REPRESENTATION SELECTIVITY IN INTERACTING SYSTEMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning representations of interacting systems requires capturing not only the individual components, but also the relations between them. This becomes particularly important under restricted representation capacity, where not all information can be preserved, and the representation must select what is relevant. However, it remains unclear whether relational information should be formed before individual subsystem representations are compressed, or reconstructed afterward. In this work, we study how this choice interacts with the prediction objective, a phenomenon we call *representation selectivity*. We show that relational quantities can be represented compactly when information from both systems is available before compression, while separate compression may require preserving richer local information. Controlled nonlinear systems allow us to disentangle relational structure, physical coupling, and representation capacity. On the same interacting physical system, representations trained to predict future states show essentially no benefit from combining the two systems before compression, despite predicting the future state substantially better than a constant baseline. In contrast, when trained to predict force and relative motion at the physical interface, early access to both systems provides a large and consistent advantage. Our results show that, under restricted capacity, learning interaction-aware representations depends not only on architecture, but critically on what the representation is trained to predict.

1 INTRODUCTION

When systems interact, their behavior is determined by the state of their individual components and by relations that emerge between them. In robotics, these relations arise naturally in manipulation, contact-rich interaction, and multimodal perception, through relative motion, contact forces, constraints, and information and energy exchange across interfaces. Efficient robot models must compress these rich observations while retaining the information needed for prediction, reasoning, and control. Under restricted representation capacity, this requires selecting what information should survive compression. Should each subsystem therefore be compressed independently and combined afterwards, or should cross-system relations be formed before compression?

What information should survive compression also depends on what the representation is trained to predict. A generic future-prediction objective may favor different information than one that explicitly depends on an interaction. We therefore study two coupled choices: *where* information from different subsystems is combined relative to compression, and *what* the learning objective makes relevant. We refer to their interplay as *representation selectivity*.

We formalize this interplay through the system partition, prediction target, and representation capacity. A structural result characterizes when a cross-system relation can be represented compactly after joint access but requires richer local information after separate compression, while a target-relative decomposition makes explicit that relevance depends on what is predicted. This leads to controlled tests that separate relational structure from physical coupling, vary the prediction target, and relax the representation bottleneck.

The central observation arises when we hold an interacting physical system fixed and change what the representation is trained to predict. In a stochastic double-well system, representations trained to predict future states substantially outperform a constant predictor, yet combining the two subsystems before compression provides essentially no advantage. When the target instead describes the

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.

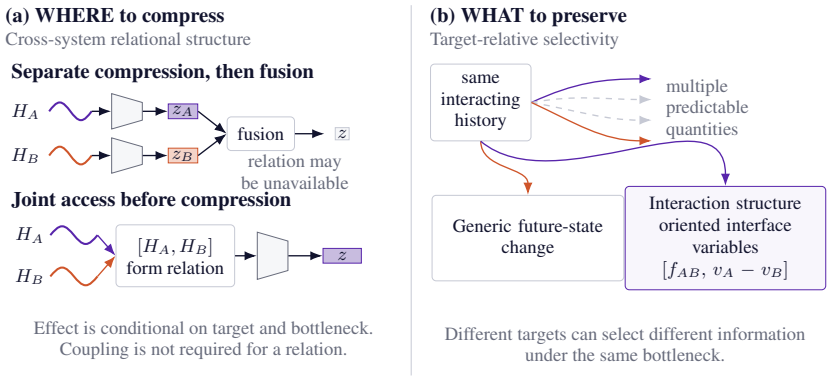

Figure 1: **Where to compress and what to preserve.** (a) H_A and H_B denote the observed histories of systems A and B. Separate encoders compress them into z_A and z_B before fusion into z; a joint encoder instead observes $[H_A, H_B]$ and can form a cross-system relation before compression. (b) The same observations can support different prediction targets. f_{AB} denotes the force across the A–B interface and $v_A - v_B$ their relative velocity. Under restricted capacity, changing the prediction target can change which information is preserved.

physical interface, predicting interaction force and relative motion produces a large and consistent joint-access advantage. The partner history is informative about these interface quantities in both settings: making interaction information available therefore does not by itself ensure that a generic predictive objective will preserve it.

Our contributions are threefold. **(i) Representation selectivity.** We formalize how system partition, prediction target, and restricted representation capacity jointly determine which cross-system information can remain accessible after compression. **(ii) Controlled dissociation of relational structure, interaction, and capacity.** Across nonlinear dynamical systems, we show that a joint-access advantage can arise without physical coupling, changes with the prediction target, and weakens when the relevant local information can be retained separately. **(iii) Target selection for interaction-aware representations.** On the same interacting physical system, generic future prediction yields essentially no joint-access advantage despite substantially outperforming a constant predictor, whereas supervision through oriented interface variables produces a large and consistent effect. A planar-contact study provides a complementary boundary: physical meaning alone does not guarantee that a prediction target preserves the directional information required downstream.

2 Foundations and Theory

We formalize representation selectivity through three coupled elements: the system partition, the prediction target, and the available representation capacity. The partition determines which relations can be formed before compression; the target determines which information is relevant; and capacity determines which of these distinctions can survive compression. We develop these elements below and then derive the experimental predictions that follow from their interaction.

2.1 System partitions and joint access

Consider two observed histories H_A and H_B and a prediction target Y. A joint encoder constructs its representation after observing both histories,

$$Z_J = E_J(H_A, H_B), \tag{1}$$

whereas a separate architecture first compresses each history independently and combines the resulting representations afterwards,

$$Z_A = E_A(H_A), \qquad Z_B = E_B(H_B), \qquad Z_S = g(Z_A, Z_B). \tag{2}$$

The two architectures differ in what must survive compression. The joint encoder can construct a relation between H_A and H_B before information is discarded. The separate architecture must preserve enough information about each subsystem for the same relation to remain recoverable after fusion.

The following simple construction makes this distinction explicit.

Proposition 1 (Joint statistics under separate compression)**.** *Let $a, b \in \mathbb{R}^m$ and consider the relational target $Y(a, b) = a^\top b$. A joint encoder can represent Y exactly with one real-valued coordinate. Suppose instead that continuous encoders $E_A : \mathbb{R}^m \to \mathbb{R}^{d_A}$ and $E_B : \mathbb{R}^m \to \mathbb{R}^{d_B}$ are applied separately and that there exists a decoder g such that*

$$a^\top b = g(E_A(a), E_B(b)) \qquad \textit{for all } a, b \in \mathbb{R}^m. \tag{3}$$

Then both E_A and E_B must be injective, which requires $d_A, d_B \geq m$.

To see this, suppose $E_A(a_1) = E_A(a_2)$ for some $a_1 \neq a_2$. Choosing $b = a_1 - a_2$ gives

$$a_1^\top b - a_2^\top b = \|a_1 - a_2\|_2^2 > 0, \tag{4}$$

although the decoder receives the same representation from subsystem A. Hence E_A must be injective, and the same argument applies to E_B. Continuity and invariance of domain then give the dimension bound. A complete proof is provided in Appendix B.

The proposition isolates the structural cost of forming a relation only after information has been compressed separately. Related lower bounds for distributed computation have a long history (Abelson, 1980; Luo & Tsitsiklis, 1993), and coding for computing similarly studies what distributed sources must communicate to recover a function (Orlitsky & Roche, 2001). Here, we use this structural distinction to study a representation-learning question: when does the location of compression change which target-relevant relations remain accessible?

Importantly, the relevant quantity is not the dimension of the raw observation itself. If each subsystem admits a compact local statistic that is sufficient for Y, preserving that statistic is enough. Conversely, a relation can be low-dimensional after joint access while requiring richer information to be retained separately on the two sides. This distinction will become central once the prediction target is allowed to change.

2.2 Target-relative relevance

What should survive compression depends on what the representation is expected to predict. A history can contain information about many aspects of the future, while a restricted representation must select which of these distinctions to retain.

Let X summarize the observed history H and consider the linear target

$$Y = \mathcal{G}X + \epsilon, \tag{5}$$

where the innovation ϵ is independent of (X, H, ξ) and ξ denotes encoder randomness. The predictable component of Y is

$$S_Y = \mathbb{E}[Y \mid X] = \mathcal{G}X, \tag{6}$$

written in nonredundant coordinates. Let $X_\perp$ complete S_Y to an invertible representation of X. In the linear-Gaussian construction used in Appendix A,

$$X_\perp = X - \mathbb{E}[X \mid S_Y], \tag{7}$$

with $X_\perp \perp S_Y$.

Let Z be a causal representation of H and assume $I(H; Z) = I(X; Z)$. Innovation independence implies $Y \perp Z \mid S_Y$. Applying the chain rule gives

$$I(H; Z) - I(Z; Y) = I(Z; S_Y \mid Y) + I(Z; X_\perp \mid S_Y). \tag{8}$$

The derivation is given in Appendix A. Gaussianity provides a convenient construction of the orthogonal residual $X_\perp$; the identity itself follows from the sufficiency and invertibility conditions above.

Equation (8) makes the dependence on the target explicit. The first term captures information retained about the predictable statistic S_Y that is not resolved by the realized noisy target. The second captures variation in the observed state that remains in Z beyond what is required to determine S_Y. Such information may be useful for a different target, but it does not improve prediction of the current one. Changing Y therefore changes both the target-predictive statistic and which retained information is relevant.

