TEMPORAL ACTION STRUCTURE CONTROLS REPRESENTATION IDENTIFIABILITY UNDER PREDICTIVE OBJECTIVES

Anonymous authors
Paper under double-blind review

ABSTRACT

World models learn from trajectories generated by a behavior policy, yet that policy is often summarized only by marginal action frequencies. We show that this summary can miss a variable that determines which equally informative state coordinates a predictive objective prefers: the temporal composition of actions. In a stationary Gaussian controlled world, the preference is governed by a spectral gap between the least predictable linear state direction and the most predictable quadratic alternative. For dimension-matched, population-whitened encoders, a positive gap uniquely selects the state up to rotation and stabilizes near-optimal solutions; a negative gap admits an arbitrarily close, smooth, invertible nonlinear improvement. We construct two full-support laws over length-five action sequences that match every interaction among up to three positions and the complete linear prediction operator, yet have opposite gap signs. Two conditionally independent futures from the same state and action sequence, available through a resettable simulator or an equivalent conditional-sampling interface, estimate this gap; a semidefinite program then designs a positive-gap distribution. In a controlled visual world, objective-only CNN training recovers the predicted coordinate switch across ten paired seeds. Starting from those selected CNNs, released LeWorldModel action-conditioning modules preserve the phase and benefit from the ordered word. Balanced open-loop ranking over the same sequences gives normalized regrets 0.073 and 0.366 for the positive- and negative-phase coordinates. The separation also persists in an explicit nonlinear Gaussian-stationary family. Thus marginal action coverage can fix what actions occur while temporal behavior controls which representation prediction rewards.

1 INTRODUCTION

Predictive representation learning seeks coordinates in which the future is easy to model. In an interactive system, those coordinates are learned from trajectories produced by a behavior policy. The policy determines not only how often each action occurs, but also how actions are composed over time. Two robot policies may issue the same number of left and right commands, for example, while one alternates them and the other executes long runs. Their action counts agree, yet they expose the learner to different multi-step transformations.

We study this temporal degree of freedom in isolation. We call a short ordered action sequence an *action word*, and the distribution from which such sequences are drawn an *action-word law*. Our question is: *with the world and learning objective fixed, can changing only the order statistics of behavior change the coordinates learned by prediction?* We ask this under strong controls: the two laws can have the same one-step frequencies, the same interactions among actions at any three positions, and even the same linear prediction operator.

This question is not settled by asking whether a representation preserves information. A rotation of the state and an invertible nonlinear change of coordinates can both retain the entire state, but they present different dynamics to a downstream model. A rotation preserves linear dynamics; a nonlinear change generally does not, even if it is smooth, well conditioned, and distribution preserving. A sufficiently expressive predictor may use either coordinate system, whereas a word-conditioned linear

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.

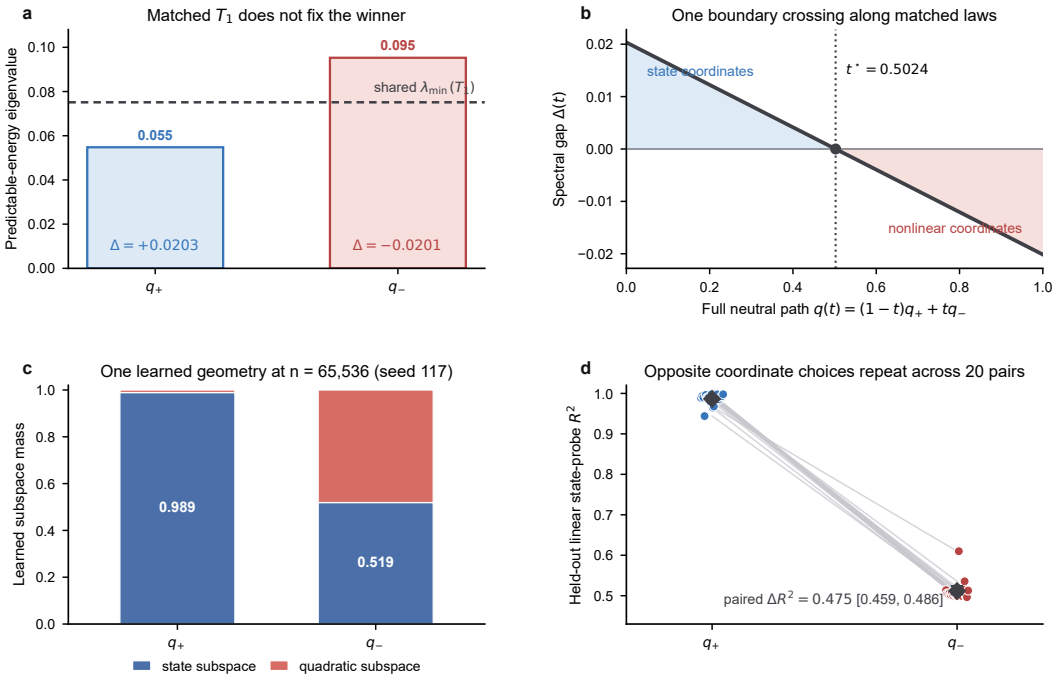

Figure 1: **Action order changes which coordinates prediction prefers. a**, The length-five behavior laws q_+ and q_- agree on every interaction among up to three positions and on all linearly predictable state signal. Nevertheless, the weakest linear direction lies above the strongest quadratic feature under q_+, and below it under q_-. **b**, Interpolating between the laws crosses this boundary once. **c**, A learner fitted only to two independent futures from the same state and action sequence recovers linear state features under q_+, but allocates about half of its variance to quadratic features under q_-. **d**, A held-out linear probe measures the resulting state recovery over 20 paired runs (mean difference 0.475, 95% interval [0.459, 0.486]).

latent predictor—one linear map for each action sequence—can be substantially less accurate in the nonlinear coordinates. Behavior can therefore change the usable structure of a representation without causing collapse or information loss.

Joint-embedding predictive architectures (JEPAs) learn by predicting target features rather than reconstructing observations (Assran et al., 2023; Bardes et al., 2024). LeJEPA and its world-model analysis show that predictive alignment combined with Gaussian output regularization can recover linear latent coordinates in an isotropic Gaussian world (Balestriero & LeCun, 2025; Klindt et al., 2026). These results clarify what the representation distribution should look like. They leave open a complementary question: which conditional transformations does the behavior policy place in the training data? Controlled-world analyses make behavior excitation explicit (Zhang et al., 2026a), but do not determine whether policies that appear identical to common coverage diagnostics can make linear state features lose to nonlinear predictive features.

We characterize this effect exactly in a stationary controlled Gaussian model. Figure 1 previews the result. Two action-word laws expose the same linear predictive signal and agree on action statistics involving any three positions. Under one law, the predictive objective uniquely selects state coordinates; under the other, it rewards a nonlinear coordinate. The matched-law difference persists after a smooth nonlinear perturbation that keeps the same Gaussian stationary state distribution. A separate analytic construction shows that even an encoder that preserves the full standard Gaussian distribution can prefer nonlinear coordinates in an anisotropic world. The missing condition is therefore not another marginal normalization constraint. It is a separation between predictable state directions and predictable nonlinear alternatives, and temporal behavior controls that separation.

1.1 CONTRIBUTIONS AND ROADMAP

The theory asks which features can carry the fixed variance of a standardized representation. Let $T_1(q)$ collect the predictable energy of linear state features under behavior law q, and let $T_2(q)$ collect the same quantity for centered quadratic features. Section 3 derives these operators from the transition model. Contractive dynamics ensure that still higher feature orders are no more predictable than the quadratic order. The decisive quantity is

$$\Delta(q) = \lambda_{\min}(T_1(q)) - \|T_2(q)\|_{\mathrm{op}}.$$

The first term is the least predictable state direction; the second is the most predictable quadratic alternative. Their difference, $\Delta(q)$, is the *linear–nonlinear predictability gap*. Our main theorem proves a sharp dichotomy. If the gap is positive, every optimal standardized encoder is a rotation of the state, and near-optimal encoders remain quantitatively close to one. If the gap is negative, a smooth invertible nonlinear coordinate can improve the objective arbitrarily close to the state representation.

A sharp representation criterion (Section 3). We prove that the sign of $\Delta(q)$ is both sufficient for linear state recovery and, in the opposite direction, certifies a smooth invertible nonlinear improvement. The positive side includes a stability bound for near-optimal encoders. The negative side shows that the problem is not information loss: the preferred nonlinear coordinate still determines the full state. We also prove that the matched-law boundary persists under sufficiently small, exactly Gaussian-stationary nonlinear perturbations. A separate anisotropic construction shows that exact Gaussian preservation alone does not rule out a nonlinear improvement.

A limitation of ordinary action coverage (Section 3.4). We construct two length-five behavior laws that agree on all action interactions up to degree three and on the complete linear prediction operator, yet have opposite gap signs. The construction isolates what the usual summaries omit: higher-order temporal composition. We then quantify a downstream consequence in Section 3.5: the nonlinear coordinate favored by the rich objective is harder to use with a word-conditioned linear latent predictor.

From a criterion to behavior design, learning, and decisions (Section 4). The evidence follows two complementary tracks. First, replicated pilot data recover the spectral ordering and use it prospectively to design a behavior law. Separately, the fixed matched laws q_+ and q_- test whether a CNN discovers the predicted coordinates from rendered images, whether released LeWorldModel action-conditioned predictor modules retain and use them, and whether they alter a balanced one-word open-loop action choice. Keeping these tracks separate distinguishes behavior design from neural coordinate discovery and separates what a probe decodes from what a learned predictor and decision rule actually use.

2 RELATED WORK

Predictive representations and world-model identifiability. JEPAs predict in representation space rather than reconstructing observations (Assran et al., 2023; Bardes et al., 2024). LeJEPA combines predictive alignment with Gaussian regularization (Balestriero & LeCun, 2025), and LeWorldModel instantiates this recipe in an action-conditioned pixel predictor (Maes et al., 2026). Recent theory proves linear identifiability and planning consequences for LeJEPA in isotropic Gaussian worlds and already notes that anisotropic transitions can interleave linear and higher-order Hermite features (Klindt et al., 2026). Other work studies finite-sample planning error (Cui et al., 2026) and separates statistical from structural identifiability more broadly (Nelson et al., 2026). We ask a different question: can temporal behavior flip this ordering when ordinary action summaries and the complete linear prediction operator are held fixed? Our matched action-word laws show that it can. A separate Gaussian-preserving example then isolates why marginal Gaussianization cannot by itself certify the missing cross-degree separation.

Identifiability under actions and behavior data. Controlled-world theory studies joint state and transition recovery under state-dependent behavior and conditional action excitation (Zhang et al., 2026a), while active experiment design chooses state-dependent policies to allocate informative

measurements (Mutny et al., 2023). Other methods alter the learned factors, prediction targets, or latent geometry to expose action sensitivity, physical variables, or goal progress (Liu et al., 2023; Zhang et al., 2026b; Tan et al., 2026; Destrade et al., 2025; Bai & Xiong, 2026). We instead keep the world, prediction target, and objective fixed, and change only a state-independent distribution over finite action sequences. Crucially, our paired laws agree on all action interactions up to three positions and on the complete linear prediction operator. The distinction is therefore not missing marginal action coverage or weak one-step excitation; it is higher-order temporal composition.

Spectral prediction and sequence design. Maximal correlation and nonlinear CCA connect prediction to spectral dependence (Hirschfeld, 1935; Gebelein, 1941; Rényi, 1959; Han et al., 2025). Slow feature analysis, VAMP, and Koopman methods use related operator views to learn predictable or invariant features (Wiskott & Sejnowski, 2002; Wu & Noé, 2020; Mardt et al., 2018; Koopman, 1931; Williams et al., 2015; Takeishi et al., 2017; Lusch et al., 2018; Haseli & Cortés, 2026). Hermite forecasting supplies the degree-wise language used in our Gaussian analysis (McElroy & Das, 2021). Classical input design asks which control sequences identify dynamics (Hjalmarsson, 2005), while predictive-state models organize controlled futures (Littman & Sutton, 2001); orthogonal arrays and trades provide distributions with matched low-order statistics (Rao, 1947; Bush, 1952; Hedayat et al., 1999; Demirkale et al., 2016). We combine these ideas for a different target—selecting representation order under fixed linear dynamics—through a sequence-level criterion that is identifiable, estimable, and optimizable.

