In *Proceedings of Machine Learning Research*, volume 267, pp. 46427–46447, 2025. URL https://proceedings.mlr.press/v267/nikulin25a.html.

Mehdi Noroozi and Paolo Favaro. Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles. In *Computer Vision – ECCV 2016*, Lecture Notes in Computer Science, pp. 69–84. Springer International Publishing, 2016. ISBN 978-3-319-46466-4. URL https://doi.org/10.1007/978-3-319-46466-4_5.

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

Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. On variational bounds of mutual information. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 5171–5180. PMLR, 2019. URL https://proceedings.mlr.press/v97/poole19a.html.

Borja Rodríguez Gálvez, Arno Blaas, Pau Rodriguez, Adam Golinski, Xavier Suau, Jason Ramapuram, Dan Busbridge, and Luca Zappella. The role of entropy and reconstruction in multi-view self-supervised learning. In *Proceedings of the 40th International Conference on Machine Learning*,

volume 202, pp. 29143–29160. PMLR, 2023. URL https://proceedings.mlr.press/v202/rodri-guez-galvez23a.html.

Geoffrey Roeder, Luke Metz, and Durk Kingma. On linear identifiability of learned representations. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 9030–9039. PMLR, 2021. URL https://proceedings.mlr.press/v139/roeder21a.html.

Nikunj Saunshi, Orestis Plevrakis, Sanjeev Arora, Mikhail Khodak, and Hrishikesh Khandeparkar. A theoretical analysis of contrastive unsupervised representation learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 5628–5637. PMLR, 2019. URL https://proceedings.mlr.press/v97/saunshi19a.html.

Ravid Shwartz Ziv and Yann LeCun. To compress or not to compress—self-supervised learning and information theory: A review. *Entropy*, 26(3):252, 2024. URL https://doi.org/10.3390/e26030252.

Ravid Shwartz-Ziv, Randall Balestriero, Kenji Kawaguchi, Tim G. J. Rudner, and Yann LeCun. An information theory perspective on variance-invariance-covariance regularization. In *Advances in Neural Information Processing Systems*, volume 36, pp. 33965–33998, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/6b1d4c03391b0aa6ddde0b807a78c950-Paper-Conference.pdf.

Henning Sprekeler, Tiziano Zito, and Laurenz Wiskott. An extension of slow feature analysis for nonlinear blind source separation. *Journal of Machine Learning Research*, 15(26):921–947, 2014. URL http://jmlr.org/papers/v15/sprekeler14a.html.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *European conference on computer vision*, pp. 776–794. Springer, 2020. URL https://doi.org/10.1007/978-3-030-58621-8_45.

Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5026–5033. IEEE, 2012. URL https://doi.org/10.1109/IROS.2012.6386109.

Michael Tschannen, Josip Djolonga, Paul K. Rubenstein, Sylvain Gelly, and Mario Lucic. On mutual information maximization for representation learning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rkxoh24FPH.

Julius von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style. In *Advances in Neural Information Processing Systems*, volume 34, pp. 16451–16467. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/8929c70f8d710e412d38da624b21c3c8-Paper.pdf.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 9929–9939. PMLR, 2020. URL https://proceedings.mlr.press/v119/wang20k.html.

Roland S. Zimmermann, Yash Sharma, Steffen Schneider, Matthias Bethge, and Wieland Brendel. Contrastive learning inverts the data generating process. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 12979–12990. PMLR, 2021. URL https://proceedings.mlr.press/v139/zimmermann21a.html.

A Appendix

A.1 Mathematical appendix

A.1.1 Detailed assumptions

Assumption 1 (Private nuisance)**.** *The stochastic part of the model factorizes as*

$$p(n, s, x_c, c) = p(n \mid s)p(s \mid c)p(x_c \mid c)p(c), \tag{12}$$

with $x = g(s, n)$, $z = f(x)$, and $z_c = f_c(x_c)$. More specifically, the nuisance is private to the current observation in the sense that

$$n \perp\!\!\!\perp c \mid s. \tag{13}$$

The privacy condition has the equivalent information-theoretic form

$$I[(s, n); c] = I[s; c] + I[n; c \mid s] = I[s; c]. \tag{14}$$

Thus the nuisance carries no information about the condition beyond the signal.