This observation is particularly important under restricted capacity. A representation cannot generally preserve every predictable property of the history, and the prediction objective determines which distinctions are rewarded for surviving compression. In our experiments, capacity is controlled through representation width rather than an explicit information-rate constraint; we return to this distinction below.

2.3 Physical interaction as structured relational supervision

Physical interaction provides a natural setting in which relevant quantities span subsystem boundaries. Port-Hamiltonian descriptions make this explicit through paired effort and flow variables at an interconnection (van der Schaft & Maschke, 2013). At a mechanical interface, force f and relative velocity v retain the orientation of the interaction, while their pairing

$$P = f^\top v \tag{9}$$

gives scalar power and its time integral gives work.

These quantities describe different aspects of the same interaction. The oriented pair (f, v) contains information that is absent from the scalar pairing: for any orthogonal transformation Q,

$$(Qf)^\top (Qv) = f^\top v. \tag{10}$$

Consequently, scalar power can be sufficient for a task concerned with energy exchange while being insufficient for a task that depends on the direction of force or motion. Physical meaning alone therefore does not determine which interaction variable a representation should preserve; the appropriate quantity remains relative to the downstream prediction problem.

We use this structure to construct controlled targets in the experiments. The oriented interface variables provide a relational target tied directly to the physical interconnection, while scalar work provides a complementary case in which physically meaningful information has been compressed further.

2.4 Representation selectivity: partition, target, and capacity

We can now combine the previous observations. Suppose a target Y depends on the two subsystem histories through local target-relevant statistics,

$$Y = \psi\big(\phi_A^Y(H_A), \phi_B^Y(H_B)\big)\,. \tag{11}$$

The superscript emphasizes that these statistics are defined relative to the target: changing Y can change what information each subsystem must preserve.

A separate architecture must retain sufficient information about $\phi_A^Y(H_A)$ and $\phi_B^Y(H_B)$ through the two compression stages for their relation to remain recoverable after fusion. A joint encoder faces a different problem. Because it observes both histories before compression, it may directly represent a lower-dimensional statistic of their relation without preserving the corresponding local information separately. Proposition 1 gives an exact instance of this distinction.

This view also makes the role of capacity explicit. When the separate representations can retain the target-relevant local statistics, fusion can in principle recover the same relation and there need not be an advantage to joint access. Under restricted capacity, however, the system partition determines which information must survive independently before the relation can be formed. In the width-constrained setting considered here, representation selectivity therefore depends on the interaction between

$$\boxed{\text{system partition} \;\times\; \text{prediction target} \;\times\; \text{representation capacity}}\,. \tag{12}$$

This formulation leads to three predictions that we test below. First, a partition effect does not require physical coupling: it can arise whenever the prediction target itself is relational. Second, changing the target can change the effect even when the observations, architecture, and physical system are fixed, because the target changes which local information must survive compression. Third, increasing the capacity of the separate representations should reduce the effect once the relevant local information can be retained before fusion.

These predictions concern the width-constrained representations studied here. Real-valued width is not an information rate, and learned representations are additionally shaped by finite data, optimization, and regularization. We therefore do not interpret representation width as a general communication budget. Within the setting considered here, however, the framework makes a concrete prediction: the value of joint access should follow the target-relevant information exposed by the system partition, rather than physical coupling alone.

3 Controlled Tests of Representation Selectivity

We first isolate the partition and target effects in two stochastic nonlinear oscillators,

$$\begin{aligned} dq_A &= p_A\,dt, \qquad dp_A = [-\omega_A^2 q_A - \alpha q_A^3 - \gamma p_A + k\tanh(q_B - q_A)]dt + \sigma dW_A, \\ dq_B &= p_B\,dt, \qquad dp_B = [-\omega_B^2 q_B - \alpha q_B^3 - \gamma p_B + k\tanh(q_A - q_B)]dt + \sigma dW_B. \end{aligned} \tag{13}$$

The Wiener processes are independent; $k = 0$ removes interaction and $k = 4$ introduces reciprocal coupling. Histories contain three observations and the prediction horizon is $0.20\,\mathrm{s}$. At equal total width d, the joint encoder observes both histories before projection, while the separate model assigns $d/2$ coordinates to each history before fusion. The architectures are matched within 1.3% in parameter count and use the same training budget. Targets are standardized using training data only, and accessibility is measured with a common nonlinear reader fitted to the frozen representations. For the partition experiment, we additionally report the models' trained prediction heads. Full architecture, optimization, and data details are given in Appendix C.

Before introducing the oscillator dynamics, we verify that the learned architectures recover the structural distinction in Proposition 1. For independent $a, b \in \mathbb{R}^4$, the models predict $\frac{1}{2}a^\top b$. The joint encoder reaches held-out MSE near 0.005 at every tested width, whereas the separate architecture remains close to the train-mean predictor (Appendix C). Thus, the expected joint-versus-separate distinction is present before dynamics, coupling, or target selection are introduced.

Where to compress. Both models first predict the standardized four-coordinate future state change. Let $G(k) = L_S(k) - L_J(k)$, so positive G favors joint access, and $\Delta_G = G(4) - G(0)$. At $d = 2$, across nine seeds, $G(0) = -0.010$, $G(4) = +0.091$, and $\Delta_G = +0.101$; all nine seeds have positive $G(4)$ and Δ_G. The trained prediction heads, which do not depend on the common reader, give the same coupling-dependent change, $\Delta_G = +0.101$. Importantly, separate encoding has a small advantage at $k = 0$, showing that joint access is not uniformly beneficial. A three-seed capacity study finds the partition contrast close to zero at $d = 4$ and $d = 16$ (Appendix C), consistent with the effect weakening once the relevant local information can be retained before fusion.

What to preserve. We next keep the joint architecture fixed and change what it is trained to predict. Define relative and symmetric future-state channels,

$$Y_{\mathrm{rel}} = \Delta(q_B - q_A, p_B - p_A), \qquad Y_{\mathrm{sym}} = \Delta(q_A + q_B, p_A + p_B). \tag{14}$$

One objective predicts Y_{rel} alone, while a broader objective predicts both. Let

$$C(k, d) = L_{\mathrm{broad}\to\mathrm{rel}}(k, d) - L_{\mathrm{rel}\to\mathrm{rel}}(k, d), \tag{15}$$

so positive C measures the loss in accessibility of the relative channel under broader supervision. At $k = 4$ and $d = 2$, $C(4, 2) = +0.1661$, with relative-channel error increasing from 0.109 to 0.275. At $d = 16$, the same intervention changes error only from 0.108 to 0.109, giving $C(4, 2) - C(4, 16) = +0.1649$. Both contrasts are positive across all five seeds. At $k = 0$, the narrow representation also shows a retention cost, $C(0, 2) = +0.125$, demonstrating that competition between targets does not require physical interaction.

Figure 2 summarizes the two interventions. Coupling changes the value of joint access for a fixed prediction target, while changing the target changes what remains accessible from a fixed architecture. These controlled results establish the two forms of selectivity that we next examine in more complex dynamical systems.

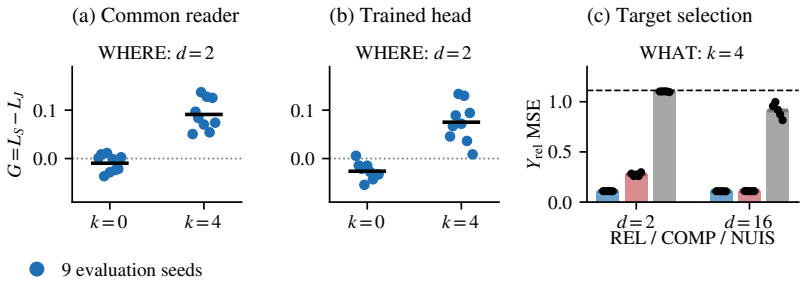

Figure 2: **Controlled partition and target effects.** (a,b) Oscillator G at $d = 2$; positive values favor joint access and black ticks denote means. Filled points are the nine evaluation seeds. (c) Relative-channel error at $k = 4$: REL, COMP, and NUIS denote the relative-only, broader, and nuisance-target conditions. All five seeds are shown; the dashed line is the training-mean predictor. Broader supervision reduces relative-channel accessibility at $d = 2$, but has minimal effect at $d = 16$.

4 Separating Relational Structure from Interaction

The structural argument in Proposition 1 depends on the target being relational, not on the two systems being dynamically coupled. We examine this distinction in two nonidentical Lorenz systems with reciprocal diffusive coupling through x (Appendix E). The systems have slightly different intrinsic parameters, and we consider both independent dynamics ($k = 0$) and weak coupling ($k = 0.5$). Histories contain three states from each system, and the prediction target is

$$Y_{\mathrm{L}} = \left[x_B - x_A,\ \dot{x}_B - \dot{x}_A\right]_{t+\tau}^{\top}, \tag{16}$$

which depends explicitly on their relative state. At total width $d = 2$, the joint encoder can construct this relation before compression, whereas the separate architecture assigns one scalar to each subsystem before fusion. Integration, data generation, and selection of the weak-coupling condition are detailed in Appendix E.