3 How temporal behavior selects a representation

We isolate the effect of the action-word law while keeping the marginal state distribution fixed. This makes it possible to ask a causal design question: if the dynamics and observation process are unchanged, can changing only the distribution of action sequences change which representation the learning objective prefers?

3.1 A controlled world that isolates temporal behavior

Let $X \in \mathbb{R}^d$ denote the current latent state, and choose coordinates so that $X \sim \mathcal{N}(0, I_d)$. Let $\mathcal{A}$ be a finite action set and let q be a distribution on $\mathcal{A}^L$. The behavior law samples a word $W = (W_1, \dots, W_L) \sim q$, independently of X. Each single-step, or *primitive*, action $a \in \mathcal{A}$ has a matrix $A_a \in \mathbb{R}^{d\times d}$. Executing a particular word $w = (w_1, \dots, w_L)$ therefore applies the ordered product $M_w = A_{w_L} \cdots A_{w_1}$.

Conditioned on the current state and word, the future state is

$$Y = M_w X + \xi_w, \qquad \xi_w \sim \mathcal{N}(0, I_d - M_w M_w^\top), \qquad \xi_w \perp X,$$

where $0 \preceq M_w M_w^\top \preceq I_d$. The noise covariance is chosen so that every word preserves the same stationary marginal, $Y \sim \mathcal{N}(0, I_d)$. This removes a simple confound: changing q changes how states are connected over time, but not which states are observed.

Our objective uses two independent futures Y_1 and Y_2 generated from the same (X, W). In a simulator or resettable environment, these are two rollouts from the same state under the same action word. The learned encoder $h : \mathbb{R}^d \to \mathbb{R}^d$ is centered and whitened at the population level,

$$\mathbb{E}h(X) = 0, \qquad \mathbb{E}h(X)h(X)^\top = I_d.$$

Whitening rules out constant and rescaled solutions, so different encoders have the same total variance and can be compared through predictability alone. The theorem is dimension matched ($d_z = d$): "full measurable" means every centered, whitened $h : \mathbb{R}^d \to \mathbb{R}^d$. If $d_z > d$, linear and nonlinear modes may coexist; the spectrum ranks leading subspaces, but the zero-sum uniqueness conclusion does not transfer unchanged. Appendix H shows that even the stronger constraint $h_\# \gamma_d = \gamma_d$ is not sufficient for an anisotropic temporal operator. The analysis is written in state coordinates, but the population L^2 conclusion transfers to nonlinear observations $O = g(X)$ whenever the renderer is invertible: a state-coordinate encoder h is implemented as $\widetilde{h} = h \circ g^{-1}$. Transferring the smooth bi-Lipschitz converse also requires the renderer and its inverse to preserve that regularity.

3.2 PREDICTABLE ENERGY SEPARATES FEATURES BY DEGREE

We first measure how much of the future representation is predictable from the full current state. For two replicated futures, define

$$F_q(h) = \mathbb{E}\big[h(Y_1)^\top h(Y_2)\big] = \mathbb{E}\big\|\mathbb{E}[h(Y) \mid X, W]\big\|^2$$

where the expectation includes X, the word $W \sim q$, and transition noise. Conditional independence of Y_1 and Y_2 gives the second equality. Thus $F_q(h)$ is precisely the squared energy of the predictable part of $h(Y)$; larger values mean that more representation variance can be forecast.

The score above gives the predictor access to the full state X. To separate predictability from information retained by the encoder, we also consider the best predictor that sees only $h(X)$ and the word:

$$B_q(h) = \mathbb{E}\big\|\mathbb{E}[h(Y) \mid h(X), W]\big\|^2.$$

The tower property gives $B_q(h) \le F_q(h)$: an encoder may discard information that was useful for predicting its own future. The two scores coincide whenever h is invertible, because then $h(X)$ contains exactly the same information as X. We optimize and analyze the replica score F_q, while B_q lets us verify that the negative result is not created by giving the predictor hidden access to the state.

The Gaussian assumption gives a concrete language for comparing linear and nonlinear coordinates. For example, when $d = 2$, the first feature degree is spanned by X_1, X_2, while the second is spanned by the centered quadratic features $(X_1^2 - 1)/\sqrt{2}$, X_1X_2, and $(X_2^2 - 1)/\sqrt{2}$. More generally, every centered square-integrable feature has the orthogonal decomposition

$$L_0^2(\gamma_d) = \bigoplus_{k \ge 1} \mathcal{H}_k, \qquad \mathcal{H}_k \simeq \mathrm{Sym}^k(\mathbb{R}^d),$$

where γ_d is the standard Gaussian measure and $\mathcal{H}_k$ is the degree-k Hermite subspace. Thus $\mathcal{H}_1$ contains the state coordinates, $\mathcal{H}_2$ contains quadratic alternatives, and larger k represents higher nonlinearity. No truncation is made: the decomposition covers the entire centered L^2 encoder class.

The action-word law acts separately on each degree. Its degree-k predictability operator is

$$T_k(q) = \sum_w q_w (M_w M_w^\top)^{\odot k}.$$

Here q_w is the probability of sequence w, and $\odot k$ denotes the k-fold symmetric tensor power in an orthonormal Hermite basis. The operator $T_k(q)$ therefore records how much degree-k future variation remains predictable after averaging over action sequences. In particular, T_1 describes linear state features and T_2 describes quadratic features.

Formally, if $P_w f = \mathbb{E}[f(Y) \mid X, W = w]$, the full predictable-energy operator is $L_q = \sum_w q_w P_w^* P_w$. It preserves feature degree and has blocks $T_1, T_2, \ldots$. This is the key simplification: an optimization over arbitrary measurable encoders becomes a comparison between feature degrees. Appendix B derives the decomposition, including the $M_w M_w^\top$ orientation, using standard Gaussian second quantization (Janson, 1997; van Neerven, 2005; Peccati & Taqqu, 2011).

Together, these operators support a three-step argument. The next theorem turns their comparison into an exact identifiability boundary over the full dimension-matched measurable encoder class. We then show that standard policy summaries do not determine which side of that boundary the data occupy, before measuring what the resulting coordinate choice costs a word-conditioned linear latent predictor.

3.3 A SHARP BOUNDARY BETWEEN LINEAR RECOVERY AND NONLINEAR SHORTCUTS

Whitening fixes the total variance of every admissible representation. The encoder can only decide how much of that variance lies in linear features and how much lies in nonlinear ones. Because the channels are contractive, predictability decreases with Hermite degree. The strongest nonlinear competitor is therefore quadratic, and the entire problem reduces to one comparison: is the weakest linear state direction still stronger than the best quadratic feature?

We measure this separation by $\Delta(q) = \lambda_{\min}(T_1(q)) - \|T_2(q)\|_{\mathrm{op}}$. We call $\Delta > 0$ the *state-recovery side* of the boundary and $\Delta < 0$ the *nonlinear-shortcut side*. Let Π_1 be the L^2 projection onto $\mathcal{H}_1$,

and define $\eta(h) = \sum_{i=1}^{d} \|(I - \Pi_1)h_i\|_{L^2}^2 \in [0, d]$. Thus $\eta(h) = 0$ exactly when a whitened encoder is an orthogonal linear copy of the state. The following theorem is the paper's central result. It shows that Δ is not merely a diagnostic: its sign determines whether state coordinates are identifiable.

Theorem 1 (Exact criterion for state-coordinate recovery)**.** *In the setting of Section 3, every centered, population-whitened measurable encoder $h : \mathbb{R}^d \to \mathbb{R}^d$ satisfies*

$$B_q(h) \le F_q(h) \le \operatorname{tr} T_1(q) - \Delta(q)\eta(h), \quad \Delta(q) = \lambda_{\min}(T_1(q)) - \|T_2(q)\|_{\mathrm{op}}.$$

State-recovery side. *If $\Delta(q) > 0$, the maximizers of both scores are exactly $h(x) = Qx$ γ_d-almost surely for $Q \in O(d)$, where $O(d)$ is the orthogonal group. For either $J \in \{F_q, B_q\}$, every $\varepsilon \ge 0$ and every encoder satisfying $\operatorname{tr} T_1(q) - J(h) \le \varepsilon$ obeys*

$$\inf_{Q \in O(d)} \mathbb{E}\|h(X) - QX\|^2 \le \frac{2\varepsilon}{\Delta(q)}.$$

Nonlinear-shortcut side. *If $\Delta(q) < 0$, no orthogonal linear encoder is a local maximizer of either score in Gaussian L^2 within the admissible exactly whitened class: every QX admits arbitrarily close centered, covariance-whitened, C^∞ global bi-Lipschitz, invertible perturbations h_t such that*

$$B_q(h_t) = F_q(h_t) > B_q(QX) = F_q(QX).$$

The two sides have different meanings. When $\Delta > 0$, moving variance away from the state costs at least Δ per unit, which yields both uniqueness and stability. When $\Delta < 0$, the better coordinate still preserves the full state; what fails is the objective's preference for linear state coordinates. At $\Delta = 0$, the feature degrees may tie and no uniqueness is claimed.

The proof has one mechanism. Contraction makes the quadratic block an upper bound for every nonlinear feature degree. If this ceiling lies below the weakest linear direction, state coordinates win globally; if the ordering reverses, moving an arbitrarily small amount of variance to the best quadratic feature improves the objective. Appendices B–G give the complete proof, approximate-whitening and operator-perturbation bounds, and a nonlinear Gaussian-stationary extension. In a separate anisotropic channel example, Appendix H constructs a smooth invertible flow that preserves the entire Gaussian distribution while making the linear identity locally suboptimal. Marginal Gaussianity is therefore not a universal certificate of linear optimality.

The theorem establishes the first link in the argument: the action-word law determines T_1 and T_2, and their margin determines the preferred coordinates. The next question is whether ordinary coverage summaries already determine that margin.

3.4 Action frequencies and three-position statistics miss the boundary

The criterion above would add little if ordinary action coverage already determined it. We therefore hold familiar statistics fixed and vary only the joint distribution over complete action sequences. The two laws match pure positional behavior moments through degree three and, separately, the complete environment-dependent linear predictive operator T_1. This audited collection is substantially stronger than action counts or one-step marginals, yet it still leaves room for T_2 to change. For binary actions, encode the action at position i as $s_i(W) \in \{-1, +1\}$. Matching through degree three means matching $\mathbb{E}_q \prod_{i \in S} s_i(W)$ for every $S \subseteq \{1, \dots, L\}$ with $|S| \le 3$.

Appendix I gives the general rank characterization of this matched-summary family and its connection to orthogonal-array trades (Rao, 1947; Bush, 1952; Hedayat et al., 1999; Demirkale et al., 2016).

Having room to vary T_2 is not enough: the variation must cross the boundary in Theorem 1. The next result gives an exact pair for which every monitored summary and the full linear predictive signal agree, but the preferred representation changes.

Proposition 2 (Identical audited coverage, opposite representation preferences)**.** *There are rational 2×2 matrices for two primitive actions and laws q_+, q_- on length-five binary words such that every word has probability at least $1/2000$, the laws agree on every positional moment through order three and on the complete T_1, but*

$$\Delta(q_+) > \frac{1}{50}, \qquad \Delta(q_-) < -\frac{1}{50}.$$

For $t \in [0, 1]$, along $q(t) = q_+ + t(q_- - q_+)$, all constrained action statistics are constant and $\Delta(q(t))$ crosses zero exactly once, at $t_0 \in (0.5024, 0.5025)$, with $\Delta'(t_0) \ne 0$.

This pair supplies the controlled intervention needed by the argument. The world, stationary distribution, degree-three positional behavior moments, and complete linear predictive signal remain fixed; only higher-order temporal dependence changes. By Theorem 1, that hidden change flips whether the objective recovers state coordinates. No diagnostic based only on the matched action statistics or on T_1 can therefore determine the representation regime.