Assumption 2 (Exponential-family prediction)**.** *Let $\tau_\star : \mathcal{S} \to \mathbb{R}^{r_s}$ be a sufficient statistic, and let $h_s(s)$ be a fixed carrier function on $\mathcal{S}$. For each condition c, let $\alpha(c) \in \mathbb{R}^{r_s}$ be its natural parameter, and let ψ be the corresponding log-partition function. The true predictive family is*

$$p(s \mid c) = h_s(s) \exp\big(\alpha(c)^\top \tau_\star(s) - \psi(\alpha(c))\big) \tag{15}$$

for almost every condition c.

Similarly, let $\tau_z : \mathcal{Z} \to \mathbb{R}^{r_z}$ be a sufficient statistic, and let $h_z(z)$ be a fixed carrier function on $\mathcal{Z}$. For each encoded condition z_c, let $\beta(z_c) \in \mathbb{R}^{r_z}$ be its learned natural parameter, and let ϕ be the corresponding log-partition function. The learned predictive model is

$$p_\theta(z \mid z_c) = h_z(z) \exp\big(\beta(z_c)^\top \tau_z(z) - \phi(\beta(z_c))\big) \tag{16}$$

for almost every encoded condition z_c.

As mentioned before, it is not necessary to assume any particular marginal distribution of the encoded conditioning variables z_c. In the joint formulation, one may set $p_\theta(z_c) = q_f(z_c)$, so the KL term compares only the predictive conditionals (Equation 2).

Assumption 3 (Distribution matching and information saturation)**.** *Let $q_f(z, z_c)$ be the encoded joint distribution and let $p_\theta(z, z_c)$ be a learned predictive model. Information-augmented predictive LDM uses the population objective*

$$\mathcal{F}(f, f_c, \theta) = -D_{\mathrm{KL}}[q_f(z, z_c) \| p_\theta(z, z_c)] + I_{q_f}[z; z_c]. \tag{17}$$

We assume that an ideal population optimum is realizable through optimizing the model p_θ, encoder f, and the condition encoder f_c, at which

$$D_{\mathrm{KL}}[q_f(z, z_c) \| p_\theta(z, z_c)] = 0, \qquad I_{q_f}[z; z_c] = I[(s, n); c] = I[s; c] < \infty. \tag{18}$$

As discussed in the main text, this assumption implicitly replaces common invertibility assumptions. It also implies that latent distributions match, and for the proof we define

$$p(z, z_c) := q_f(z, z_c) = p_\theta(z, z_c). \tag{19}$$

Assumption 4 (Condition diversity)**.** *Even after excluding an arbitrary zero-probability subset of conditions, one can choose $c_0, c_1, \ldots, c_{r_s}$ such that*

$$D := \begin{bmatrix} (\alpha(c_1) - \alpha(c_0))^\top \\ \vdots \\ (\alpha(c_{r_s}) - \alpha(c_0))^\top \end{bmatrix} \in \mathbb{R}^{r_s \times r_s} \tag{20}$$

is invertible.

Invertability requires the condition variation to span all r_s natural-parameter directions. In the proof, these directions supply r_s independent likelihood-ratio equations that recover every coordinate of the source sufficient statistic.

Assumption 5 (No affine redundancy in the source statistic)**.** *For every condition c for which Equation 15 holds, the random vector $\tau_\star(s)$ with $s \sim p(s \mid c)$ is not almost surely contained in a proper affine hyperplane of $\mathbb{R}^{r_s}$.*

This assumption complements Assumption 4 and says that the sufficient statistic also genuinely uses all r_s coordinates.

A.1.2 FULL PROOF OF THEOREM 1

Theorem 1 (Restated) (Exponential-family readout in InfoLDM)**.** *Suppose Assumptions 1, 2, 3, 4, and 5 hold, and the InfoLDM goal function (Equation 1) is fully maximized.*

Then the dimension of the model sufficient statistic $r_z \geq r_s$, and there is a full-row-rank matrix $A \in \mathbb{R}^{r_s \times r_z}$ and a vector $b \in \mathbb{R}^{r_s}$ such that

$$\boxed{\tau_\star(s) = A\tau_z(z) + b} \qquad \textit{almost surely}. \tag{21}$$

If $\tau_\star$ is injective, then s is almost surely a measurable function of z. If $r_z = r_s$, then A is invertible and the two sufficient statistics are related by an invertible affine transformation.