Figure 3a shows a clear joint-access advantage at weak coupling. At $k = 0.5$, the converged common reader gives $\overline{G}_{\mathrm{L}} = 0.106 \pm 0.047$, with all five seeds favoring joint access; the trained prediction heads give 0.121 ± 0.061. The joint representations predict the relation with standardized MSE 0.001–0.003, compared with 0.054–0.157 for the separate representations.

Crucially, the same effect appears without interaction. At $k = 0$, the same target gives $\overline{G}_{\mathrm{L}} = 0.120 \pm 0.041$, again with all five seeds favoring joint access. The difference between the coupled and uncoupled means is only -0.014, smaller than the variation across seeds in either condition. The joint-access advantage therefore follows the relational structure of the target rather than physical coupling in this experiment. This separates two notions that become important in the physical study below: physical interaction can define meaningful relational quantities, but relational representation does not itself require interaction.

An earlier Lorenz experiment using a generic future-state target at stronger coupling did not provide a stable partition comparison under its original evaluation protocol; we report it in Appendix F.

5 Target Selectivity in Physical Interaction

The controlled experiments separate relational structure from physical coupling. We now bring the two together and ask whether changing the prediction target changes the value of joint access while the interacting physical system is held fixed.

5.1 A stochastic double-well interaction

We consider two unit-mass particles moving in two dimensions, with $q_i = (x_i, y_i)$ and velocity v_i. Each particle evolves in a nonidentical double-well potential,

$$\begin{aligned} V_A(x, y) &= \tfrac{1}{2}(x^2 - 1)^2 + \tfrac{1}{2}y^2 + 0.1y^4, \\ V_B(x, y) &= 0.6(x^2 - 0.95^2)^2 + 0.65y^2 + 0.08y^4, \end{aligned} \tag{17}$$

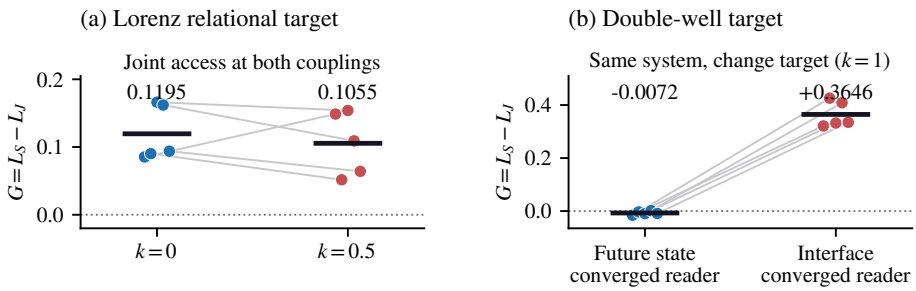

Figure 3: **Relational and target selectivity at $d = 2$.** Points show five optimization seeds, lines connect matched seeds, and black ticks denote means. (a) Lorenz relational prediction: joint access provides a similar advantage with weak coupling ($k = 0.5$) and without coupling ($k = 0$), showing that the effect follows the relational target rather than interaction itself. (b) Double-well at $k = 1$: under the same converged-reader protocol, generic future-state supervision gives essentially no joint-access advantage ($\overline{G} = -0.0072$), whereas supervision through the physical interface gives a large advantage ($\overline{G} = +0.3646$). Trained prediction heads show the same contrast (-0.0100 versus $+0.3606$).

and the pair interacts through

$$U_k(q_A, q_B) = k\left(\sqrt{1 + \|q_B - q_A\|_2^2} - 1\right). \tag{18}$$

The resulting reciprocal force is

$$f_{AB} = k\frac{q_B - q_A}{\sqrt{1 + \|q_B - q_A\|_2^2}}, \qquad f_{BA} = -f_{AB}. \tag{19}$$

The particles have different local damping and independent stochastic forcing. Histories contain three complete local states, and the future-state prediction horizon is 1.0 s. Full dynamics and data-generation details are given in Appendix G.

We first train the $d = 2$ joint and separate representations to predict the standardized eight-dimensional future state change,

$$Y_{\text{end}} = \Delta(s_A, s_B). \tag{20}$$

Both representations predict this target substantially better than the constant predictor, but joint access provides essentially no advantage. The converged common reader gives

$$\overline{G}_{\text{end}} = -0.0072 \pm 0.0067, \tag{21}$$

with mean errors 0.658 and 0.651 for the joint and separate codes, respectively, compared with 1.003 for the constant predictor. The trained prediction heads give the same result, $\overline{G}^{\text{head}}_{\text{end}} = -0.0100 \pm 0.0065$. Thus, successfully predicting the future dynamics does not by itself make early access to both systems beneficial.

5.2 Making the Physical Interface Relevant

We next hold the physical system, observations, data, architectures, representation budget, and optimization seeds fixed and change what the representation is trained to predict. The new target describes the oriented physical interface at the history anchor,

$$Y_{\text{port}}(t) = \begin{bmatrix} f_{AB}(t) \\ v_A(t) - v_B(t) \end{bmatrix} \in \mathbb{R}^4. \tag{22}$$

It combines the interaction force with the relative flow across the interface.

We first verify that the partner history carries information about this target that is not available locally. A common ridge diagnostic gives standardized test MSE 0.5895 from the best single-system history and 0.0289 from the joint history, while replacing the partner with a shuffled trajectory removes the gain. The interaction information is therefore present in the observations before learning.

Making this information relevant produces a markedly different partition effect. With the same converged-reader protocol,

$$\overline{G}_{\mathrm{port}} = +0.3646, \tag{23}$$

with all five seeds favoring joint access. Joint representations reach standardized MSE 0.156–0.186, compared with 0.503–0.582 for the separate representations. The trained prediction heads show the same contrast, $\overline{G}^{\mathrm{head}}_{\mathrm{port}} = +0.3606$, again with all five seeds positive. Figure 3b summarizes the matched future-state and interface comparisons.

Two additional controls clarify this result. First, the mean interface advantage remains 0.348 at total width $d = 4$, but falls to 0.0003 at $d = 8$, where both architectures predict the interface target nearly perfectly (Appendix G.6). This is consistent with the prediction from Section 2.4: the partition matters when target-relevant local information competes for restricted separate capacity. Second, without retraining the future-state encoders, the interface target is only modestly more accessible from the joint representation, with $\overline{G} = +0.025$. Generic future prediction therefore preserves some additional interface information under joint access, but far less than when preserving it is directly rewarded by the learning objective.

The future-state and interface experiments share the physical system, observations, architectures, and representation budget, but they are not a pure semantic target ablation. Future-state supervision predicts a stochastic change over $1.0\,\mathrm{s}$, whereas the interface target is evaluated at the history anchor. We therefore interpret the comparison as evidence of target-relative representation selectivity without attributing the entire difference to one aspect of the prediction problem.

As a final relational control, at $k = 0$ the interaction force vanishes while relative flow remains relational. Joint access still helps for this flow-only target ($\overline{G} = +0.170$), reinforcing the Lorenz result that relational structure and physical coupling are distinct.

5.3 A BOUNDARY: PHYSICAL RELEVANCE IS TARGET-DEPENDENT

The interface result does not imply that any physically meaningful interaction target will preserve the information required downstream. We examine this boundary in a planar pushing system, where the representation is trained on signed scalar work-like quantities obtained by pairing interface force with contact-point motion. Directional object motion remains predictable from the raw history (standardized MSE 0.302) and from the joint encoder's recurrent state (0.296), but is no longer accessible from the deployed work-supervised representation (1.045, compared with 1.047 for the constant predictor).

This experiment is a boundary case rather than a controlled comparison between oriented and scalar interaction targets: the work signal is sparse, training is contrastive, and the evaluated representation stages differ in dimension. Its implication is therefore narrower. Physical structure provides meaningful candidates for what a representation should preserve, but physical meaning alone does not determine which information will be sufficient for a downstream task. The complete contact experiment and diagnostics are reported in Appendix D.

Taken together, these experiments expose a distinction between *access* to interaction information and *selection* of that information by the learning objective. In the double-well system, the partner history contains substantial information about the physical interface, yet generic future-state prediction provides essentially no advantage to early joint access. When the interface itself becomes relevant, where compression occurs matters strongly. Under restricted representation capacity, learning interaction-aware representations therefore depends not only on which information the architecture can access, but also on which relations the prediction objective rewards for preserving.

6 RELATED WORK

Relational and object-centric dynamics. Interaction Networks, neural relational inference, relational potentials, and graph simulators make interactions explicit in learned dynamics (Battaglia et al., 2016; Kipf et al., 2018; Comas et al., 2023; Sanchez-Gonzalez et al., 2020); related structure appears in manipulation and contact models (Ye et al., 2020; Tekden et al., 2024; Mosbach et al., 2025). These works ask how to represent interactions. We ask when a cross-system relation must be formed relative to restrictive compression.

Distributed compression and synergy. The information bottleneck makes relevance target-dependent (Tishby et al., 1999; Alemi et al., 2017). Distributed information bottlenecks study multiple encoders that separately compress observations while collectively retaining target information (Estella Aguerri & Zaidi, 2018); coding for computing and distributed algebraic computation ask what separated sources must communicate to recover a function (Orlitsky & Roche, 2001; Abelson, 1980; Luo & Tsitsiklis, 1993). Partial information decomposition provides a complementary language for redundant, unique, and synergistic information from multiple sources (Williams & Beer, 2010). Our setting compares separate compression with an encoder that can form the relation before compression.