Appendix I gives the exact matrices, probabilities, rank calculation, and rational certificate, and proves that the crossing persists in a neighborhood of the primitive actions. Near the boundary, distinguishing gaps $+\gamma$ and $-\gamma$ requires $\Theta(\gamma^{-2})$ samples, a rate attained by replicated-future cross moments in fixed dimension (Appendices J–L) (Tsybakov, 2009; Catoni, 2012; Minsker, 2018).

3.5 Why the selected coordinates matter to a linear latent predictor

The preceding results identify which coordinates maximize predictable energy. We now ask whether two invertible coordinate systems are equally easy for a restricted world model to use. Our restricted model has one linear latent map for each action sequence; it is the simplest action-conditioned predictor that can represent linear controlled dynamics exactly. For a centered, whitened encoder, let $S_q^{\mathrm{lin}}(h)$ be the future variance explained by this best word-conditioned linear predictor:

$$S_q^{\mathrm{lin}}(h) = \sum_w q_w \left\| \mathbb{E}[h(Y)h(X)^\top \mid W = w] \right\|_F^2 .$$

The matrix inside the norm is the present–future cross-covariance under sequence w, so the best linear prediction error is $d - S_q^{\mathrm{lin}}(h)$. Appendix C proves an exact projection identity: the difference $F_q(h) - S_q^{\mathrm{lin}}(h)$ is precisely the predictable future signal that leaves the linear span of the learned coordinates. This exposes the tension: a nonlinear coordinate can increase total predictable energy while making less of that energy accessible to a linear latent model.

Proposition 3 (Predictor-capacity reversal)**.** *For the explicit $d = 2$ law q_-, there is a one-parameter family of smooth, invertible, exactly whitened coordinates h_t, starting from the linear state coordinates at $t = 0$, with the following property. The rich score F_q prefers h_t for every $0 < t < 1$, whereas the best word-conditioned linear predictor prefers the original state coordinates for $0 < t < 0.9523$. At the point of largest linear-score decrease, the rich score increases by 0.00882, while the linearly explained score decreases by 0.03357, or 14.9% of the baseline state predictable energy $\mathrm{tr}(T_1) = 0.2254825\ldots$. Both h_t and its inverse have global Lipschitz constants below three at this point.*

This is the operational consequence of the spectral boundary. Every h_t retains exactly the same state information, yet it presents different dynamics to a restricted predictor: the rich objective values a predictable nonlinear mode whose future is not fully contained in the coordinate span available to a linear latent head. The coordinate choice in Theorem 1 can therefore change the effective complexity of a world model without discarding state information. Appendix C evaluates the projection identity along the explicit path and verifies the numerical interval and the conditioning of the map and its inverse.

4 From the spectral criterion to behavior design and learned decisions

Theorem 1 yields two experimentally distinct consequences. The margin should be estimable well enough to design behavior, and its sign should predict what fixed matched laws do to learned coordinates and their downstream use. We test the first claim with replicated pilot data and the second with the predefined pair q_+, q_-. This separation avoids treating the designed law as the input to the neural experiments. Within the neural track, readouts become progressively more operational: probes identify coordinates, held-out risk asks whether a learned head uses them, and regret asks whether the difference changes an action choice.

4.1 Recovering the spectral phase from finite data

We begin in feature space, where estimation can be separated from neural optimization. The learner receives an orthogonal mixture that conceals which input features are linear and which are quadratic.

From two independent futures generated from the same current state and action word, it estimates the replicated-future cross moment and selects its leading two-dimensional subspace. A held-out *state-probe* R^2 then asks whether the two state coordinates survived: a value near one means that both are linearly recoverable, whereas a value near one half means that one has been replaced. The complementary *quadratic-feature fraction* asks whether the replacement is the specific second-order feature predicted by the theorem. These diagnostic labels are revealed only after subspace selection.

At $n = 65{,}536$, state-probe R^2 is 0.9865 under q_+ and 0.5117 under q_-, and all 20 paired runs follow the sign of the population gap (Fig. 1c–d). The quadratic-feature fraction identifies the missing direction as the specific second-order feature predicted by the theorem. Hence finite data preserve both parts of the claim: which side of the boundary is selected and what replaces the state direction.

4.2 Using the margin to design behavior

The same estimator can be used prospectively, provided a simulator can reset to a sampled current state and generate conditionally independent futures (or the required coordinates have been independently certified). It is not an online estimator from one unreset trajectory, and the finite-sample design guarantee is not a plug-in guarantee for an uncertified pixel encoder. We hold every positional action moment through degree three equal to the negative-phase reference and maximize the estimated gap over behavior laws. With 4,096 replica pairs per word, the selected law has true gap 0.02468 (95% CI $[0.02432, 0.02504]$) across 20 independent runs, improving the reference by 0.04482. Its mean regret to the population-oracle design is 4.12×10^{-4}. Optimizing only the linear operator T_1 yields a smaller gap of 0.01656; the paired advantage of optimizing the full linear–nonlinear gap is 0.00812 (95% CI $[0.00061, 0.01563]$). Thus the gap is not only a diagnostic: it selects a behavior law that preserves the required low-order action statistics while improving the representation phase. The comparison with T_1-only design shows that maximizing linear predictability alone is suboptimal for the full margin: it gives a narrower positive gap and weaker finite-sample recovery than joint T_1–T_2 optimization. The semidefinite program and its statistical guarantees are given in Appendix K.

These two tests close the loop in feature space: the phase margin can be recovered from finite samples and used prospectively to improve the behavior law (Appendix Fig. 6). They do not assume that a neural encoder will find the same coordinates. The next experiments return to the fixed matched pair q_+, q_-, remove the supplied feature family, and follow its coordinates through image learning, prediction, and choice.

4.3 From pixels to learned prediction and choice

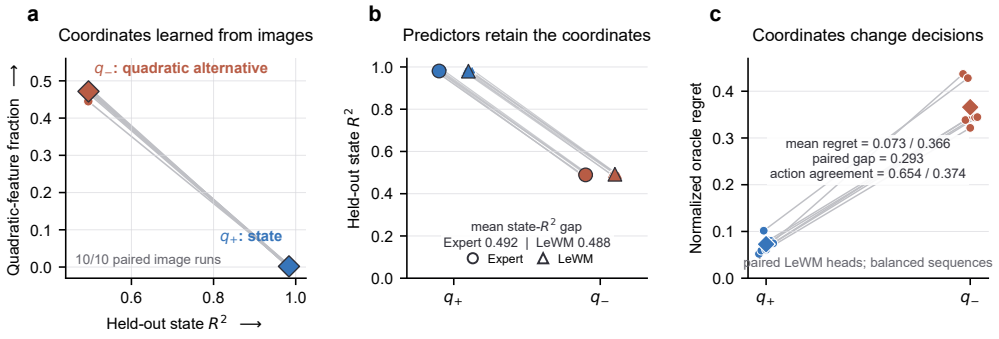

Figure 2: **Replica-selected image coordinates persist through predictor training and alter one-word open-loop decisions.** **a**, All ten label-free CNN pairs switch from state under q_+ to the predicted quadratic alternative under q_-. **b**, After joint training with released LeWorldModel head modules, the held-out state-R^2 gap remains in the same controlled visual world. **c**, Fresh instances of the same head see all 32 words equally, yet q_+ coordinates yield lower action-selection regret. Panels b–c show all seven held-out pairs.

Discovering the coordinates from pixels. The preceding learner was given the relevant feature family. The next test removes that advantage and trains an encoder directly from observations. A deterministic nonlinear, invertible renderer maps each state to a 64×64 RGB image, and a 689,586-parameter CNN maximizes an empirical replica objective estimating F_q, with hard whitening enforcing zero mean and identity covariance. The encoder is learned directly from pixels; omitting a predictor head at this stage isolates coordinate discovery. Checkpoints are selected using held-out F_q alone; state and Hermite labels are used only after selection to identify the learned coordinates.

All ten paired runs recover the difference predicted by Theorem 1. The mean state-R^2 gap between q_+ and q_- is 0.4882 (95% CI $[0.4863, 0.4900]$), and the quadratic-feature fraction changes by 0.4699. Under q_-, the selected representation also exceeds the linear state coordinates in held-out F_q by 0.0160 (Fig. 2a). Jointly, the probe, feature fraction, and held-out objective distinguish a failed state-recovery run from the predicted nonlinear optimum: the CNN discovers the coordinate that the population objective ranks above the state without being supplied Hermite features.

Using the coordinates for learned prediction. Coordinate discovery alone does not show that an action-conditioned world model can use the result. We therefore couple the selected encoders to released LeWorldModel action-embedding and AdaLN predictor modules within the same controlled visual world (Maes et al., 2026). Each run starts from its objective-selected CNN checkpoint from the preceding experiment; the predictor was not used to choose that checkpoint. Given $h(X)$ and the ordered action word W, the head predicts $h(Y)$. During joint training, the predictor minimizes single-future latent risk, while the encoder retains its replica objective plus a weight-0.1 predictor term. The protocol used three disclosed development pairs and then seven frozen held-out pairs, not a post-hoc subset. On the latter, the state-R^2 gap remains 0.4880 (Fig. 2b); preserving word order reduces risk by 0.0977 under q_+ and 0.0930 under q_-. Because this expressive head uses both coordinate systems, the later regret gap measures surrogate geometry, not predictor incapacity. The linear-head reversal is theoretical (Proposition 3, Appendix C); Appendix M instead compares LeWorldModel with Expert, an unshared 38,976-parameter per-word MLP bank.

A decision consequence. Probes and prediction risk are still intermediate readouts: neither shows whether a coordinate choice changes action. We fit fresh heads on balanced data over all 32 words and evaluate 4,096 tasks with $x, x_{\mathrm{goal}} \overset{\mathrm{iid}}{\sim} \mathcal{N}(0, I_2)$. Each head ranks words by predicted latent squared distance plus a fitted residual trace. We normalize the selected word's excess state cost by $|\mathcal{W}|^{-1} \sum_v J_x(v) - \min_v J_x(v)$, the uniform-word excess baseline. The q_+ representation selects a lowest-cost word in 65.4% of tasks, versus 37.4% for q_-. Mean normalized regret rises from 0.073 to 0.366, a paired gap of 0.293 (95% CI $[0.250, 0.336]$), with the same direction in all seven pairs (Fig. 2c). Because both heads receive equal exposure to every word, missing actions cannot explain the gap. Appendix D gives the distortion lemma; Appendix M, Table 6, gives the neural protocol and metric.

5 Discussion and conclusion

Within the controlled class, one spectral margin determines whether prediction identifies linear state coordinates or favors an invertible nonlinear alternative. Neither degree-three positional statistics nor the complete linear prediction operator determines it; replicated pilots can estimate and optimize it. Fixed opposite-phase laws yield the predicted coordinate, risk, and one-word open-loop differences.

The scope is precise. Pixel encoders are selected by a replicated-future objective; LeWorldModel training begins from those checkpoints and retains that objective. These experiments therefore do not establish that an ordinary single-future end-to-end world model will discover both phases from random initialization. We rank one length-five word open-loop, not closed-loop MPC, and extend only to an explicit Gaussian-stationary radial-twist family rather than generic non-Gaussian dynamics.

Thus temporal behavior can control representation geometry even when familiar marginal summaries match, making temporal-structure design part of representation learning.

AI USE STATEMENT

Generative AI tools assisted with literature search and organization; brainstorming and refinement of hypotheses, theoretical models, mathematical claims, and proof strategies; experiment design and interpretation; code and Lean-formalization development and debugging; figure and artifact preparation; Chinese–English translation; and manuscript structure and readability edits. All AI-assisted claims, proofs, code, citations, and results were checked against mathematical derivations, primary sources, executable tests and certificates, and original experiment outputs, as applicable. The authors reviewed all AI-assisted work and take full responsibility for the final content, including text, claims, code, and artifacts.

REPRODUCIBILITY STATEMENT

Appendix M reports the experiments' seeds, sample sizes, compute, and solver diagnostics; Appendix N states the boundary of the Lean formalization. The anonymous supplement contains fixed rational witnesses, exact verification scripts, raw JSON results, environment specifications, and replay commands. The exact certificates require only CPU computation. The ten-seed pixel confirmation ran on three RTX 2080 Ti GPUs; the LeWorldModel predictor and decision confirmations, as well as the earlier development experiments, ran on a single RTX 4060 Ti. The exact nonlinear stationary confirmation is CPU-only.