Proof. The proof proceeds in four steps. In Step 1, we show that MI maximization implies that the learned context z_c and true context c are interchangeable for latent prediction. In Step 2, we show that it also implies that it is possible to equate the likelihood ratios of the learned representation z and the of true signal s with respect to some reference condition c_0. This is the key identity that explains why representations have to recover the signal.

The following two steps are closely related to previous exponential family identifiability proofs (Khemakhem et al., 2020). In Step 3 we use the above likelihood ratio, the exponential family assumption, and the condition diversity assumption to establish that there is an affine relation between the statistics $\tau_z(z)$ and $\tau_\star(s)$ when z and s are sampled under c_0. Finally, we extend this result in Step 4 to the entire $p(c)$, and derive the full-rank and invertibility relating z and s.

Step 1: Information saturation and transfer. The chain rule of mutual information lets us compute

$$\begin{aligned} I[s;c] - I[z;z_c] &= \big(I[s;c] - I[z;c]\big) + \big(I[z;c] - I[z;z_c]\big) \\ &= \big(I[s,z;c] - I[z;c]\big) + \big(I[z;c,z_c] - I[z;z_c]\big) \\ &= I[s;c \mid z] + I[z;c \mid z_c]. \end{aligned} \tag{22}$$

The second step follows, since from the generative model (Figure 1) and Assumption 1 we know that $z \perp\!\!\!\perp c \mid s$ and $z \perp\!\!\!\perp z_c \mid c$. Equation 22, together with Assumption 3 implies $s \perp\!\!\!\perp c \mid z$ and $z \perp\!\!\!\perp c \mid z_c$. Consequently, for almost every jointly occurring pair (c, z_c),

$$p(z \mid c) = p(z \mid c, z_c) = p(z \mid z_c). \tag{23}$$

Thus $p(z \mid c)$ is a member of the learned exponential family for almost every condition c.

Step 2: Likelihood-ratio equality. The two conditional independences involving s tell us that the conditional distributions relating s and z are independent of c, i.e., $p(z \mid s, c) = p(z \mid s)$ and $p(s \mid z, c) = p(s \mid z)$. Therefore, the following two factorizations hold for almost every condition c:

$$p(s, z \mid c) = p(z \mid s)p(s \mid c) = p(s \mid z)p(z \mid c). \tag{24}$$

Because the conditional distributions between s and z in Equation 24 do not depend on c, they cancel when the distributions at c and some reference c_0 are compared

$$\frac{p(s \mid c)}{p(s \mid c_0)} = \frac{p(z \mid c)}{p(z \mid c_0)} \qquad \text{for almost every } (s, z) \text{ generated under } c_0. \tag{25}$$

The ratios are well defined because all selected family members are strictly positive wherever their common carrier functions are positive.

Step 3: Locally affine relation. We now want to select a number of r_s different contexts c_i that constrain the relation between s and z. By using Equation 25 for any c_i we get

$$\frac{p(s \mid c_i)}{p(s \mid c_0)} = \frac{p(z \mid c_i)}{p(z \mid c_0)} \qquad \text{for almost every } (s, z) \text{ generated under } c_0. \tag{26}$$

Taking logarithms, and using the two exponential-family forms (Assumption 2) we get

$$(\alpha(c_i)^\top - \alpha(c_0)^\top)\tau_\star(s) - [\psi(\alpha(c_i)) - \psi(\alpha(c_0))] = (\beta_i - \beta_0)^\top \tau_z(z) - [\phi(\beta_i) - \phi(\beta_0)]. \tag{27}$$

Here we chose parameters β_i for which (by the result of Step 1)

$$p(z \mid c_i) = h_z(z) \exp\big(\beta_i^\top \tau_z(z) - \phi(\beta_i)\big)\,. \tag{28}$$