Predictive representations and physical interaction. World models and predictive objectives determine which future distinctions are retained (Hafner et al., 2019; 2025; van den Oord et al., 2018; Yan et al., 2021; Assran et al., 2023; Bardes et al., 2024); state-predictive and control-oriented representations make this dependence explicit (Federici et al., 2024; Wang & Tiwary, 2021; Gelada et al., 2019; Zhang et al., 2021). In robotics, the same architectural question appears when object, arm, tactile, visual, or proprioceptive streams are encoded separately before interaction. Port-Hamiltonian models provide structured interface variables through effort–flow pairs (van der Schaft & Maschke, 2013), while contact-learning work incorporates geometry and interaction structure (Pfrommer et al., 2020; Bianchini et al., 2023; Zhang et al., 2026). We use these physical quantities as controlled targets rather than as a universal prescription.

7 Discussion and Conclusion

Our results point to a distinction that is easy to overlook when designing representations for interacting systems: making information available to a model is not the same as making that information relevant to what the model learns to preserve. In the double-well system, the partner history contains substantial information about the physical interface, and both representations predict future states substantially better than a constant baseline. Yet early joint access provides essentially no advantage under this generic predictive objective. When the interface itself becomes the prediction target, the same architectural comparison produces a large and consistent effect. The choice of learning objective can therefore determine whether interaction structure survives compression, even when that structure is already present in the observations.

This changes how we should think about fusion in learned models. Early versus late fusion is often treated primarily as an architectural choice. Our results show that its value cannot be assessed independently of the target and the representation budget. Relational targets favor joint access even without physical coupling, while increasing separate capacity removes the advantage once the relevant local information can survive compression. The relevant design object is therefore not the fusion architecture alone, but the combination of *where* relations can be formed, *what* the objective rewards the representation for preserving, and *how much* information can be retained before fusion.

This perspective is particularly relevant for embodied learning. Robot world models increasingly combine information across objects, arms, sensory modalities, and contacts, while being expected to support many downstream predictions and actions. Simply scaling a predictive model or exposing it to all available observations does not guarantee that interaction variables needed downstream will remain accessible in its representation. For bimanual manipulation, multi-object prediction, or multimodal contact reasoning, the important question is which relations should be constructed before local information is compressed and which can safely be reconstructed afterwards. As contacts and tasks change, these computational boundaries may need to change with them.

Our experiments deliberately use controlled systems in which partition, target, coupling, and capacity can be varied independently. This allows us to identify representation selectivity rather than conflate it with model scale or task complexity. The present results establish the phenomenon for width-constrained representations and do not imply an information-rate advantage once local sufficient statistics can already be retained. Extending the same question to learned object representations, high-dimensional multimodal observations, and adaptive interaction graphs is an important next step. More broadly, representation learning for embodied systems may need to determine not only *what* to compress, but also *when information should interact before it is compressed*.

AI USE STATEMENT

Generative AI tools assisted with hypothesis formulation, theoretical modeling and review of mathematical claims and proofs, experimental design and implementation, analysis and interpretation, literature discovery, figure production, manuscript drafting and editing, and scientific auditing. The authors are responsible for checking the equations, data, citations, and claims and for the final submission.

ETHICS STATEMENT

The controlled experiments use simulated dynamical systems and a simulated planar contact environment. No human-participant data are reported. Physical deployment would require additional evaluation of failure modes in contact estimation and control.

REPRODUCIBILITY STATEMENT

The theoretical derivations and proof are in Appendices A and B. Appendix C specifies the controlled oscillator protocol; Appendix D gives the contact boundary protocol; Appendices E and F document the Lorenz studies; Appendix G documents the double-well future-state/interface intervention and capacity relaxation; and Appendix H collects common evaluation and reproducibility conventions. Figures report saved experiment results.

REFERENCES

Harold Abelson. Lower bounds on information transfer in distributed computations. *Journal of the ACM*, 27(2):384–392, 1980. doi: 10.1145/322186.322200. URL https://doi.org/10.1145/322186.322200.

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations*, 2017. URL https://arxiv.org/abs/1612.00410.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15619–15629, 2023. URL https://arxiv.org/abs/2301.08243.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=QaCCuDfBk2.

Peter W. Battaglia, Razvan Pascanu, Matthew Lai, Danilo Rezende, and Koray Kavukcuoglu. Interaction networks for learning about objects, relations and physics. In *Advances in Neural Information Processing Systems*, 2016. URL https://arxiv.org/abs/1612.00222.

Bibit Bianchini, Mathew Halm, and Michael Posa. Simultaneous learning of contact and continuous dynamics. In *Proceedings of the 7th Conference on Robot Learning*, pp. 3966–3978, 2023. URL https://proceedings.mlr.press/v229/bianchini23a.html.

Armand Comas, Yilun Du, Christian Fernandez Lopez, Sandesh Ghimire, Mario Sznaier, Joshua B. Tenenbaum, and Octavia Camps. Inferring relational potentials in interacting systems. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 6364–6383, 2023. URL https://proceedings.mlr.press/v202/comas23a.html.

Iñaki Estella Aguerri and Abdellatif Zaidi. Distributed information bottleneck method for discrete and Gaussian sources. In *International Zurich Seminar on Information and Communication (IZS)*, 2018. doi: 10.3929/ethz-b-000245048. URL https://doi.org/10.3929/ethz-b-000245048.

Marco Federici, Patrick Forré, Ryota Tomioka, and Bastiaan S. Veeling. Latent representation and simulation of markov processes via time-lagged information bottleneck. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2309.07200.

Carles Gelada, Saurabh Kumar, Jacob Buckman, Ofir Nachum, and Marc G. Bellemare. Deepmdp: Learning continuous latent space models for representation learning. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 2170–2179, 2019. URL https://proceedings.mlr.press/v97/gelada19a.html.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *Proceedings of the 36th International Conference on Machine Learning*, 2019. URL https://proceedings.mlr.press/v97/hafner19a.html.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640:647–653, 2025. doi: 10.1038/s41586-025-08744-2. URL https://doi.org/10.1038/s41586-025-08744-2.

Allen Hatcher. *Algebraic Topology*. Cambridge University Press, 2002. URL https://pi.math.cornell.edu/~hatcher/AT/AT.pdf.

Thomas Kipf, Ethan Fetaya, Kuan-Chieh Wang, Max Welling, and Richard Zemel. Neural relational inference for interacting systems. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2688–2697. PMLR, 2018. URL https://proceedings.mlr.press/v80/kipf18a.html.

Zhi-Quan Luo and John N. Tsitsiklis. On the communication complexity of distributed algebraic computation. *Journal of the ACM*, 40(5):1019–1047, 1993. doi: 10.1145/174147.174149.

Malte Mosbach, Jan Niklas Ewertz, Angel Villar-Corrales, and Sven Behnke. SOLD: Slot object-centric latent dynamics models for relational manipulation learning from pixels. In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 44911–44935, 2025. URL https://proceedings.mlr.press/v267/mosbach25a.html.

Alon Orlitsky and James R. Roche. Coding for computing. *IEEE Transactions on Information Theory*, 47(3):903–917, 2001. doi: 10.1109/18.915643.

Samuel Pfrommer, Mathew Halm, and Michael Posa. Contactnets: Learning discontinuous contact dynamics with smooth, implicit representations. In *Proceedings of the Conference on Robot Learning*, 2020. URL https://arxiv.org/abs/2009.11193.

Alvaro Sanchez-Gonzalez, Jonathan Godwin, Tobias Pfaff, Rex Ying, Jure Leskovec, and Peter W. Battaglia. Learning to simulate complex physics with graph networks. In *Proceedings of the 37th International Conference on Machine Learning*, 2020. URL https://arxiv.org/abs/2002.09405.

Ahmet Ercan Tekden, Aykut Erdem, Erkut Erdem, Tamim Asfour, and Emre Ugur. Object and relation centric representations for push effect prediction. *Robotics and Autonomous Systems*, 174:104632, 2024. doi: 10.1016/j.robot.2024.104632. URL https://arxiv.org/abs/2102.02100.

Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. *Proceedings of the 37th Allerton Conference on Communication, Control and Computing*, 1999. URL https://arxiv.org/abs/physics/0004057.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. URL https://arxiv.org/abs/1807.03748.

A. J. van der Schaft and B. M. Maschke. Port-Hamiltonian systems on graphs. *SIAM Journal on Control and Optimization*, 51(2):906–937, 2013. doi: 10.1137/110840091. URL https://doi.org/10.1137/110840091.

Dedi Wang and Pratyush Tiwary. State predictive information bottleneck. *The Journal of Chemical Physics*, 154:134111, 2021. doi: 10.1063/5.0038198. URL https://arxiv.org/abs/2011.10127.

Paul L. Williams and Randall D. Beer. Nonnegative decomposition of multivariate information. *arXiv preprint arXiv:1004.2515*, 2010.

Wilson Yan, Ashwin Vangipuram, Pieter Abbeel, and Lerrel Pinto. Learning predictive representations for deformable objects using contrastive estimation. In *Proceedings of the 2020 Conference on Robot Learning*, pp. 564–574, 2021. URL https://proceedings.mlr.press/v155/yan21a.html.