REFERENCES

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, June 2023. doi: 10.1109/CVPR52729.2023.01499. URL https://doi.org/10.1109/CVPR52729.2023.01499.

Jiaxin Bai and Jiaxuan Xiong. Temporal-distance JEPA: Plan-aware representation learning for latent world model predictive control. *arXiv preprint arXiv:2607.25337*, 2026. URL https://arxiv.org/abs/2607.25337.

Randall Balestriero and Yann LeCun. LeJEPA: Provable and scalable self-supervised learning without the heuristics. *arXiv preprint arXiv:2511.08544*, 2025. URL https://arxiv.org/abs/2511.08544.

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. *arXiv preprint arXiv:2404.08471*, 2024. URL https://arxiv.org/abs/2404.08471.

K. A. Bush. Orthogonal arrays of index unity. *The Annals of Mathematical Statistics*, 23(3):426–434, 1952. doi: 10.1214/aoms/1177729387. URL https://doi.org/10.1214/aoms/1177729387.

Olivier Catoni. Challenging the empirical mean and empirical variance: A deviation study. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 48(4):1148–1185, 2012. doi: 10.1214/11-AIHP454. URL https://doi.org/10.1214/11-AIHP454.

Jingyi Cui, Qi Zhang, Hongwei Wen, and Yisen Wang. A generalization theory for JEPA-based world models. *arXiv preprint arXiv:2606.27014*, 2026. URL https://arxiv.org/abs/2606.27014.

Fatih Demirkale, Diane M. Donovan, Selda Küçükçifçi, and Emine Şule Yazıcı. Orthogonal trades and the intersection problem for orthogonal arrays. *Graphs and Combinatorics*, 32(3):903–912, 2016. doi: 10.1007/s00373-015-1638-y. URL https://doi.org/10.1007/s00373-015-1638-y.

Matthieu Destrade, Oumayma Bounou, Quentin Le Lidec, Jean Ponce, and Yann LeCun. Value-guided action planning with JEPA world models. *arXiv preprint arXiv:2601.00844*, 2025. URL https://arxiv.org/abs/2601.00844.

Hans Gebelein. Das statistische problem der korrelation als variations- und eigenwertproblem und sein zusammenhang mit der ausgleichsrechnung. *Zeitschrift für Angewandte Mathematik und Mechanik*, 21(6):364–379, 1941. doi: 10.1002/zamm.19410210604. URL https://doi.org/10.1002/zamm.19410210604.

Zhiwei Han, Stefan Matthes, and Hao Shen. Provable affine identifiability of nonlinear CCA under latent distributional priors. *arXiv preprint arXiv:2510.04758*, 2025. URL https://arxiv.org/abs/2510.04758.

Masih Haseli and Jorge Cortés. Modeling nonlinear control systems via koopman control family: Universal forms and subspace invariance proximity. *Automatica*, 185:112722, 2026. doi: 10.1016/j.automatica.2025.112722. URL https://arxiv.org/abs/2307.15368.

A. S. Hedayat, N. J. A. Sloane, and John Stufken. *Orthogonal Arrays: Theory and Applications*. Springer Series in Statistics. Springer New York, 1999. ISBN 9780387987668. doi: 10.1007/978-1-4612-1478-6. URL https://doi.org/10.1007/978-1-4612-1478-6.

H. O. Hirschfeld. A connection between correlation and contingency. *Mathematical Proceedings of the Cambridge Philosophical Society*, 31(4):520–524, 1935. doi: 10.1017/S0305004100013517. URL https://doi.org/10.1017/S0305004100013517.

Håkan Hjalmarsson. From experiment design to closed-loop control. *Automatica*, 41(3):393–438, 2005. doi: 10.1016/j.automatica.2004.11.021. URL https://doi.org/10.1016/j.automatica.2004.11.021.

Svante Janson. *Gaussian Hilbert Spaces*, volume 129 of *Cambridge Tracts in Mathematics*. Cambridge University Press, 1997. ISBN 9780521561280. doi: 10.1017/CBO9780511526169. URL https://doi.org/10.1017/CBO9780511526169.

David Klindt, Yann LeCun, and Randall Balestriero. When does LeJEPA learn a world model? *arXiv preprint arXiv:2605.26379*, 2026. URL https://arxiv.org/abs/2605.26379.

B. O. Koopman. Hamiltonian systems and transformation in hilbert space. *Proceedings of the National Academy of Sciences*, 17(5):315–318, 1931. doi: 10.1073/pnas.17.5.315. URL https://doi.org/10.1073/pnas.17.5.315.

Michael L. Littman and Richard S. Sutton. Predictive representations of state. In *Advances in Neural Information Processing Systems*, volume 14, 2001. URL https://proceedings.neurips.cc/paper/2001/hash/1e4d36177d71bbb3558e43af9577d70e-Abstract.html.

Yu-Ren Liu, Biwei Huang, Zhengmao Zhu, Honglong Tian, Mingming Gong, Yang Yu, and Kun Zhang. Learning world models with identifiable factorization. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://openreview.net/forum?id=6JJq5TW9Mc.

Bethany Lusch, J. Nathan Kutz, and Steven L. Brunton. Deep learning for universal linear embeddings of nonlinear dynamics. *Nature Communications*, 9:4950, 2018. doi: 10.1038/s41467-018-07210-0. URL https://www.nature.com/articles/s41467-018-07210-0.

Lucas Maes, Quentin Le Lidec, Damien Scieur, Yann LeCun, and Randall Balestriero. LeWorldModel: Stable end-to-end joint-embedding predictive architecture from pixels. *arXiv preprint arXiv:2603.19312*, 2026. URL https://arxiv.org/abs/2603.19312.

Andreas Mardt, Luca Pasquali, Hao Wu, and Frank Noé. VAMPnets for deep learning of molecular kinetics. *Nature Communications*, 9:5, 2018. doi: 10.1038/s41467-017-02388-1. URL https://doi.org/10.1038/s41467-017-02388-1.

Tucker McElroy and Srinjoy Das. Nonlinear prediction via Hermite transformation. *Statistical Theory and Related Fields*, 5(1):49–54, 2021. doi: 10.1080/24754269.2020.1856589. URL https://doi.org/10.1080/24754269.2020.1856589.

Stanislav Minsker. Sub-gaussian estimators of the mean of a random matrix with heavy-tailed entries. *The Annals of Statistics*, 46(6A):2871–2903, 2018. doi: 10.1214/17-AOS1642. URL https://doi.org/10.1214/17-AOS1642.

Mojmir Mutny, Tadeusz Janik, and Andreas Krause. Active exploration via experiment design in markov chains. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 7349–7374, 2023. URL https://proceedings.mlr.press/v206/mutny23a.html.

Walter Nelson, Marco Fumero, Theofanis Karaletsos, and Francesco Locatello. Statistical and structural identifiability in representation learning. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=Wa3cfE3Iay.

Giovanni Peccati and Murad S. Taqqu. *Wiener Chaos: Moments, Cumulants and Diagrams: A Survey with Computer Implementation*. Bocconi & Springer Series. Springer Milan, 2011. ISBN 9788847016798. doi: 10.1007/978-88-470-1679-8. URL https://doi.org/10.1007/978-88-470-1679-8.

C. Radhakrishna Rao. Factorial experiments derivable from combinatorial arrangements of arrays. *Journal of the Royal Statistical Society: Series B*, 9(1):128–139, 1947. doi: 10.2307/2983576. URL https://doi.org/10.2307/2983576.

Alfréd Rényi. On measures of dependence. *Acta Mathematica Academiae Scientiarum Hungaricae*, 10(3–4):441–451, 1959. doi: 10.1007/BF02024507. URL https://doi.org/10.1007/BF02024507.

Naoya Takeishi, Yoshinobu Kawahara, and Takehisa Yairi. Learning Koopman invariant subspaces for dynamic mode decomposition. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3a835d3215755c435ef4fe9965a3f2a0-Abstract.html.

Kaizhen Tan, Xin Xu, Siru Tao, Yixiao Li, Hanzhe Hong, Yang Feng, and Heqing Du. What can latent world models know? physical parameter identifiability in multimodal predictive representations. *arXiv preprint arXiv:2607.27017*, 2026. URL https://arxiv.org/abs/2607.27017.

The mathlib Community. The Lean mathematical library. In *Proceedings of the 9th ACM SIGPLAN International Conference on Certified Programs and Proofs*, pp. 367–381, 2020. doi: 10.1145/3372885.3373824. URL https://doi.org/10.1145/3372885.3373824.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer New York, 2009. doi: 10.1007/b13794. URL https://doi.org/10.1007/b13794.

Jan M. A. M. van Neerven. Second quantization and the L^p-spectrum of nonsymmetric ornstein–uhlenbeck operators. *Infinite Dimensional Analysis, Quantum Probability and Related Topics*, 8 (3):473–495, 2005. doi: 10.1142/S0219025705002074. URL https://doi.org/10.1142/S0219025705002074.

Matthew O. Williams, Ioannis G. Kevrekidis, and Clarence W. Rowley. A data-driven approximation of the koopman operator: Extending dynamic mode decomposition. *Journal of Nonlinear Science*, 25(6):1307–1346, 2015. doi: 10.1007/s00332-015-9258-5. URL https://doi.org/10.1007/s00332-015-9258-5.

Laurenz Wiskott and Terrence J. Sejnowski. Slow feature analysis: Unsupervised learning of invariances. *Neural Computation*, 14(4):715–770, 2002. doi: 10.1162/089976602317318938. URL https://doi.org/10.1162/089976602317318938.

Hao Wu and Frank Noé. Variational approach for learning markov processes from time series data. *Journal of Nonlinear Science*, 30(1):23–66, 2020. doi: 10.1007/s00332-019-09567-y. URL https://doi.org/10.1007/s00332-019-09567-y.

Xiangteng Zhang, Yang Guan, Bo Zhang, Hongyang Li, Ya-Qin Zhang, and Shengbo Eben Li. On the identifiability of controlled world models. *arXiv preprint arXiv:2607.22430*, 2026a. URL https://arxiv.org/abs/2607.22430.

Zhenghao Zhang, Yuanxiang Wang, Zhenyu Guan, Yujia Yang, Bingkang Shi, Tianyu Zong, Hongzhu Yi, Guoqing Chao, Xingchen Chen, Tiankun Yang, Chenxi Bao, Tao Yu, Jingjing Zhou, and Jungang Xu. Delta-JEPA: Learning action-sensitive world models via latent difference decoding. *arXiv preprint arXiv:2606.31232*, 2026b. URL https://arxiv.org/abs/2606.31232.

CONTENTS OF APPENDIX

I Representation phase and robustness . . . 15

A Guide to the appendix . . . 15

A.1 Standing notation and proof conventions . . . 15

B Proof of the sharp representation phase . . . 16

B.1 Overview of the argument . . . 16

B.2 Conditional replicas and the full Hermite expansion . . . 16

B.3 Why the strongest nonlinear competitor is quadratic . . . 17

B.4 The positive phase: global optimality and stability . . . 18

B.5 The negative phase: a smooth improving coordinate . . . 19

C Proof of the predictor-capacity reversal . . . 20

C.1 Overview . . . 20

C.2 The linear predictor is an orthogonal projection . . . 21

C.3 An explicit whitened and invertible cosine path . . . 21

C.4 Exact rich and linear score polynomials . . . 22

C.5 Certified reversal for the exact negative-phase law . . . 22

D From surrogate cost distortion to action-word regret . . . 23

E Robustness to approximate population whitening . . . 24

F Stability under approximate block closure . . . 26

G Exact nonlinear dynamics with a shared Gaussian stationary law . . . 27

H Exact Gaussianization does not guarantee identifiability . . . 29

II Matched laws, statistical limits, and design . . . 32

I Matched summaries, horizon sharpness, and open families . . . 32

I.1 Neutral directions are a kernel computation . . . 32

I.2 Lifting a separation to every longer horizon . . . 33

I.3 Why horizon four is the existential minimum . . . 33

I.4 The exact length-five witness . . . 33

I.5 An ambient-open family of witnesses . . . 35

J Local testing lower bound . . . 36

K Action-word design as a semidefinite program . . . 37

L Conditional estimation and data-driven word-law design . . . 37

III Reproducibility and machine-checked scope . . . 41

M Reproducibility and compute details . . . 41

N Lean verification . . . 46

Part I
Representation phase and robustness

A GUIDE TO THE APPENDIX

The main text follows one argument: the action-word law changes a spectral margin, the margin changes the preferred representation, and the representation changes what a word-conditioned linear latent predictor can express. The appendix follows the same order. We first develop the analytic ingredients shared by the phase results, then prove the positive and negative regimes, and only afterward construct matched laws and study what can be learned from finite data. Each long proof opens with the obstacle it must overcome and a short route through the needed lemmas, so the argument can be read independently of the implementation.