We now define $d_i := \alpha(c_i) - \alpha(c_0)$, and $e_i := \beta_i - \beta_0$. With this notation Equation 27 becomes

$$d_i^\top \tau_\star(s) = e_i^\top \tau_z(z) + h_i, \qquad h_i := \psi(\alpha(c_i)) - \psi(\alpha(c_0)) - \phi(\beta_i) + \phi(\beta_0). \tag{29}$$

Let D have rows $d_i^\top$, E have rows $e_i^\top$, and let $h = (h_1, \ldots, h_{r_s})^\top$. Thus, stacking Equation 29 gives

$$D\tau_\star(s) = E\tau_z(z) + h \qquad \text{for almost every } (s, z) \text{ generated under } c_0. \tag{30}$$

Finally, Assumption 4 lets us choose conditions $c_0, c_1, \ldots, c_{r_s}$ for which D is invertible. This proves the readout for samples generated under c_0, with

$$A := D^{-1}E, \qquad b := D^{-1}h. \tag{31}$$

Step 4: Globalization, rank, and reconstruction. What remains is to extend this reference-condition identity to the full joint distribution, prove that A has full row rank, and derive the reconstruction conclusions.

For globalization, we note that for almost every condition c, the distribution $p(s \mid c)$ has the same zero-probability sets as $p(s \mid c_0)$ because both are strictly positive on the common carrier. Multiplying these distributions with the same $p(z \mid s)$ preserves this property for the conditional joint distributions of $p(s, z \mid c)$, i.e., any set of probability zero under $p(s, z \mid c_0)$ has probability zero under $p(s, z \mid c)$. In particular, the set $\{(s, z) : \tau_\star(s) \neq A\tau_z(z) + b\}$ has probability zero under $p(s, z \mid c_0)$ and hence under $p(s, z \mid c)$. The readout therefore holds conditionally for almost every condition c. Averaging over $p(c)$ gives Equation 8 under the original joint distribution.

To show that A is full rank, assume it were not, i.e., $\mathrm{rank}(A) < r_s$. In this case the reference-condition identity (Equation 30) places $\tau_\star(s)$ in the proper affine subspace $b + \mathrm{im}(A)$ under $p(s \mid c_0)$. Because $p(s \mid c_0)$ is strictly positive on the common carrier, this contradicts Assumption 5. Thus $\mathrm{rank}(A) = r_s$ and $r_z \geq r_s$. When $r_z = r_s$, the full-row-rank matrix A is square and hence invertible.

To reconstruct s from z we have to invert $\tau_\star$. If $\tau_\star$ is injective, the Lusin–Souslin theorem supplies a measurable inverse on its image, which may be extended measurably by a fixed value outside that image, so