Yufei Ye, Dhiraj Gandhi, Abhinav Gupta, and Shubham Tulsiani. Object-centric forward modeling for model predictive control. In *Proceedings of the Conference on Robot Learning*, pp. 100–109, 2020. URL https://proceedings.mlr.press/v100/ye20a.html.

Amy Zhang, Rowan McAllister, Roberto Calandra, Yarin Gal, and Sergey Levine. Learning invariant representations for reinforcement learning without reconstruction. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2006.10742.

Zhiyuan Zhang, Pokuang Zhou, Kaidi Zhang, Adeesh Desai, Temitope Amosa, Davood Soleymanzadeh, Jiuzhou Lei, Minghui Zheng, and Yu She. Contactworld: What representations matter in vision-tactile world models for contact-rich manipulation, 2026. URL https://arxiv.org/abs/2606.13877.

A TARGET-RELATIVE INFORMATION DECOMPOSITION

A.1 DERIVATION OF EQ. (8)

Recall the setting of Section 2.2. Let

$$Y = \mathcal{G}X + \epsilon, \tag{24}$$

where ϵ is independent of (X, H, ξ), and let Z be generated causally from the observed history H and encoder randomness ξ. Define the target-predictive statistic

$$S_Y = \mathbb{E}[Y \mid X] \tag{25}$$

in nonredundant coordinates. In the linear-Gaussian construction, let

$$X_\perp = X - \mathbb{E}[X \mid S_Y]. \tag{26}$$

Then $(S_Y, X_\perp)$ determines X, and Gaussian regression gives $X_\perp \perp S_Y$. Hence

$$I(X; Z) = I(S_Y; Z) + I(X_\perp; Z \mid S_Y). \tag{27}$$

Since $Y = S_Y + \epsilon$ and the innovation is independent of the observed history and encoder randomness,

$$Y \perp Z \mid S_Y. \tag{28}$$

The mutual-information chain rule therefore gives

$$I(Z; S_Y) - I(Z; Y) = I(Z; S_Y \mid Y) - I(Z; Y \mid S_Y) \tag{29}$$

$$= I(Z; S_Y \mid Y). \tag{30}$$

Using $I(H; Z) = I(X; Z)$ and substituting Eq. (30) into Eq. (27) yields

$$I(H; Z) - I(Z; Y) = I(Z; S_Y \mid Y) + I(Z; X_\perp \mid S_Y), \tag{31}$$

which is Eq. (8). Both terms are nonnegative conditional mutual informations.

A sufficient condition for $I(H; Z) = I(X; Z)$ is that X is a measurable function of H and the encoder channel factors through X,

$$q(z \mid h) = q(z \mid X(h)). \tag{32}$$

The decomposition itself assumes the information equality directly; the neural experiments do not require learned encoders to satisfy this factorization.

A.2 REPRESENTATION WIDTH AND INFORMATION RATE

Equation (8) is stated for finite information quantities. For continuous inputs, a deterministic real-valued representation may have infinite mutual information with its input even when its dimension is small. Representation width should therefore not be identified with an information rate.

The neural experiments instead use width as an architectural constraint. Their purpose is to test whether changing the target or the location of compression changes what remains accessible when representational capacity is restricted. A noisy or finite-resolution channel would be required to turn this architectural constraint into an explicit information-rate constraint.

A.3 LINEAR DYNAMICAL REALIZATION

For a linear-Gaussian dynamical system,

$$X_{t+\Delta} = A_\Delta X_t + \eta_\Delta, \tag{33}$$

with future innovation η_Δ independent of the observed history, consider

$$Y = CX_{t+\Delta} = CA_\Delta X_t + C\eta_\Delta. \tag{34}$$

Then

$$\mathcal{G} = CA_\Delta, \qquad \epsilon = C\eta_\Delta \tag{35}$$

satisfy the assumptions above. This provides the direct dynamical interpretation used in Section 2.2: changing C, and hence the prediction target, changes the target-predictive statistic S_Y and therefore which retained information is relevant.

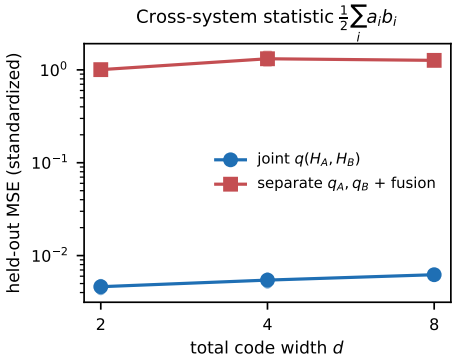

Figure 4: **Learned structural control.** Held-out MSE for the cross-system target $\frac{1}{2}a^\top b$ across three optimization seeds. The joint encoder forms the scalar relation before compression; the separate architecture approaches the same error once each branch has sufficient width to retain its four-dimensional local input.

B Joint Access under Separate Compression

B.1 Proof of Proposition 1

For

$$Y(a, b) = a^\top b, \tag{36}$$

a one-dimensional joint representation can output Y directly.

Suppose instead that continuous messages $E_A(a)$ and $E_B(b)$ allow a fusion function to recover Y exactly for every $a, b \in \mathbb{R}^m$. If

$$E_A(a_1) = E_A(a_2) \tag{37}$$

for distinct a_1, a_2, choose $b = a_1 - a_2$. The fused output must then be identical for the two inputs, although

$$Y(a_1, b) - Y(a_2, b) = (a_1 - a_2)^\top b \tag{38}$$

$$= \|a_1 - a_2\|_2^2 > 0, \tag{39}$$

a contradiction. Hence E_A is injective. Exchanging the roles of a and b proves that E_B is also injective.

It remains to relate injectivity to representation dimension. By invariance of domain, an embedding of an open subset of $\mathbb{R}^m$ into $\mathbb{R}^m$ has open image (Hatcher, 2002, Theorem 2B.3). Suppose $d_A < m$ and compose

$$E_A : \mathbb{R}^m \to \mathbb{R}^{d_A} \tag{40}$$

with the coordinate inclusion $\mathbb{R}^{d_A} \hookrightarrow \mathbb{R}^m$. Restrict the resulting continuous injective map to a closed ball. Compactness makes the restriction an embedding, and its restriction to the interior is therefore an embedding of an open subset of $\mathbb{R}^m$. Its image lies in a proper coordinate subspace and cannot be open, a contradiction. Thus $d_A \geq m$, and the same argument gives $d_B \geq m$.

The proposition concerns continuous representations, exact recovery, and the full input domain. It does not provide a lower bound on approximation error for a finite data distribution or identify real-valued dimension with communication rate.

B.2 Learned structural control

The structural control uses $m = 4$ and the rescaled target

$$Y = \tfrac{1}{2}a^\top b. \tag{41}$$

The scale factor does not affect Proposition 1. At total widths $d = 2$ and $d = 4$, each branch of the separate architecture has fewer than four coordinates. At $d = 8$, each branch has four coordinates and can in principle retain its complete local input.

The learned models recover the qualitative transition predicted by the proposition and use the same encoder families as the dynamical experiments. The experiment is an architectural control rather than an estimate of the exact-recovery bound.

C Controlled Oscillator: Experimental Details

C.1 Environment and data

The oscillator dynamics are given in Eq. (13), with

$$\omega_A = 1.0, \qquad \omega_B = 1.15, \qquad \alpha = 0.30, \qquad \gamma = 0.18, \qquad \sigma = 0.8, \tag{42}$$

and coupling

$$k \in \{0, 4\}. \tag{43}$$

The two oscillators are driven by independent Wiener processes.

Episodes last 6 s and are integrated with Euler–Maruyama at 0.01 s. States are logged every second integration step, giving a 0.02 s sampling interval. Initial positions are sampled independently from $\mathrm{Unif}[-1, 1]$ and initial momenta from $\mathrm{Unif}[-0.6, 0.6]$. Parameters are fixed within each coupling condition; only initialization and process noise vary across episodes.

For each coupling level, the data comprise 400 training, 100 selection, and 200 held-out episodes. All windows from one episode remain in the same split. Each trajectory contains 301 logged states. For anchors $t = 2, \ldots, 290$, the input history is (X_{t-2}, X_{t-1}, X_t), the ten-step future packet is $(X_{t+1}, \ldots, X_{t+10})$, and the future-state target is $X_{t+10} - X_t$. This gives 289 windows per episode and a prediction horizon of 0.20 s.

Oscillator states occupy four channels of the common 13-channel input format; unused channels are zero. Input channels are standardized using training histories only, with population standard deviation and denominator $\mathrm{sd} + 10^{-6}$. Constant channels are set to zero, and the same statistics are applied to selection and held-out data.

C.2 Representation architectures

The joint encoder observes both subsystem histories. A two-layer per-step MLP feeds a two-layer GRU with 128 hidden units, followed by a linear projection to representation width d.

The separate architecture uses two recurrent encoders, each with 91 hidden units and projection to width $d/2$, followed by the common fusion network

$$g : \mathbb{R}^d \to \mathbb{R}^{64} \to \mathbb{R}^d. \tag{44}$$

The per-step MLP is $13 \to 128 \to 128$ for the joint encoder and $6 \to 91 \to 91$ or $9 \to 91 \to 91$ for the two separate branches. Linear layers use ReLU activations. GRUs are two-layer, unidirectional, and initialized with zero hidden state. No dropout or normalization is used in the encoders.