Three reading paths. For the central theorem, read Appendices B and I. For the finite-data argument, continue with Appendices J–L. Readers interested in robustness can move from Appendix B directly to Appendices E–F. Experimental and formalization details are self-contained in Part III.

A.1 STANDING NOTATION AND PROOF CONVENTIONS

Table 1 collects the symbols used across more than one proof. Local variables are introduced where they first appear.

Symbol	Meaning
Controlled Gaussian world	
$X, Y \in \mathbb{R}^d$	Current and future latent state; both have marginal law $\gamma_d = \mathcal{N}(0, I_d)$.
$a \in \mathcal{A}$	Primitive action, with linear transition matrix A_a.
$w = (a_1, \dots, a_L)$	Length-L action word; W denotes the sampled word.
$M_w = A_{a_L} \cdots A_{a_1}$	Ordered state-transition matrix induced by word w.
$q = (q_w)_w$	Action-word law, $q_w = \Pr(W = w)$; the behavior-design variable.
Representations and Gaussian chaos	
$h : \mathbb{R}^d \to \mathbb{R}^d$	Encoder, centered and population-whitened unless stated otherwise.
$\mathcal{H}_k$	Degree-k Hermite chaos in $L^2(\gamma_d)$; $\mathcal{H}_1$ is the linear state space.
Φ_k, C_k	Orthonormal degree-k feature map and the corresponding coefficient matrix of h.
Conditional operators and scores	
P_w	Conditional-mean operator, $P_w f(x) = \mathbb{E}[f(Y) \mid X = x, W = w]$.
$D_w = M_w M_w^\top$	First-chaos predictable-covariance operator for word w.
$T_k(q) = \sum_w q_w D_w^{\odot k}$	Word-averaged predictable-energy operator on $\mathcal{H}_k$.
$L_q = \sum_w q_w P_w^* P_w$	Full population predictable-energy operator.
$F_q(h)$	Rich conditional predictable score, estimated by replicated futures.
$S_q^{\mathrm{lin}}(h)$	Best score attainable by a word-conditioned linear latent predictor.
$\Delta(q)$	Phase margin $\lambda_{\min}(T_1(q)) - \lambda_{\max}(T_2(q))$.

Table 1: Notation used throughout the appendix. The symmetric tensor power is written $\odot k$; all operator norms use the orthonormal Hermite/Fock normalization.

All inner products and norms without further qualification are taken in $L^2(\gamma_d)$, where $\gamma_d = \mathcal{N}(0, I_d)$. The scalar Hermite chaoses are denoted by $\mathcal{H}_k$, with the orthogonal decomposition

$$L^2(\gamma_d) = \mathcal{H}_0 \oplus \mathcal{H}_1 \oplus \mathcal{H}_2 \oplus \cdots, \qquad \mathcal{H}_k \simeq \mathrm{Sym}^k(\mathbb{R}^d). \tag{A.1}$$

The identification in (A.1) uses the orthonormal Hermite, or symmetric Fock, normalization. This qualification is important: operator norms of normalized Hermite blocks are not the norms of raw monomial coefficient matrices. For a vector-valued encoder $h = (h_1, \dots, h_d)$, we apply (A.1) to

each output coordinate. The matrix C_k collects the coefficients of all output coordinates in $\mathcal{H}_k$; its rows therefore index encoder outputs and its columns index an orthonormal basis of $\mathcal{H}_k$.

For a word w, let $P_w f(x) = \mathbb{E}[f(Y) \mid X = x, W = w]$, put $D_w = M_w M_w^\top$, and define

$$T_k(q) = \sum_w q_w D_w^{\odot k}, \qquad L_q = \sum_w q_w P_w^* P_w. \tag{A.2}$$

The symbol $\odot k$ denotes the symmetric tensor power in the normalization of (A.1). Since $0 \preceq D_w \preceq I$, all operators in (A.2) are positive contractions. Equalities involving encoders are understood γ_d-almost surely unless pointwise regularity is explicitly stated. The exact rational constructions are finite-dimensional and are checked by rational arithmetic; decimal values are printed only to make the magnitude of the certified inequalities easy to read.

B Proof of the sharp representation phase

What must be proved. Theorem 1 makes a statement over the entire centered, population-whitened $L^2(\gamma_d)$ encoder class. When $\Delta(q) > 0$, every maximizer must use only first-order Hermite coordinates, and near-optimal encoders must be close to an orthogonal copy of the state. When $\Delta(q) < 0$, the converse must remain valid even after the competitor is required to be smooth, globally invertible, exactly whitened, and arbitrarily close to a linear coordinate system.

Why the conclusion is not a covariance calculation. A tempting argument is to compare T_1 and T_2 and stop. That would only rule out one quadratic candidate. A general measurable encoder can distribute its variance over infinitely many Hermite orders and mix those orders across output coordinates. The proof must therefore control every order at once, identify the equality case, and then build a nonlinear competitor that obeys the global geometric constraints. The lemmas below address these three issues in that order.

B.1 Overview of the argument

The proof has four steps. First, conditional replicas turn the population score into a squared conditional mean. Second, Gaussian second quantization diagonalizes that conditional mean by Hermite degree. Third, contraction of the word channels makes the block norms decrease with degree, reducing the global positive-phase proof to a comparison between T_1 and T_2. Finally, in the negative phase we replace one weak linear coordinate by a smooth even feature and verify directly that this replacement is whitened, invertible, arbitrarily small, and strictly improving.

B.2 Conditional replicas and the full Hermite expansion

We first record the exact population identity used throughout the paper.

Lemma 4 (Replica identity)**.** *Let Y_1, Y_2 be conditionally independent copies of Y given (X, W). For every square-integrable vector-valued encoder h,*

$$\mathbb{E}[h(Y_1)^\top h(Y_2) \mid X, W] = \|\mathbb{E}[h(Y) \mid X, W]\|^2. \tag{A.3}$$

Consequently,

$$F_q(h) = \sum_w q_w \sum_{j=1}^{d} \|P_w h_j\|_2^2 = \sum_w q_w \sum_{j=1}^{d} \langle h_j, P_w^* P_w h_j \rangle. \tag{A.4}$$

Proof. Condition on (X, W). Conditional independence factorizes every product of coordinates:

$$\mathbb{E}[h_j(Y_1) h_j(Y_2) \mid X, W] = \mathbb{E}[h_j(Y_1) \mid X, W] \mathbb{E}[h_j(Y_2) \mid X, W].$$

The two conditional expectations are the same. Summing over j proves (A.3), and integrating first over X and then over W proves (A.4). No Gaussian assumption is needed for this lemma; Gaussianity enters only when we diagonalize P_w. □

Lemma 5 (Gaussian channel on each chaos)**.** *For the stationary Gaussian channel $Y = M_w X + \xi_w$, the conditional-mean operator preserves Hermite degree and obeys*

$$P_w|_{\mathcal{H}_k} = (M_w^\top)^{\odot k}, \qquad P_w^* P_w|_{\mathcal{H}_k} = (M_w M_w^\top)^{\odot k}. \tag{A.5}$$

Proof. For $t \in \mathbb{R}^d$, introduce the normalized Gaussian exponential

$$\mathcal{E}_t(x) = \exp\left(t^\top x - \frac{1}{2}\|t\|^2\right).$$

Its Taylor expansion in t generates the orthonormal Hermite tensors of all degrees. Using the Gaussian moment-generating function and $\mathrm{Cov}(\xi_w) = I - M_w M_w^\top$,

$$\begin{aligned} P_w \mathcal{E}_t(x) &= \mathbb{E}\left[\exp\left(t^\top (M_w x + \xi_w) - \frac{1}{2}\|t\|^2\right)\right] \\ &= \exp\left(t^\top M_w x - \frac{1}{2} t^\top M_w M_w^\top t\right) = \mathcal{E}_{M_w^\top t}(x). \end{aligned}$$

Comparing homogeneous terms of degree k gives the first identity in (A.5). Taking the Hilbert adjoint and multiplying gives the second. On the first chaos, the same orientation can be seen without notation: for $f_a(y) = a^\top y$, $P_w f_a(x) = a^\top M_w x = (M_w^\top a)^\top x$, and therefore $\|P_w f_a\|_2^2 = a^\top M_w M_w^\top a$. This explains why the relevant positive operator is $M_w M_w^\top$, including when M_w is not symmetric. □

Proposition 6 (Non-truncated predictable-energy expansion)**.** *If h is centered and square integrable, write $h = \sum_{k\geq 1} C_k \Phi_k$ in orthonormal Hermite coordinates. Then*

$$F_q(h) = \sum_{k\geq 1} \mathrm{tr}\big(C_k T_k(q) C_k^\top\big). \tag{A.6}$$

If, in addition, h is population-whitened, then

$$\sum_{k\geq 1} C_k C_k^\top = I_d, \qquad \sum_{k\geq 1} \|C_k\|_F^2 = d. \tag{A.7}$$

Proof. Centering removes the constant chaos. Lemma 5 and Parseval give, for each coordinate h_j,

$$\|P_w h_j\|_2^2 = \sum_{k\geq 1} \langle c_{j,k}, D_w^{\odot k} c_{j,k}\rangle.$$

Every summand is nonnegative. Hence Tonelli's theorem permits exchanging the infinite sum with the finite sums over words and output coordinates. Summing the resulting quadratic forms gives (A.6). The series is finite because P_w is a contraction and $\sum_j \|h_j\|_2^2 < \infty$.

For (A.7), apply Parseval to the cross moment of output coordinates:

$$\mathbb{E}[h_i(X) h_j(X)] = \sum_{k\geq 1} \langle c_{i,k}, c_{j,k}\rangle.$$

The matrix of these identities is precisely $\sum_k C_k C_k^\top = I_d$; taking its trace yields the second equality. Thus the expansion is over all Hermite orders and does not rely on a finite degree cutoff. □

B.3 Why the strongest nonlinear competitor is quadratic

The next lemma is the step that collapses an infinite-dimensional comparison to the first two chaos levels.

Lemma 7 (All-order tail monotonicity)**.** *For every $k \geq 1$,*

$$\|T_{k+1}(q)\|_{\mathrm{op}} \leq \|T_k(q)\|_{\mathrm{op}}. \tag{A.8}$$

In particular, $\sup_{k\geq 2} \|T_k(q)\|_{\mathrm{op}} = \|T_2(q)\|_{\mathrm{op}}$.

Proof. Let $J_k : \mathrm{Sym}^{k+1}(\mathbb{R}^d) \to \mathrm{Sym}^k(\mathbb{R}^d) \otimes \mathbb{R}^d$ be the canonical isometric inclusion. In orthonormal occupation-number coordinates, write e_α for the basis vector indexed by $\alpha \in \mathbb{N}^d$ with $|\alpha| = k+1$, and set

$$J_k e_\alpha = \sum_{i:\alpha_i>0} \sqrt{\frac{\alpha_i}{k+1}}\, e_{\alpha-e_i} \otimes e_i.$$

Images belonging to different occupation indices are orthogonal, and for each α the squared coefficients sum to $\sum_i \alpha_i/(k+1) = 1$. Thus $J_k^* J_k = I$.