$$s = \tau_\star^{-1}(A\tau_z(z) + b) \qquad \text{almost surely.} \tag{32}$$

□

A.2 LEAN FORMALIZATION

To verify the proof we translated it to compiling Lean code using generative AI. The declaration `statisticReadout_of_exponentialFamily_nuisanceLDM` formalizes Theorem 1 after the mutual-information argument. Below is the header of the main function which includes the main assumptions and the conclusion. The line numbers refer to the actual Lean source, the full declaration and proof can be found in the source code at `Lean/ExponentialFamilyReadout.lean`.

```
theorem statisticReadout_of_exponentialFamily_nuisanceLDM
    {Ω H V X Y : Type*}
    [MeasurableSpace Ω] [StandardBorelSpace Ω]
    [MeasurableSpace H] [StandardBorelSpace H] [Nonempty H]
    [MeasurableSpace V] [StandardBorelSpace V] [Nonempty V]
    [MeasurableSpace X] [StandardBorelSpace X] [Nonempty X]
    [MeasurableSpace Y] [StandardBorelSpace Y] [Nonempty Y]
    (P : Measure Ω) [IsProbabilityMeasure P]
    (history : Ω → H) (view : Ω → V)
    (signal : Ω → X) (code : Ω → Y)
    (hhistory : Measurable history) (hview : Measurable view)
    (hsignal : Measurable signal) (hcode : Measurable code)
    (signal_history_given_code : signal ⟂ᵢ[code, hcode; P] history)
    (code_history_given_signal : code ⟂ᵢ[signal, hsignal; P] history)
    (code_history_given_view : code ⟂ᵢ[view, hview; P] history)
    (code_view_given_history : code ⟂ᵢ[history, hhistory; P] view)
    (sourceCarrier : Measure X) (codeCarrier : Measure Y)
    (sourceStatistic : X → EuclideanSpace ℝ ι)
```

```
776     (codeStatistic : Y → EuclideanSpace ℝ κ)
777     (hsourceStatistic : Measurable sourceStatistic)
778     (hcodeStatistic : Measurable codeStatistic)
779     (sourcePartition : EuclideanSpace ℝ ι → ℝ)
780     (codePartition : EuclideanSpace ℝ κ → ℝ)
781     (sourceParameter : H → EuclideanSpace ℝ ι)
782     (learnedParameter : V → EuclideanSpace ℝ κ)
783     (hsourceConditional :
784       condDistrib signal history P =ᵐ[P.map history]
785         fun c ↦ statisticNaturalExpMeasure sourceCarrier sourceStatistic
786           sourcePartition (sourceParameter c))
787     (hlearnedConditional :
788       condDistrib code view P =ᵐ[P.map view]
789         fun v ↦ statisticNaturalExpMeasure codeCarrier codeStatistic
790           codePartition (learnedParameter v))
791     (hhistoryDiversity :
792       ∀ good : Set H, (∀ᵐ c ∂P.map history, c ∈ good) →
793         ∃ (c₀ : H) (c : ι → H),
794           c₀ ∈ good ∧
795           (∀ i, c i ∈ good) ∧
796           Submodule.span ℝ
797             (Set.range fun i ↦ sourceParameter (c i) - sourceParameter c₀) = ⊤)
798     (hsourceNoAffineRedundancy :
799       ∀ (w : EuclideanSpace ℝ ι) (a : ℝ),
800         (fun x ↦ inner ℝ w (sourceStatistic x)) =ᵐ[sourceCarrier]
801             (fun _ ↦ a) →
802           w = 0) :
803     ∃ (A : EuclideanSpace ℝ κ →L[ℝ] EuclideanSpace ℝ ι)
804         (b : EuclideanSpace ℝ ι),
805       (fun ω ↦ sourceStatistic (signal ω)) =ᵐ[P]
806           (fun ω ↦ A (codeStatistic (code ω)) + b) ∧
807       Function.Surjective A ∧
808       Module.finrank ℝ (EuclideanSpace ℝ ι) ≤
809         Module.finrank ℝ (EuclideanSpace ℝ κ) ∧
810       (Function.Injective sourceStatistic →
811         ∃ decoder : Y → X, Measurable decoder ∧
812           signal =ᵐ[P] fun ω ↦ decoder (code ω)) ∧
813       (Module.finrank ℝ (EuclideanSpace ℝ κ) =
814           Module.finrank ℝ (EuclideanSpace ℝ ι) →
815         Function.Bijective A) := by
```

This declaration reads in blocks as follows.

Lines 759–769: probability space and variables. `Omega` is the underlying probability space, and `H`, `V`, `X`, and `Y` are the state spaces of c, z_c, s, z. The standard-Borel and measurability assumptions are the formal version of the measurable setup in Section 1. The finite types ι and κ index the coordinates of $\tau_\star$ and τ_z, so their `finranks` are r_s and r_z.

Lines 770–773: the four conditional independences. In the order shown, these are

$$s \perp\!\!\!\perp c \mid z, \qquad z \perp\!\!\!\perp c \mid s, \qquad z \perp\!\!\!\perp c \mid z_c, \qquad z \perp\!\!\!\perp z_c \mid c. \tag{33}$$

The second and fourth are the Markov consequences of Assumption 1. The first and third follow in the paper from

$$I[s;c] - I[z;z_c] = I[s;c \mid z] + I[z;c \mid z_c] = 0. \tag{34}$$

Thus Lean starts at the exact conditional-independence consequences of MI saturation, rather than formalizing mutual information and the optimization problem.