The deployed representation parameter counts are $217{,}028$ (joint) and $219{,}725$ (separate) at $d = 2$; $217{,}544$ and $220{,}167$ at $d = 4$; and $220{,}640$ and $222{,}819$ at $d = 16$. The compared encoder families therefore differ by at most approximately 1.3% in parameter count.

The regression head is $d \to 64 \to m$ with a ReLU hidden layer, where $m = 4$ for future-state and COMP prediction and $m = 2$ for the single-channel objectives.

C.3 Prediction objectives

For the partition comparison, the target is the standardized four-coordinate state change over the 0.20 s horizon.

For the target-selection comparison, the relative and symmetric channels are defined in Eq. (14). The REL objective predicts Y_{rel} alone, COMP predicts $(Y_{\mathrm{rel}}, Y_{\mathrm{sym}})$, and the symmetric-only condition predicts Y_{sym} as a diagnostic. Target coordinates are standardized using training data only.

For COMP, the two relative and two symmetric coordinates are standardized separately before concatenation and receive equal coordinate weight in the MSE, so the two channels contribute equally to the objective.

The development stage analysis additionally considers regression on the ten-step future packet and a contrastive objective. The future packet contains all four oscillator coordinates at ten future times,

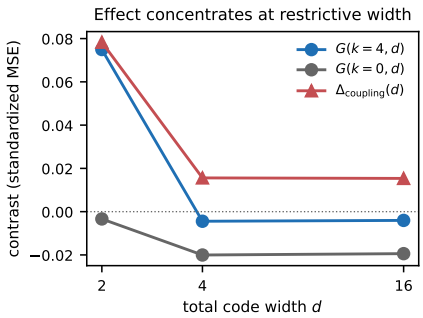

Figure 5: **Oscillator capacity diagnostic.** Three-seed means of the partition contrast $G(k, d)$ and coupling contrast $G(4, d) - G(0, d)$ under future-state prediction. The joint-access advantage at $k = 4$ is concentrated at the narrowest representation; the width study is descriptive.

flattened to 40 dimensions. For the contrastive model, the query embedding $\phi([r, u])$ and packet embedding $\psi(F)$ have width 32, with $u = 0$. For a batch of $N = 400$ distinct episodes, one anchor per episode defines the positive pair and the remaining 399 packets act as negatives. With $s_{ij} = \phi([r_i, u_i])^\top \psi(F_j)$, the loss is

$$\frac{1}{N} \sum_{i=1}^{N} \left[\log \sum_{j=1}^{N} e^{s_{ij}} - s_{ii} \right] - \log N. \tag{45}$$

No temperature or logit normalization is used.

C.4 Optimization and evaluation

All oscillator encoders are trained with Adam at learning rate 10^{-3} for 10,000 updates. Each update uses all 400 training episodes with one uniformly sampled anchor per episode. Adam uses $(\beta_1, \beta_2) = (0.9, 0.999)$, $\epsilon = 10^{-8}$, zero weight decay, and no AMSGrad, learning-rate schedule, or gradient clipping. Reported representations use the final checkpoint.

The common frozen-code reader is a two-hidden-layer MLP with widths $128, 128$, ReLU activations, and dropout 0.1 during fitting. It uses Adam at learning rate 10^{-3}, weight decay 10^{-4}, and 100 full-batch updates. Reader inputs and targets are standardized using training data only.

For the common-reader measure, squared error is averaged over output coordinates and all 289 windows within each of the 200 held-out episodes, then across episodes. The trained-head measure uses eight fixed held-out monitor batches, each containing one anchor per episode. Both use the final encoder checkpoint.

For the partition comparison at $d = 2$, nine evaluation seeds 33004–33012 are paired across $k = 0, 4$. The tests of $G(4) > 0$ and $G(4) - G(0) > 0$ are one-sided Student t tests across the nine paired seed contrasts. The earlier seeds 33001–33003 and the width study at $d = 4, 16$ are descriptive. The target-selection comparison uses five seeds 33013–33017; tests of $C(4, 2) > 0$ and $C(4, 2) - C(4, 16) > 0$ are paired by seed and one-sided. The symmetric-only condition and $k = 0$ target-selection contrasts are descriptive.

D Planar Contact Boundary Study

D.1 Environment and interaction target

The contact study considers an actuated cylindrical body A pushing a passive box B across a frictional support. The task is planar, while MuJoCo 3.9.0 retains full constrained rigid-body contact dynamics. Body A moves through two horizontal slide joints with direct-force actuators bounded at ± 8 N. The cylinder has mass 0.35 kg and radius 0.025 m; the box half-sizes are $(0.05, 0.025, 0.025)$ m. Gravity is $(0, 0, -9.81)$ m s^{-2}. The cylinder–box and box–support pairs use three-dimensional contact with an elliptic friction cone, `solref` $(0.01, 1)$, and `solimp` $(0.9, 0.95, 0.001)$. The `implicitfast` integrator uses a 0.00025 s physical step and 30 solver iterations.

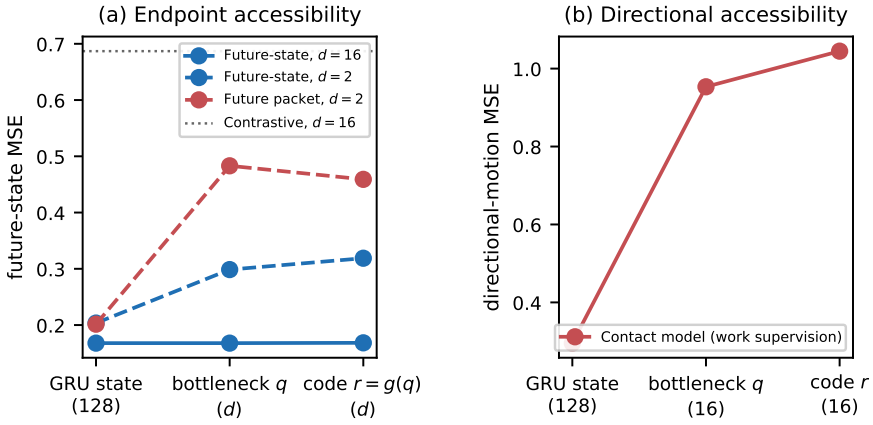

Figure 6: **Accessibility across representation stages.** (a) Oscillator future-state error from the joint recurrent state h, projected state q, and deployed representation r under future-state, future-packet, and contrastive supervision. (b) For the contact model, directional velocity-change MSE is 0.296 from the recurrent state, 0.953 from q, and 1.045 from r. Stage dimensions and decoder parameter counts differ, so these measurements localize where accessibility changes without isolating its cause.

For each episode, box mass is sampled uniformly from $[0.8, 1.2)$ kg, support friction from $[0.25, 0.45)$, and cylinder–box friction from $[0.35, 0.65)$. Initial yaw is sampled from $[-\pi, \pi)$, the lateral offset from $[-0.65, 0.65)$ times 0.025 m, the initial gap from $[0.006, 0.025)$ m, and command magnitude from $[1.4, 2.6)$ N. After a 0.20 s settling period, each episode runs for 1.20 s with commands held over 0.02 s intervals. Four command programs generate straight, paused, oblique, and reversing pushes; their complete specification is included with the released environment.

Rather than retaining the oriented force–relative-flow pair used in Section 5.2, the contact representation is trained on scalar force–motion pairings. For interface force f_{AB} and contact-point displacements $\Delta r_A, \Delta r_B$, we define

$$T_t = \tfrac{1}{2} f_{AB}^{\top}(\Delta r_A + \Delta r_B), \qquad R_t = f_{AB}^{\top}(\Delta r_A - \Delta r_B). \tag{46}$$

These quantities describe signed work-like exchange across the interface without retaining the complete orientation of force and motion.

Equation (46) is accumulated from the physical solver steps within each 0.02 s bin. For each cylinder–box contact, the solved wrench is oriented as the force f_{AB} on B due to A. If v_A^-, v_A^+ and v_B^-, v_B^+ denote the material contact-point velocities immediately before and after a physical step, we approximate

$$\Delta r_A = \tfrac{1}{2}(v_A^- + v_A^+)h, \qquad \Delta r_B = \tfrac{1}{2}(v_B^- + v_B^+)h, \tag{47}$$

with $h = 0.00025$ s. The corresponding work contributions are

$$W_A = -f_{AB}^{\top}\Delta r_A, \qquad W_B = f_{AB}^{\top}\Delta r_B. \tag{48}$$

After summing over contacts and physical steps within the bin,

$$T = (W_B - W_A)/2, \qquad R = -(W_A + W_B). \tag{49}$$

Bins without cylinder–box contact have $T = R = 0$.

The resulting target is sparse. Figure 7a counts a window as small when every packet component has magnitude below 10^{-4} J; 75% of the training windows satisfy this criterion. The reported correlation is computed over paired active bins satisfying $|T| > 10^{-6}$ J and $|R| > 10^{-6}$ J.