The intertwining relation can be seen first on pure symmetric powers:

$$J_k(x^{\odot(k+1)}) = x^{\odot k} \otimes x,$$

and therefore, for every linear map D,

$$J_k D^{\odot(k+1)} x^{\odot(k+1)} = (D^{\odot k} \otimes D) J_k x^{\odot(k+1)}.$$

Pure powers span $\mathrm{Sym}^{k+1}(\mathbb{R}^d)$ by polarization, so this identity holds on the whole space. Applying J_k^* on the left and using $J_k^* J_k = I$ gives, for each word,

$$D_w^{\odot(k+1)} = J_k^* (D_w^{\odot k} \otimes D_w) J_k. \tag{A.9}$$

Because $0 \preceq D_w \preceq I$, positivity of the tensor product gives

$$0 \preceq D_w^{\odot k} \otimes D_w \preceq D_w^{\odot k} \otimes I.$$

Multiply by $q_w \geq 0$, sum over words, and compress by J_k. For every unit $u \in \mathrm{Sym}^{k+1}(\mathbb{R}^d)$,

$$\begin{aligned}\langle u, T_{k+1} u\rangle &\leq \langle J_k u, (T_k \otimes I)\, J_k u\rangle \\ &\leq \|T_k \otimes I\|_{\mathrm{op}} \|J_k u\|^2 = \|T_k\|_{\mathrm{op}}.\end{aligned}$$

Taking the supremum over unit u proves (A.8). Iterating (A.8) shows that the largest norm among all nonlinear blocks occurs at $k = 2$. □

B.4 The positive phase: global optimality and stability

Let

$$\alpha = \lambda_{\min}(T_1), \qquad \nu = \|T_2\|_{\mathrm{op}}, \qquad \eta(h) = \sum_{k\geq 2} \|C_k\|_F^2 = d - \|C_1\|_F^2. \tag{A.10}$$

The last equality follows from whitening. The number $\eta(h)$ is the total encoded variance outside the linear state space.

Lemma 8 (First-chaos deficit)**.** *For every centered, whitened $h : \mathbb{R}^d \to \mathbb{R}^d$,*

$$\mathrm{tr}(C_1 T_1 C_1^\top) \leq \mathrm{tr}(T_1) - \alpha\eta(h). \tag{A.11}$$

Proof. Whitening implies $C_1 C_1^\top \preceq I$. Since C_1 is square, all its singular values are at most one and hence also $C_1^\top C_1 \preceq I$. Cyclicity of the trace and $T_1 \succeq \alpha I$ give

$$\begin{aligned}\mathrm{tr}(C_1 T_1 C_1^\top) &= \mathrm{tr}(T_1) - \mathrm{tr}\{T_1(I - C_1^\top C_1)\} \\ &\leq \mathrm{tr}(T_1) - \alpha\,\mathrm{tr}(I - C_1^\top C_1).\end{aligned}$$

The last trace is $d - \|C_1\|_F^2 = \eta(h)$, proving (A.11). □

Proof of the positive-phase part of Theorem 1. Proposition 6, Lemma 7, and Lemma 8 imply

$$\begin{aligned}F_q(h) &= \mathrm{tr}(C_1 T_1 C_1^\top) + \sum_{k\geq 2} \mathrm{tr}(C_k T_k C_k^\top) \\ &\leq \mathrm{tr}(T_1) - \alpha\eta(h) + \nu\eta(h) \\ &= \mathrm{tr}(T_1) - \Delta(q)\eta(h).\end{aligned} \tag{A.12}$$

To compare the two population scores, view conditional expectation as an orthogonal projection in L^2. Since $\sigma(h(X), W) \subseteq \sigma(X, W)$, projecting onto the smaller sigma-field cannot increase squared norm. Therefore $B_q(h) \le F_q(h)$.

Every $h(x) = Qx$, $Q \in O(d)$, is feasible and has all its mass in the first chaos. It attains $F_q(h) = \mathrm{tr}(T_1)$; because it is invertible, it also attains $B_q(h) = F_q(h)$. Suppose now that $\Delta > 0$ and either score attains this value. Inequality (A.12) forces $\eta(h) = 0$. Thus $h(X) = C_1 X$ almost surely, and whitening gives $C_1 C_1^\top = I$. Since the matrix is square, $C_1 \in O(d)$. This proves the claimed equality characterization for both scores.

For the quantitative claim, let $J \in \{F_q, B_q\}$ and assume $\mathrm{tr}(T_1) - J(h) \le \varepsilon$. The same inequality gives $\eta(h) \le \varepsilon/\Delta$. Choose an orthogonal polar factor Q of C_1. If $s_1, \dots, s_d \in [0, 1]$ are the singular values of C_1, then

$$\|C_1 - Q\|_F^2 = \sum_i (1 - s_i)^2 \le \sum_i (1 - s_i^2) = \eta(h).$$

The linear and nonlinear chaos components are orthogonal, so

$$\mathbb{E}\|h(X) - QX\|^2 = \|C_1 - Q\|_F^2 + \eta(h) \le 2\eta(h) \le \frac{2\varepsilon}{\Delta(q)}. \tag{A.13}$$

This proves global uniqueness and the stated stability bound over the complete centered, whitened L^2 encoder class. □

B.5 The negative phase: a smooth improving coordinate

The converse requires more than inserting a quadratic feature: the improved encoder must remain exactly whitened and globally invertible. We first replace the unbounded quadratic eigenfunction by a bounded smooth approximation, then use it in a triangular coordinate change.

Lemma 9 (A bounded even feature with the required Rayleigh quotient)**.** *If $\nu = \|T_2\|_{\mathrm{op}} > \alpha = \lambda_{\min}(T_1)$, there exists a bounded C^∞ function ϕ such that*

$$\mathbb{E}\phi(X) = 0, \qquad \|\phi\|_2 = 1, \qquad \phi \perp \mathcal{H}_1, \qquad \langle \phi, L_q \phi \rangle > \alpha, \qquad \|\nabla \phi\|_\infty < \infty. \tag{A.14}$$

Proof. Let $\psi \in \mathcal{H}_2$ be a unit top eigenfunction of T_2. A centered quadratic Gaussian chaos can be diagonalized as

$$\psi(x) = \sum_{j=1}^{d} c_j \{(v_j^\top x)^2 - 1\},$$

after absorbing normalization into the coefficients. For $\omega > 0$, set

$$g_\omega(x) = \sum_{j=1}^{d} c_j \{\cos(\omega v_j^\top x) - e^{-\omega^2/2}\}. \tag{A.15}$$

Each summand is centered under γ_d, bounded, smooth, and even. Consequently it is orthogonal to every linear Gaussian coordinate. Taylor's formula and Gaussian domination give

$$\omega^{-2} g_\omega \longrightarrow -\frac{1}{2}\psi \quad \text{in } L^2(\gamma_d). \tag{A.16}$$

After normalization, a choice of sign converges to ψ. The operator L_q is bounded, so the Rayleigh quotient is continuous under this L^2 convergence. Since $\langle \psi, L_q \psi \rangle = \nu > \alpha$, a sufficiently small fixed ω yields a normalized ϕ satisfying the strict inequality in (A.14). Being a finite cosine sum, this ϕ is bounded and has bounded gradient. □

Proof of the negative-phase part of Theorem 1. Rotate the input coordinates so that x_d is a unit eigenfunction of T_1 with eigenvalue α, and choose ϕ from Lemma 9. For $|\varepsilon| < 1$, let $a_\varepsilon = \sqrt{1 - \varepsilon^2}$ and define

$$h_\varepsilon(x) = (x_1, \dots, x_{d-1}, a_\varepsilon x_d + \varepsilon \phi(x)). \tag{A.17}$$

Step 1: exact centering and whitening. The first $d-1$ coordinates are unchanged. The last coordinate is centered because both x_d and ϕ are centered. It is uncorrelated with every linear coordinate because $\phi \perp \mathcal{H}_1$, and its variance is $a_\varepsilon^2 + \varepsilon^2 = 1$. Hence (A.17) is exactly population-whitened, not merely whitened to first order.

Step 2: global invertibility. Choose $|\varepsilon|$ small enough that

$$m := a_\varepsilon - |\varepsilon| \|\partial_d \phi\|_\infty > 0, \qquad K := |\varepsilon| \|\nabla_\perp \phi\|_\infty. \tag{A.18}$$

For each fixed $x_\perp = (x_1, \ldots, x_{d-1})$, the derivative of the last output coordinate with respect to x_d is at least m. The fiber map is therefore strictly increasing. Because ϕ is bounded, its leading term is $a_\varepsilon x_d$, so it tends to $\pm\infty$ with x_d. Each fiber is thus a bijection $\mathbb{R} \to \mathbb{R}$, and the first $d-1$ output coordinates already reveal $x_\perp$. This proves global bijectivity.

For a quantitative inverse bound, the mean-value theorem gives, for all $x, y \in \mathbb{R}^d$,

$$|x_d - y_d| \le m^{-1}(|h_{\varepsilon,d}(x) - h_{\varepsilon,d}(y)| + K\|x_\perp - y_\perp\|)\,. \tag{A.19}$$

Together with $h_{\varepsilon,\perp}(x) = x_\perp$, this yields, for example,

$$\mathrm{Lip}(h_\varepsilon^{-1}) \le \sqrt{1 + \frac{1+K^2}{m^2}}. \tag{A.20}$$

The triangular Jacobian has last diagonal entry at least m, so the implicit function theorem makes the inverse smooth. Since $\nabla\phi$ is bounded, h_ε is also globally Lipschitz. Thus (A.17) is a global smooth bi-Lipschitz diffeomorphism.

Step 3: exact score gain. The operator L_q preserves Hermite order. Hence its cross term between the linear coordinate x_d and the even function ϕ vanishes. Only the last output coordinate changes, and direct expansion gives

$$F_q(h_\varepsilon) - F_q(I) = \varepsilon^2\{\langle \phi, L_q\phi\rangle - \alpha\} > 0. \tag{A.21}$$

Moreover,

$$\|h_\varepsilon - I\|_{L^2}^2 = (a_\varepsilon - 1)^2 + \varepsilon^2 \longrightarrow 0, \tag{A.22}$$

so the improvement occurs arbitrarily close to the linear encoder within the exactly whitened feasible class.

Step 4: transfer to the profiled score. Global invertibility implies $\sigma(h_\varepsilon(X), W) = \sigma(X, W)$. Conditioning on the learned representation therefore loses no information, and

$$B_q(h_\varepsilon) = F_q(h_\varepsilon) > F_q(I) = B_q(I).$$

Finally, left composition by any $Q \in O(d)$ preserves centering, whitening, both scores, and bi-Lipschitz regularity. It transports the same construction to an arbitrarily small neighborhood of every orthogonal state encoder. This completes the proof of Theorem 1. □

C Proof of the predictor-capacity reversal

C.1 Overview

The representation theorem compares two ideal population scores. This section asks a different question: after the representation has been chosen, how much of its predictable signal can a linear latent dynamics model express? The proof has three parts. We first identify the gap to the Bayes score as a Hilbert-space projection residual. We then construct an exactly whitened two-dimensional cosine warp for which both the rich and linear scores are explicit quadratic polynomials. Finally, certified interval arithmetic fixes their opposite rankings for the exact negative-phase word law.

The distinction is one of information access. The rich score conditions on the full current state, whereas the linear predictor can use only linear combinations of the coordinates already exposed by $h(X)$. An invertible warp may therefore preserve all state information and improve the rich score, yet rotate the conditional mean outside the finite-dimensional coordinate span seen by the linear predictor. The projection identity below isolates exactly that lost component before any special construction is introduced.

C.2 THE LINEAR PREDICTOR IS AN ORTHOGONAL PROJECTION

Fix a centered, whitened encoder and write $V_h = \operatorname{span}\{h_1, \dots, h_d\}$. Its coordinates form an orthonormal basis of V_h. For word w, define $C_w = \mathbb{E}[h(Y)h(X)^\top \mid W = w]$.