Lines 774–790: the two exponential families. The names `sourceCarrier`, `sourceStatistic`, `sourcePartition`, and `sourceParameter` correspond to $h_s, \tau_\star, \psi, \alpha$. Their code-side counterparts correspond to h_z, τ_z, ϕ, β. The two hypotheses `hsourceConditional` and `hlearnedConditional` are the conditional families

$$p(s \mid c) = h_s(s) \exp\big(\alpha(c)^\top \tau_\star(s) - \psi(\alpha(c))\big), \tag{35}$$

$$p(z \mid z_c) = h_z(z) \exp\big(\beta(z_c)^\top \tau_z(z) - \phi(\beta(z_c))\big), \tag{36}$$

where `statisticNaturalExpMeasure` is the Lean implementation of the exponential-family distribution in Equation 15 and Equation 16. Lean's eventual-equality notation means that the displayed equality may fail only on a zero-probability set.

Lines 791–802: diversity and nonredundancy. The hypothesis `hconditionDiversity` is Assumption 4 written without coordinates: after removing any zero-probability set of conditions, it selects c_0 and a family (c_i) whose differences $\alpha(c_i) - \alpha(c_0)$ span the whole source-statistic space. The hypothesis `hsourceNoAffineRedundancy` is Assumption 5: if $\langle w, \tau_\star(s) \rangle$ is almost surely constant on the carrier, then $w = 0$.

Lines 803–815: conclusion. Lean returns a continuous linear map $A : \mathbb{R}^{r_z} \to \mathbb{R}^{r_s}$ and an offset b with

$$\tau_\star(s) = A\tau_z(z) + b \qquad \text{almost surely.} \tag{37}$$

`Function.Surjective A` is the full-row-rank conclusion $\mathrm{rank}(A) = r_s$, and the `finrank` inequality is $r_s \leq r_z$. The last two lines add the two corollaries in Theorem 1: an injective $\tau_\star$ gives a measurable decoder $z \mapsto s$, and equal statistic dimensions make A bijective. The proof first establishes the likelihood-ratio identity at a reference condition and then extends it to the full model.

A.3 Simulation details

Numerical evaluation. We generated 100000 training and 10000 test sequences of length five. The fixed observation map consisted of three affine layers with leaky-ReLU activations and mapped the concatenated signal and nuisance variables to 100-dimensional observations. The encoder was a five-layer MLP with 200 hidden units per layer and ReLU activations. Its output dimension matched the signal dimension, which was varied over $d_S \in \{5, 10, 15, 20\}$. A single-layer LSTM with 10 hidden units summarized the observation history, and a linear prediction head predicted the next latent state.

Models were trained for 20 epochs using Adam and a linearly decaying learning rate. Batch size was 256 for the kNN and log-determinant estimators. In this and the following experiments, entropy estimation was implemented as described by Mikulasch & Zenke (2026). We found that KDE performed poorly in this experiment and increased batchsize to 5096 (similar to von Kügelgen et al., 2021) and the number of samples to 5×10^6 to improve performance.

Causal3DIdent. We based our experiment on the data of von Kügelgen et al. (2021) but resized images to 224×224 pixels. We used a ResNet-18 encoder producing a 10-dimensional representation, followed by a single-layer LSTM with 64 hidden units and a linear prediction head. Models were trained for 25 epochs with Adam. Batch sizes were 256 for the kNN and KDE estimators and 64 for the log-determinant estimator.

MuJoCo Hopper. We generated 100000 training and 1000 test rollouts from the DM-Control Hopper environment. Each sequence contained eight 128×128 rendered frames, separated by three control steps. Actions were produced by a pretrained controller (Nikulin et al., 2025) and perturbed with Gaussian exploration noise of standard deviation 0.1. Additional unobserved actuator noise with standard deviation 0.5 was applied before executing each action, making the controlled dynamics stochastic. Note that adding Gaussian noise every simulation step, not every render step, results in a system that does not fully follow the assumptions of Theorem 1. At every rendered timestep, independent nuisance variables modified the body and accent colors, camera position and field of view, illumination, and a procedurally generated colored background. The physical signal used for evaluation consisted of seven position variables and seven velocity variables.