D.2 Data and representation learning

The dataset contains 400 training, 100 selection, and 200 held-out episodes. The 11 observed state coordinates comprise the planar position and velocity of A, together with the planar position, sine and cosine of yaw, planar velocity, and yaw rate of B. Two past command coordinates are appended

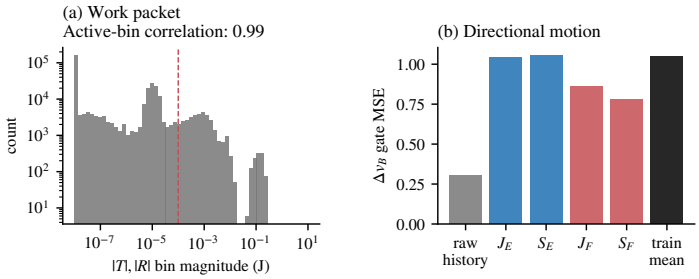

Figure 7: **Contact boundary measurements.** (a) Magnitude distribution of the interface work-like packet. (b) Standardized MSE for directional velocity-change prediction from raw history, intermediate representation stages, the deployed work-supervised representations, and the constant predictor. Directional information is accessible from the history and recurrent state but not from the deployed representation under this supervision. J and S denote joint and separate encoding; subscripts E and F denote supervision with the ten-bin interface-work packet and the ten-step future-observation packet, respectively. Thus J_E, S_E are work-supervised and J_F, S_F are future-observation-supervised representations.

at each time step, giving 13 input channels for the joint encoder. The two separate branches receive their corresponding local state coordinates together with the same commands.

For anchors $n = 10, \ldots, 50$, the input contains observations $n-9, \ldots, n$ and commands $n-10, \ldots, n-1$. The future packet spans bins $n, \ldots, n+9$, yielding 41 windows per episode. Windows are retained regardless of whether contact occurs.

The representation is trained contrastively on the 20-coordinate packet obtained by interleaving T and R over ten future bins. Inputs, future commands, and work channels are standardized using training data only. The joint encoder uses a $13 \to 128 \to 128$ per-step MLP, a two-layer 128-unit GRU, and a $128 \to 16$ projection. The separate branches use 6 and 9 input channels, respectively, two-layer 91-unit GRUs, and eight projected coordinates each, followed by the common $16 \to 64 \to 16$ fusion network.

The query head embeds the 16-coordinate representation together with 20 future-command coordinates using $36 \to 128 \to 128 \to 32$ layers; the packet head uses $20 \to 128 \to 128 \to 32$. Training minimizes dot-product InfoNCE with one anchor from each of 400 distinct training episodes and the other 399 episodes as negatives. No temperature or logit normalization is used. The encoder is trained for 10,000 Adam updates at learning rate 10^{-3}.

D.3 Directional-motion diagnostic

We evaluate whether the learned representation preserves the passive object's directional velocity change over 0.20 s. The two target coordinates are standardized using the training split. Downstream decoders do not receive future commands or the contrastive target embedding.

Directional velocity change is predictable directly from the observed history, with standardized MSE 0.302. For the joint work-supervised encoder, the 128-coordinate recurrent state gives MSE 0.296, the 16-coordinate projection gives 0.953, and the deployed representation gives 1.045, compared with 1.047 for the constant predictor (Figures 7 and 6). Thus the directional information remains accessible before projection but is largely absent from the deployed representation under this supervision.

These measurements use one trained contact model rather than independent encoder-training seeds. In addition, representation dimension and decoder input dimension differ across stages. The experiment therefore localizes where directional accessibility changes, but does not isolate scalarization, sparsity, or projection as its unique cause.

E LORENZ RELATIONAL STUDY

E.1 ENVIRONMENT AND DATA

We use two nonidentical Lorenz systems with reciprocal diffusive coupling through the x coordinates,

$$\dot{x}_A = 10(y_A - x_A) + k(x_B - x_A), \tag{50}$$
$$\dot{y}_A = x_A(28 - z_A) - y_A, \tag{51}$$
$$\dot{z}_A = x_A y_A - \tfrac{8}{3} z_A, \tag{52}$$
$$\dot{x}_B = 10.5(y_B - x_B) + k(x_A - x_B), \tag{53}$$
$$\dot{y}_B = x_B(29 - z_B) - y_B, \tag{54}$$
$$\dot{z}_B = x_B y_B - 2.8 z_B. \tag{55}$$

The two systems therefore have distinct intrinsic dynamics even at $k = 0$.

Initial conditions are sampled independently for each subsystem,

$$x_i(0), y_i(0) \sim \mathcal{U}(-15, 15), \qquad z_i(0) \sim \mathcal{U}(5, 35). \tag{56}$$

The coupled six-dimensional system is integrated simultaneously using classical RK4 with step 0.002. We discard the first 20 time units and retain 16 time units, sampled every 0.02.

Each input contains three consecutive complete states from each subsystem. The relational target is

$$Y_{\mathrm{L}} = \left[x_B - x_A,\ \dot{x}_B - \dot{x}_A\right]^{\top}_{t+\tau}, \tag{57}$$

with prediction horizon $\tau = 0.20$ time units. The target is relational but is not interpreted as a mechanical effort–flow port.

We compare independent dynamics, $k = 0$, with weak reciprocal coupling, $k = 0.5$. The weak-coupling condition was selected before representation training using a partner-information diagnostic that compared prediction from each local history, the true joint history, and a same-age shuffled partner. The purpose of this diagnostic was to ensure that the coupled condition retained trajectory-specific information from both systems rather than entering a strongly redundant regime.

E.2 REPRESENTATION COMPARISON

The total representation width is $d = 2$. The joint encoder observes both histories before projection. The separate architecture allocates one coordinate to each subsystem before fusion. Five optimization seeds are used for both coupling conditions.

At $k = 0.5$, the converged common reader gives

$$\overline{G}_{\mathrm{L}} = 0.106 \pm 0.047, \tag{58}$$

with all five seeds favoring joint access. The five seed-level contrasts are $G = \{0.1486, 0.0516, 0.1539, 0.1090, 0.0640\}$. The corresponding trained-head contrast is 0.121 ± 0.061. Joint representations reach standardized MSE 0.001–0.003, compared with 0.054–0.157 for the separate representations.

At $k = 0$, the same relational target gives

$$\overline{G}_{\mathrm{L}} = 0.120 \pm 0.041, \tag{59}$$

again with all five seeds favoring joint access. The five seed-level contrasts are $G = \{0.0853, 0.0902, 0.1661, 0.1619, 0.0938\}$. The difference between the coupled and uncoupled mean contrasts is -0.014, smaller than the seed-to-seed variation in either condition. This is the basis for the conclusion in Section 4 that the partition effect follows the relational target rather than physical coupling in this experiment.

F STRONG-COUPLING LORENZ FUTURE-STATE STUDY

Before the relational study above, we evaluated the same architectural comparison using a six-dimensional future-state-change target at $k = 0$ and $k = 2$. The numerical result was reproducible, but subsequent diagnostics showed that this experiment did not provide a reliable comparison of the representations.

Three effects were relevant. First, the original common reader was still improving at its 100-update cutoff; extending reader optimization reduced the seed-level spread substantially. Second, encoder optimization variance was large relative to the observed partition contrast. Third, at $k = 2$ the diffusive coupling makes the two Lorenz trajectories substantially more redundant, reducing the amount of partner-specific information available to distinguish the architectures.

For these reasons, we do not use the strong-coupling future-state experiment as evidence for or against the partition effect. We retain it here because it motivated the explicit partner-information control and converged-reader evaluation used in the subsequent experiments.

G DOUBLE-WELL PHYSICAL INTERACTION STUDY

G.1 ENVIRONMENT AND DATA

Each subsystem has state

$$s_i = (x_i, y_i, v_{x,i}, v_{y,i}), \qquad q_i = (x_i, y_i), \tag{60}$$

with local potentials

$$V_A(x, y) = 0.5(x^2 - 1)^2 + 0.5y^2 + 0.1y^4, \tag{61}$$

$$V_B(x, y) = 0.6(x^2 - 0.95^2)^2 + 0.65y^2 + 0.08y^4. \tag{62}$$

The interaction potential is

$$U_k(q_A, q_B) = k\left(\sqrt{1 + \|q_B - q_A\|_2^2} - 1\right), \tag{63}$$

giving the reciprocal force

$$f_{AB} = k\frac{q_B - q_A}{\sqrt{1 + \|q_B - q_A\|_2^2}}, \qquad f_{BA} = -f_{AB}. \tag{64}$$

The continuous-time stochastic dynamics are

$$dq_A = v_A\, dt, \tag{65}$$

$$dv_A = [-\nabla V_A(q_A) + f_{AB} - 0.8v_A]\, dt + \sqrt{0.96}\, dW_A, \tag{66}$$

$$dq_B = v_B\, dt, \tag{67}$$

$$dv_B = [-\nabla V_B(q_B) - f_{AB} - v_B]\, dt + \sqrt{1.2}\, dW_B, \tag{68}$$

where the two-dimensional Wiener processes W_A and W_B are independent.

We integrate the dynamics using BAOAB splitting with step $h = 0.005$. Each trajectory discards a 20-time-unit burn-in and retains 60 time units sampled every 0.05. Initial positions are drawn near the two wells,

$$x_i(0) = \varepsilon_i s_i + \eta_i, \qquad \Pr(\varepsilon_i = \pm 1) = \tfrac{1}{2}, \qquad \eta_i \sim \mathcal{N}(0, 0.1^2), \tag{69}$$

with

$$y_i(0) \sim \mathcal{N}(0, 0.3^2), \qquad v_{x,i}(0), v_{y,i}(0) \sim \mathcal{N}(0, 0.6). \tag{70}$$

Here the second argument denotes the variance. Here the well locations are $s_A = 1$ and $s_B = 0.95$.