Lemma 10 (Projection identity)**.** *For every centered, whitened encoder,*

$$S_q^{\mathrm{lin}}(h) = \sum_w q_w \|\Pi_{V_h} P_w|_{V_h}\|_{\mathrm{HS}}^2, \qquad F_q(h) = \sum_w q_w \|P_w|_{V_h}\|_{\mathrm{HS}}^2,$$

and consequently

$$F_q(h) - S_q^{\mathrm{lin}}(h) = \sum_w q_w \|(I - \Pi_{V_h}) P_w|_{V_h}\|_{\mathrm{HS}}^2. \tag{A.23}$$

The minimum word-conditioned linear prediction error is

$$\inf_{A_w} \mathbb{E}[\|h(Y) - A_w h(X)\|^2 \mid W = w] = d - \|C_w\|_F^2. \tag{A.24}$$

Proof. The (i, j) entry of C_w is

$$\mathbb{E}[h_i(Y)h_j(X) \mid W = w] = \langle P_w h_i, h_j \rangle.$$

Thus C_w, up to the harmless choice of transpose, is the matrix of $\Pi_{V_h} P_w|_{V_h}$ in the orthonormal coordinate basis. Its squared Frobenius norm is the corresponding squared Hilbert–Schmidt norm. Meanwhile, $\sum_i \|P_w h_i\|_2^2 = \|P_w|_{V_h}\|_{\mathrm{HS}}^2$. Since $P_w|_{V_h} = \Pi_{V_h} P_w|_{V_h} + (I - \Pi_{V_h}) P_w|_{V_h}$ is an orthogonal decomposition, Pythagoras proves (A.23).

For (A.24), expand the conditional squared error:

$$\mathbb{E}\|h(Y) - A_w h(X)\|^2 = d - 2\operatorname{tr}(A_w C_w^\top) + \operatorname{tr}(A_w A_w^\top),$$

where whitening gives both marginal covariance matrices equal to I. Completing the square shows that the unique minimum is attained at $A_w = C_w$, with value $d - \|C_w\|_F^2$. If h is invertible, conditioning on $h(X)$ is equivalent to conditioning on X, so the Bayes error is $d - F_q(h)$. Equation (A.23) is therefore exactly the excess error caused by restricting the latent predictor to be linear. □

C.3 AN EXPLICIT WHITENED AND INVERTIBLE COSINE PATH

We now evaluate this residual on the graph path in Proposition 3. Put $U = u^\top X$, $V = v^\top X$, and define

$$\phi(z) = \frac{\cos(z/\sqrt{2}) - e^{-1/4}}{(1 - e^{-1/2})/\sqrt{2}}, \qquad h_t(x) = \left(v^\top x, \sqrt{1-t}\, u^\top x + \sqrt{t}\, \phi(v^\top x)\right), \quad 0 \le t < 1.$$

Here u is the weakest first-chaos direction and $v \perp u$. Under the standard Gaussian state, U, V are independent $\mathcal{N}(0, 1)$.

Lemma 11 (Normalization and invertibility of the cosine path)**.** *For every $0 \le t < 1$, h_t is centered, exactly population-whitened, smooth, and globally invertible. Its inverse in the (V, U) coordinates is*

$$h_t^{-1}(z_1, z_2) = \left(z_1, \frac{z_2 - \sqrt{t}\, \phi(z_1)}{\sqrt{1-t}}\right). \tag{A.25}$$

Proof. For $G \sim \mathcal{N}(0, 1)$,

$$\mathbb{E}\cos(G/\sqrt{2}) = e^{-1/4}, \qquad \mathbb{E}\cos^2(G/\sqrt{2}) = \frac{1 + e^{-1}}{2}.$$

Therefore the numerator defining ϕ is centered and has variance $(1 - e^{-1/2})^2/2$, exactly the square of its denominator. Hence $\mathbb{E}\phi(G) = 0$ and $\mathbb{E}\phi(G)^2 = 1$. Since ϕ is even, $\mathbb{E}[G\phi(G)] = 0$. Independence of U, V now shows that the two coordinates of $h_t(X)$ are centered, have unit variance, and are uncorrelated. This is exact whitening for every t, not an asymptotic calculation.

The first output coordinate reveals V. Substituting it into the second coordinate recovers U by (A.25). The denominator is positive for $t < 1$, so this is a smooth global inverse. □

C.4 Exact rich and linear score polynomials

For each word define

$$r_w = u^\top M_w u, \quad b_w = v^\top M_w u, \quad c_w = \langle \phi, P_w \phi \rangle, \quad s_w = \|P_w \phi\|_2^2.$$

The quantity s_w is the full predictable energy of the cosine feature, whereas c_w^2 is the part that remains in its one-dimensional span. Odd and even Hermite orders are orthogonal, so all mixed linear–cosine terms that are not displayed below vanish.

Lemma 12 (Score polynomials along the graph path)**.** *Define*

$$\delta_q = \sum_w q_w s_w - \lambda_{\min}(T_1),$$

$$K_q = \sum_w q_w \{b_w^2 + (r_w - c_w)^2 + s_w - c_w^2\}, \qquad D_q = \sum_w q_w (r_w - c_w)^2, \tag{A.26}$$

Then

$$F_q(h_t) - F_q(h_0) = t\delta_q, \qquad F_q(h_t) - S_q^{\mathrm{lin}}(h_t) = tK_q - t^2 D_q. \tag{A.27}$$

Consequently,

$$S_q^{\mathrm{lin}}(h_t) - S_q^{\mathrm{lin}}(h_0) = t(\delta_q - K_q) + t^2 D_q. \tag{A.28}$$

Proof. Write $a = \sqrt{1-t}$ and $b = \sqrt{t}$. The second learned coordinate is $aU + b\phi(V)$, an orthogonal mixture of a first-chaos function and an even function. Since every Gaussian conditional-mean operator preserves Hermite degree, its predictable squared norm is $a^2\|P_w U\|_2^2 + b^2\|P_w \phi(V)\|_2^2$. The first term removes $t\lambda_{\min}(T_1)$ after averaging over w, while the second adds $t\sum_w q_w s_w$, proving the first identity in (A.27).

For the linear score, expand the 2×2 cross-covariance matrix between the two coordinates of $h_t(Y)$ and $h_t(X)$. To expose every term, put

$$d_w = v^\top M_w v, \qquad \ell_w = u^\top M_w v, \qquad e_w = \|P_w U\|_2^2 = r_w^2 + \ell_w^2.$$

The last equality uses that (u, v) is the chosen orthonormal basis in the $d = 2$ construction. Odd–even orthogonality removes every inner product between the linear functions $P_w U, P_w V$ and the even functions $\phi, P_w \phi$. Hence the wordwise cross-covariance matrix is exactly

$$C_w(t) = \begin{pmatrix} d_w & ab_w \\ a\ell_w & a^2 r_w + b^2 c_w \end{pmatrix}, \qquad a = \sqrt{1-t}, \quad b = \sqrt{t}. \tag{A.28a}$$

At $t = 0$, this is $C_w(0) = \left[\begin{smallmatrix} d_w & b_w \\ \ell_w & r_w \end{smallmatrix}\right]$. Expanding the two Frobenius norms gives

$$\|C_w(t)\|_F^2 - \|C_w(0)\|_F^2 = -t(b_w^2 + \ell_w^2) + 2tr_w(c_w - r_w) + t^2(c_w - r_w)^2. \tag{A.28b}$$

The corresponding full-score change is $t(s_w - e_w)$. Subtracting (A.28b) from it, and using $e_w = r_w^2 + \ell_w^2$, yields the wordwise linearity tax

$$t\{b_w^2 + (r_w - c_w)^2 + s_w - c_w^2\} - t^2(r_w - c_w)^2.$$

Summing with weights q_w proves the second identity. The term $s_w - c_w^2 = \|(I - \Pi_{\mathrm{span}\{\phi\}})P_w\phi\|_2^2$ is nonnegative by Pythagoras. Thus (A.27) is exactly the linear-closure residual from Lemma 10, rather than a formal subtraction of two unrelated objectives. Subtracting the two identities in (A.27) gives (A.28). □

C.5 Certified reversal for the exact negative-phase law

For the exact negative law, outward-rounded interval arithmetic gives

$$\delta_- = 0.0185162388\ldots, \qquad K_- = 0.1595089643\ldots, \qquad D_- = 0.1480502037\ldots.$$

The linear score is below its state value for $0 < t < (K_- - \delta_-)/D_- = 0.9523305\ldots$, and its minimum occurs at

$$t_\star = \frac{K_- - \delta_-}{2D_-} = 0.4761653\ldots.$$

Table 2: Outward-rounded interval bounds for the exact negative-phase law. Every endpoint is rounded away from the certified value.

Quantity	Certified enclosure	Role in the proof
δ_-	$[0.01851623882,\ 0.01851623883]$	rich-score slope
K_-	$[0.15950896431,\ 0.15950896432]$	linearity-tax slope
D_-	$[0.14805020372,\ 0.14805020373]$	quadratic correction
$(K_- - \delta_-)/D_-$	$[0.95233050644,\ 0.95233050645]$	reversal endpoint
$t_\star$	$[0.47616525322,\ 0.47616525323]$	linear-score minimizer
rich-score change	$[0.00881678954,\ 0.00881678956]$	strict improvement
linear-score change	$[-0.03356791843,\ -0.03356791841]$	strict degradation
$\mathrm{Lip}(h_{t_\star})$	$[2.14463,\ 2.14465]$	conditioning check
$\mathrm{Lip}(h_{t_\star}^{-1})$	$[2.96317,\ 2.96319]$	inverse conditioning

Table 2 records the outward enclosures used in this conclusion. The table summarizes the interval calculation used below; it is not a separate floating-point experiment.

At this point, (A.27)–(A.28) yield the two score changes stated in the proposition. Because the state score is $\mathrm{tr}(T_1) = 0.2254825\ldots$, the linear loss is 14.887% of its predictable energy.

Finally, $\|\phi'\|_\infty \le L_\phi = (1 - e^{-1/2})^{-1} = 2.541494\ldots$. In the orthonormal (U, V) coordinates, the triangular Jacobian and its inverse satisfy

$$\mathrm{Lip}(h_t) \le \sqrt{1 + (1-t) + tL_\phi^2}, \qquad \mathrm{Lip}(h_t^{-1}) \le \sqrt{1 + \frac{1 + tL_\phi^2}{1-t}}, \tag{A.28c}$$

which give the displayed constants at $t_\star$: $\mathrm{Lip}(h_{t_\star}) < 2.15$ and $\mathrm{Lip}(h_{t_\star}^{-1}) < 2.97$. The conditional cosine needed by the Bayes head is explicit:

$$P_w\phi(V)(x) = \frac{e^{-(1-\|M_w^\top v\|^2)/4}\cos((M_w^\top v)^\top x/\sqrt{2}) - e^{-1/4}}{(1 - e^{-1/2})/\sqrt{2}}.$$

Together with the triangular inverse, this realizes the Bayes response with a constant-width linear–cosine head. Nested raw-channel Gauss–Hermite quadrature agrees with (A.27)–(A.28) to 2.1×10^{-16}, while the reported signs, threshold, and Lipschitz bounds are checked by outward-rounded intervals.

To finish the proposition, differentiate (A.28). Since $D_- > 0$, the linear score change is a convex quadratic in t, negative precisely on $0 < t < (K_- - \delta_-)/D_-$, and minimized at $t_\star = (K_- - \delta_-)/(2D_-)$. The rich score change is the positive linear function $t\delta_-$. Substitution at $t_\star$ gives the reported gain $0.0088167\ldots$ and linear-score decrease $-0.0335679\ldots$. Dividing the latter magnitude by $\mathrm{tr}(T_1) = 0.2254825\ldots$ gives 14.887%. Lemma 11 and the Lipschitz bounds above prove that this ranking reversal occurs without discarding state information or using an ill-conditioned coordinate map.

Figure 3 summarizes this capacity reversal alongside the separate finite-sample recovery experiment.

D From surrogate cost distortion to action-word regret

The representation phase concerns coordinates, whereas the experiment in Section 4 evaluates a discrete decision. The following elementary lemma makes the positive-phase connection explicit without claiming that every nonlinear reparameterization must change the selected word.