Images were encoded by a ResNet-18 into a 16-dimensional latent state. A single-layer LSTM with 64 hidden units summarized the latent history. The three controller commands between consecutive frames were encoded by a GRU with 16 hidden units and supplied, together with the LSTM state, to a one-hidden-layer prediction MLP. For image decoding we used 4-layer upsampling CNNs trained with L2 loss. With attached decoder loss we approximately balanced latent prediction and image prediction loss via a Lagrange multiplier. For the observation prediction model we used the exact same setup without latent prediction loss and the decoder attached to the prediction module. Models were trained for 40 epochs using Adam. Batch sizes were 128, 512, 32, and 128 for the kNN, KDE, log-determinant estimators, and next observation prediction, respectively.

A.4 Additional results

A.4.1 Causal3DIdent

Table 3: Nonlinear R^2 for the Causal3DIdent experiment (Table 2), computed with trained MLP predictor on held-out dataset.

Scenario	Entr. est.	$p_{\text{obj-}x}$	$p_{\text{obj-}y}$	$p_{\text{obj-}z}$	$\theta_{\text{obj-}\alpha}$	$\theta_{\text{obj-}\beta}$	$\theta_{\text{obj-}\gamma}$	h_{obj}	p_{spot}	h_{spot}	h_{back}
Environment signal	kde	0.35	0.52	0.34	0.07	0.10	0.16	0.53	**0.99**	**0.97**	**0.99**
Environment signal	knn	0.21	0.47	0.34	0.08	0.06	0.08	0.16	**0.98**	**0.96**	**0.98**
Environment signal	logdet	−0.02	−0.02	−0.03	−0.02	−0.01	−0.00	−0.01	**0.99**	0.57	**0.98**
Object signal	kde	**0.98**	**0.98**	**0.97**	**0.94**	**0.97**	0.85	**0.91**	0.14	0.03	0.36
Object signal	knn	**0.97**	**0.96**	**0.95**	0.87	**0.92**	0.29	**0.91**	0.14	−0.04	0.69
Object signal	logdet	**0.99**	**0.99**	**0.98**	**0.95**	**0.97**	**0.96**	**0.92**	0.16	−0.04	−0.04

A.4.2 Hopper

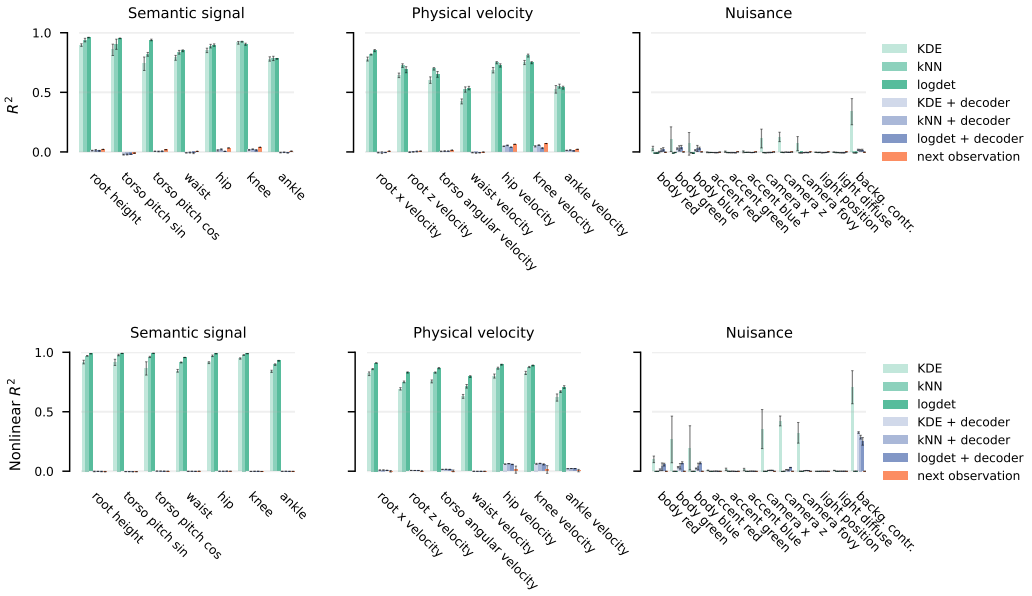


Figure 4: Linear (Top) and nonlinear (Bottom) R^2 with physical (signal) and nuisance variables for the hopper experiment. Velocity is read out from LSTM state. Background pattern nuisance is not included.