Histories contain three consecutive complete local states. For future-state prediction, the target is the concatenated state change over $\tau = 1.0$,

$$Y_{\text{end}}(t) = \begin{bmatrix} s_A(t + \tau) - s_A(t) \\ s_B(t + \tau) - s_B(t) \end{bmatrix}. \tag{71}$$

Table 1: **Reader-matched future-state evaluation at $k = 1, d = 2$.** Standardized held-out MSE with $G = L_S - L_J$.

	Converged reader			Trained head		
seed	L_J	L_S	G	L_J	L_S	G
11	0.6593	0.6436	-0.0157	0.6629	0.6440	-0.0189
12	0.6638	0.6605	-0.0033	0.6666	0.6587	-0.0079
13	0.6518	0.6421	-0.0097	0.6526	0.6411	-0.0116
14	0.6580	0.6599	+0.0019	0.6595	0.6586	-0.0009
15	0.6583	0.6493	-0.0090	0.6589	0.6483	-0.0106
mean G			-0.0072			-0.0100
sd(G)			0.0067			0.0065

For interface supervision, the target at the history anchor is

$$Y_{\text{port}}(t) = \begin{bmatrix} f_{AB}(t) \\ v_A(t) - v_B(t) \end{bmatrix}. \tag{72}$$

Inputs and targets are standardized using training data only.

G.2 Reader-matched future-state evaluation

The original future-state analysis used a 100-update common reader. To match the interface experiment, we re-evaluated the same frozen encoders using the converged-reader protocol. No encoder, data split, normalization, checkpoint, or optimization seed was changed.

The constant-predictor standardized MSE is 1.0031. Mean converged-reader losses are 0.6582 for the joint representation and 0.6511 for the separate representation. Both therefore predict the future-state target substantially better than the constant predictor, while their partition contrast remains close to zero.

For comparison, the original 100-update reader gave $\overline{G} = -0.0063$ with standard deviation 0.0067. Reader convergence changes the mean contrast by only -0.0009. All ten converged readers stopped through the common convergence criterion after 207–264 updates; none reached the update limit. The one-sided test in the joint-favoring direction gives $p = 0.96$. We report this test for completeness rather than interpreting nonsignificance as evidence of equivalence.

G.3 Interface supervision

The interface experiment uses the same $k = 1$ trajectories, train/test splits, architecture family, total representation width $d = 2$, optimizer, training budget, and five optimization seeds as the future-state experiment. Only the prediction target changes, together with the temporal definition of that target: future-state prediction uses the change over the subsequent 1.0 time unit, whereas Y_{port} is evaluated at the history anchor.

Before representation training, we evaluated whether the partner history contains information about Y_{port} that is not available from either subsystem alone. Ridge predictors were fitted using H_A, H_B, the true joint history (H_A, H_B), and a control in which the partner history was replaced by a trajectory from another episode at the same anchor age. At $k = 1$, the best local predictor gives standardized test MSE 0.5895, whereas the true joint history gives 0.0289. The shuffled-partner predictor gives 0.5895. Thus the reduction in prediction error requires the trajectory-specific partner history rather than the additional input dimension alone.

For the converged common reader, the five interface contrasts are

$$G_{\text{port}} = \{0.3213,\ 0.4259,\ 0.3322,\ 0.4087,\ 0.3349\}, \tag{73}$$

with

$$\overline{G}_{\text{port}} = 0.3646. \tag{74}$$

All five seeds favor joint access. The corresponding trained-head contrasts are

$$\{0.3165,\ 0.4249,\ 0.3272,\ 0.4052,\ 0.3294\}, \tag{75}$$

with mean 0.3606, again with all five seeds positive.

The converged common reader reaches standardized MSE 0.156–0.186 on the joint representations and 0.503–0.582 on the separate representations. The one-sided paired test across the five seed contrasts gives $p = 3.8 \times 10^{-5}$ for the common reader and $p = 4.5 \times 10^{-5}$ for the trained heads. These tests quantify consistency across optimization seeds for the fixed dataset; they do not represent variation across independently generated datasets.

G.4 Interface accessibility after future-state supervision

To ask whether generic future-state supervision preserves any of the same interface structure, we freeze the future-state-trained encoders and fit the same converged reader to predict Y_{port}. No representation is retrained for this comparison.

The resulting mean contrast is

$$\overline{G}_{\mathrm{port}\leftarrow\mathrm{end}} = 0.0252, \tag{76}$$

with all five seed-level contrasts positive. Thus the joint future-state representation retains modestly more accessible interface information than the separate representation, despite their nearly identical performance on the future-state target itself. The difference is, however, much smaller than after direct interface supervision (0.0252 versus 0.3646).

This cross-decoding result is descriptive. It shows that the future-state representations are not identical with respect to the interface target, but does not identify which properties of the learned codes account for the difference.

G.5 Relational reference without physical coupling

At $k = 0$, Eq. (19) gives $f_{AB} = 0$, while the relative velocity $v_A - v_B$ remains a relational quantity. We therefore use the flow-only component as a reference for whether joint access can be useful without physical interaction.

On the trained representations, the mean joint-access advantage for this flow target is

$$\overline{G}_{\mathrm{flow}}(0) = 0.170. \tag{77}$$

This result is consistent with the Lorenz control in Appendix E: a relational target can favor joint access even when the two underlying systems are dynamically independent. We therefore do not interpret the $k = 0$ result as evidence about physical interaction itself.

G.6 Capacity relaxation

We next increase the total representation width while keeping the $k = 1$ data, interface target, architecture family, optimizer, $10{,}000$-update training budget, batching, final-checkpoint rule, and converged-reader protocol fixed. The $d = 4$ and $d = 8$ experiments use three optimization seeds distinct from the five $d = 2$ seeds; the width comparison is therefore descriptive rather than paired.

For the converged common reader, the mean interface contrasts are

d	2	4	8
$\overline{G}_{\mathrm{port}}$	0.3646	0.3481	0.0003

(78)

and the corresponding trained-head contrasts are

d	2	4	8
$\overline{G}^{\mathrm{head}}_{\mathrm{port}}$	0.3606	0.3514	0.0005.

(79)

The effect remains large at $d = 4$ but is essentially absent at $d = 8$. At $d = 8$, each branch of the separate architecture has four coordinates, matching the dimension of its local state, and both architectures predict the four-dimensional interface target nearly perfectly. This behavior is consistent with the structural account in Section 2.4: once the relevant local information can be retained separately, the advantage of forming the relation before compression disappears in this system. Because the wider experiments use three distinct seeds rather than the five $d = 2$ seeds, we do not interpret the width comparison as a paired monotonic effect or as an information-rate threshold.

G.7 Scope of the target intervention

The future-state and interface experiments match the physical system, observations, data, architecture family, representation width, optimization budget, and seeds at $d = 2$. They differ, however, in more than the semantic identity of the target. The future-state objective predicts a stochastic state change over 1.0 time unit, whereas the interface target is evaluated at the history anchor. The comparison therefore changes relational structure, temporal horizon, and stochasticity together.

The experiment establishes that the learned partition effect is target-relative on a fixed interacting system: successful future-state prediction yields essentially no joint-access advantage, while direct interface supervision yields a large one. It does not isolate which individual property of the two prediction problems accounts for the full difference.

H Additional Reproducibility Information

H.1 Standardization and evaluation conventions

Unless stated otherwise, all input and target normalization statistics are computed from the corresponding training split and then held fixed for selection and held-out evaluation. Reported MSE values for standardized targets are averaged over target coordinates and evaluation windows before aggregation over held-out source episodes.

The partition contrast throughout the paper is

$$G = L_S - L_J, \tag{80}$$

where L_J and L_S denote losses from the joint and separate representations under the same evaluation protocol. Positive G therefore indicates an advantage for joint access.

For comparisons across optimization seeds, architectures within a seed use the corresponding matched data and evaluation protocol. Seed-level statistical tests are used only where explicitly reported. Descriptive capacity studies that use different seed sets are not treated as paired comparisons.

H.2 Reader protocols

The controlled oscillator study uses the fixed 100-update common reader specified in Appendix C. The later Lorenz relational and double-well interface studies use the converged-reader protocol introduced after the strong-coupling Lorenz diagnostic.

For the double-well reader-matching experiment, convergence is evaluated using the same stopping rule for joint and separate representations. All ten future-state readers stop after 207–264 updates, and none reaches the maximum update budget. The frozen encoders and evaluation data remain unchanged.

The trained prediction heads provide a complementary readout of the same learned representations. They are reported as robustness measurements and are not independent representation-training replications.

H.3 Interpretation of representation width

All width comparisons in this paper concern the dimensionality of real-valued learned representations. As discussed in Appendix A, this is an architectural constraint rather than a finite communication rate. In particular, the transition observed in the double-well experiment at $d = 8$ should be interpreted relative to the local state and target dimensions of that experiment, not as a universal capacity threshold.

The experiments therefore test the structural prediction that separate compression becomes less restrictive once the target-relevant local information can be retained before fusion. Extending this comparison to explicit rate-constrained, noisy, or finite-resolution representations is left for future work.