Lemma 13 (Uniform cost distortion controls selection regret)**.** *Let $\mathcal{W}$ be a finite nonempty word set with state costs $J_x : \mathcal{W} \to \mathbb{R}$ and surrogate costs $J_h : \mathcal{W} \to \mathbb{R}$. Choose $\widehat{w} \in \arg\min_w J_h(w)$ and $w^\star \in \arg\min_w J_x(w)$. If $a > 0$, $\epsilon \ge 0$, and $b \in \mathbb{R}$ satisfy*

$$\sup_{w \in \mathcal{W}} |J_h(w) - (aJ_x(w) + b)| \le \epsilon,$$

then

$$0 \le J_x(\widehat{w}) - J_x(w^\star) \le \frac{2\epsilon}{a}.$$

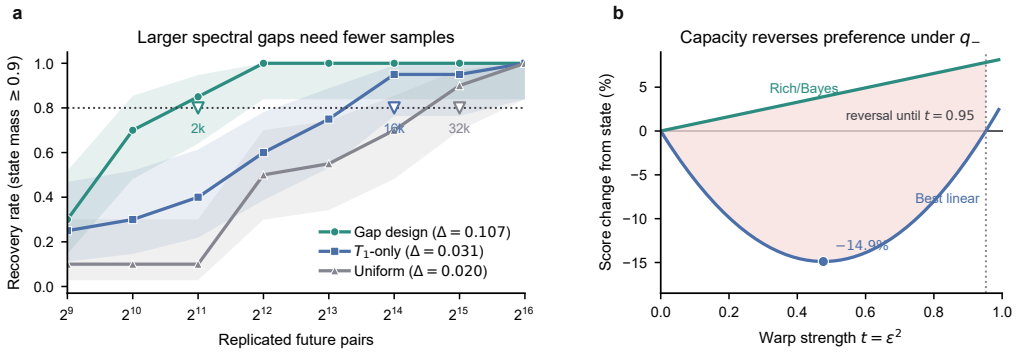

Figure 3: **Finite-sample recovery and the predictor-capacity reversal. a**, Fraction of 20 runs recovering the two-dimensional state subspace as the number of replica pairs grows. Optimizing Δ reaches 80% recovery with eight times fewer samples than optimizing linear signal alone and sixteen times fewer than uniform behavior; shading gives 95% Wilson intervals. This panel uses the fixed-action unconstrained-simplex design, distinct from the matched-moment prospective design in Figure 6. **b**, Along the invertible coordinate path under q_-, the rich objective increases while the future variance accessible to a word-conditioned linear predictor decreases. Scores are normalized by the predictable energy of the original state coordinates.

If $R_x = \max_w J_x(w) - \min_w J_x(w) > 0$*, the normalized regret is at most* $\min\{1, 2\epsilon/(aR_x)\}$.

Proof. Optimality of $\widehat{w}$ for J_h, followed by the two uniform approximation inequalities, gives

$$aJ_x(\widehat{w}) + b \le J_h(\widehat{w}) + \epsilon \le J_h(w^\star) + \epsilon \le aJ_x(w^\star) + b + 2\epsilon.$$

Division by $a > 0$ proves the upper bound, while optimality of $w^\star$ proves nonnegativity. Dividing by R_x gives the normalized bound; it is also at most one because both selected costs lie between the minimum and maximum state costs. □

For the costs used in Section 4,

$$J_x(w) = \mathbb{E}\|Y_w - x_{\text{goal}}\|^2.$$

An exact positive-phase coordinate has $\bar{h}(z) = Qz$ with $Q^\top Q = I$, and hence

$$J_h(w) = \mathbb{E}\|Q(Y_w - x_{\text{goal}})\|^2 = J_x(w).$$

Lemma 13 then applies with $(a, b, \epsilon) = (1, 0, 0)$: the population-optimal positive-phase coordinate selects a state-optimal word. Learned encoders need not satisfy a uniform cost approximation merely from an L^2 representation bound, so the nonzero negative-phase regret remains an empirical consequence rather than a theorem forced by $\Delta < 0$. Figure 2c and Table 6 report the seven-pair pixel–LeWorldModel cohort; Figure 7 is a separate five-pair direct-observation quadrature control cohort.

E Robustness to approximate population whitening

Exact population whitening makes the phase theorem transparent, but practical objectives enforce it only approximately. The result below separates the normalization issue from the representation issue: first rewhiten the encoder at the population level, apply Theorem 1, and then propagate the covariance distortion through the score. The proof also records the equality case, which is stronger than a one-sided robustness bound.

Corollary 14 (Approximate population whitening)**.** *Keep the assumptions of Theorem 1 except exact whitening, and fix* $0 \le \tau < 1$*. Let* $\mathcal{C}_\tau$ *contain the centered encoders whose population covariance* Σ_h *satisfies* $\|\Sigma_h - I\|_{\text{op}} \le \tau$*, and define* $\bar{h} = \Sigma_h^{-1/2} h$*. If* $\Delta(q) > 0$*, then, for* $J \in \{F_q, B_q\}$*,*

$$J(h) \le (1 + \tau)\{\operatorname{tr} T_1(q) - \Delta(q)\eta(\bar{h})\}.$$

The maximum is $(1+\tau)\operatorname{tr} T_1(q)$*, attained exactly by* $h(x) = \sqrt{1+\tau}\,Qx$ *almost surely,* $Q \in O(d)$*. If a feasible* h *is* ε*-suboptimal relative to this ceiling, then*

$$\inf_{Q\in O(d)} \mathbb{E}\|\bar{h}(X) - QX\|^2 \le \frac{2\varepsilon}{(1+\tau)\Delta(q)}.$$

For the same minimizing Q*,*

$$\mathbb{E}\|h(X) - \Sigma_h^{1/2} QX\|^2 \le \frac{2\varepsilon}{\Delta(q)}.$$

The negative-phase improving path remains feasible after isotropic rescaling.

Proof. **Step 1: exact rewhitening.** Let $R = \Sigma_h^{1/2}$. The condition $\tau < 1$ makes R invertible, and $\bar{h} = R^{-1}h$ is exactly centered and covariance-whitened. For the replicated-future score, put

$$m_F = \mathbb{E}[\bar{h}(Y) \mid X, W].$$

Linearity of conditional expectation gives

$$F_q(h) = \mathbb{E}[m_F^\top \Sigma_h m_F] \le (1+\tau) F_q(\bar{h}).$$

For the profiled score, invertibility implies $\sigma(h(X), W) = \sigma(\bar{h}(X), W)$. Hence, with $m_B = \mathbb{E}[\bar{h}(Y) \mid \bar{h}(X), W]$,

$$B_q(h) = \mathbb{E}[m_B^\top \Sigma_h m_B] \le (1+\tau) B_q(\bar{h}).$$

Applying Theorem 1 to $\bar{h}$ proves the displayed upper bound in Corollary 14. The encoder $\sqrt{1+\tau}\,QX$ attains the resulting ceiling.

Step 2: characterize equality. If equality holds, the exact-whitening theorem first forces $\bar{h} = QX$. The conditional-mean covariance is then $QT_1Q^\top$, which is positive definite because $\Delta > 0$ implies $\lambda_{\min}(T_1) > 0$. Equality in

$$\operatorname{tr}\{\Sigma_h Q T_1 Q^\top\} \le (1+\tau)\operatorname{tr} T_1$$

therefore forces $\Sigma_h = (1+\tau)I$, proving the equality characterization.

To see the last implication explicitly, diagonalize the positive-definite matrix $QT_1Q^\top$. The constraint $\Sigma_h \preceq (1+\tau)I$ makes every summand in $\operatorname{tr}\{((1+\tau)I - \Sigma_h)QT_1Q^\top\}$ nonnegative. Equality of the trace therefore forces $(1+\tau)I - \Sigma_h = 0$. Thus the only maximizers are the isotropically enlarged orthogonal state coordinates stated in the corollary.

Step 3: transfer the quantitative bound. If the ceiling deficit is at most ε, the same upper bound yields

$$\eta(\bar{h}) \le \frac{\varepsilon}{(1+\tau)\Delta}.$$

Equation (A.13) gives the first distance estimate. For its minimizing Q,

$$\mathbb{E}\|h - \Sigma_h^{1/2} QX\|^2 = \mathbb{E}\|R(\bar{h} - QX)\|^2 \le (1+\tau)\mathbb{E}\|\bar{h} - QX\|^2,$$

which gives the second.

Step 4: preserve the negative phase. Finally, when $\Delta < 0$, let $\bar{h}_t$ be the exactly whitened improving path from the proof of Theorem 1. For any $c \in [1-\tau, 1+\tau]$, the path $h_t = \sqrt{c}\,\bar{h}_t$ stays in $\mathcal{C}_\tau$, preserves smooth global bi-Lipschitz state-sufficiency, and satisfies

$$J(h_t) = cJ(\bar{h}_t) > cJ(QX) = J(\sqrt{c}\,QX)$$

for $J \in \{F_q, B_q\}$. Centering and $\tau < 1$ are essential: without centering, adding an arbitrary constant increases both scores, while at $\tau = 1$ the covariance may lose a state direction. □

F STABILITY UNDER APPROXIMATE BLOCK CLOSURE

The exact Gaussian analysis makes the first chaos and the nonlinear tail orthogonal invariant blocks. The following operator statement records what survives when this block structure is only approximate. It is a stability support for Theorem 1, not an additional identifiability theorem for a particular non-Gaussian model.

Proposition 15 (Approximate-block stability)**.** *Let $\mathcal{H} = \mathcal{S} \oplus \mathcal{N}$ be a Hilbert space with $\dim \mathcal{S} = d < \infty$, and let the bounded self-adjoint operator*

$$\mathcal{L} = \begin{bmatrix} A & B \\ B^* & C \end{bmatrix}$$

satisfy $A \succeq aI_{\mathcal{S}}$, $C \preceq cI_{\mathcal{N}}$, and $g = a - c > 0$. Put $b = \|B\|_{\mathrm{op}}$. Then $\mathcal{L}$ has a unique upper spectral subspace $\mathcal{U}$ of dimension d, separated from its orthogonal complement by a gap at least g, and

$$\theta_{\max}(\mathcal{U}, \mathcal{S}) \le \frac{1}{2} \arctan \frac{2b}{g}, \qquad \|P_{\mathcal{U}} - P_{\mathcal{S}}\|_{\mathrm{op}} \le \sqrt{\frac{1 - g/\sqrt{g^2 + 4b^2}}{2}}. \tag{F.1}$$

Both bounds in (F.1) *are attained by a two-dimensional scalar-block example.*

If $\widehat{\mathcal{L}} = \mathcal{L} + E$, where $E = E^*$ and $e = \|E\|_{\mathrm{op}} < g/2$, its upper d-dimensional spectral subspace $\widehat{\mathcal{U}}$ remains unique, its gap is at least $g - 2e$, and

$$\theta_{\max}(\widehat{\mathcal{U}}, \mathcal{U}) \le \frac{1}{2} \arcsin \frac{2e}{g}. \tag{F.2}$$

The gap constant 2 and the angle in (F.2) *are sharp.*

Proof. Fix $\lambda \in (c, a)$ and write $A_\lambda = A - \lambda I \succ 0$. The Schur factorization

$$\mathcal{L} - \lambda I = \begin{bmatrix} I & 0 \\ B^* A_\lambda^{-1} & I \end{bmatrix} \begin{bmatrix} A_\lambda & 0 \\ 0 & C - \lambda I - B^* A_\lambda^{-1} B \end{bmatrix} \begin{bmatrix} I & A_\lambda^{-1} B \\ 0 & I \end{bmatrix}$$

has a strictly positive d-dimensional first block and a strictly negative second block. Hence $\sigma(\mathcal{L}) \cap (c, a) = \varnothing$, and the spectral projection above any point in this interval has rank d. Equivalently, the upper cluster lies in $[a, \infty)$, the remaining spectrum lies in $(-\infty, c]$, and the intervening gap is at least g.

We next justify the graph representation used below. Let $\pi_{\mathcal{S}} : \mathcal{U} \to \mathcal{S}$ be the coordinate projection. If $u = (0, n) \in \mathcal{U}$, then $\langle u, \mathcal{L}u \rangle = \langle n, Cn \rangle \le c\|n\|^2$, contradicting that every nonzero vector in the upper spectral subspace has Rayleigh quotient at least a. Hence $\pi_{\mathcal{S}}$ is injective. Because $\dim \mathcal{U} = \dim \mathcal{S} = d$, it is bijective, and the upper spectral subspace is the graph of a unique bounded operator $X : \mathcal{S} \to \mathcal{N}$. Let $t = \|X\|_{\mathrm{op}}$, and choose a maximizing singular pair s, n with $Xs = tn$ and $X^* n = ts$. The corresponding principal vectors are

$$u = \frac{(s, tn)}{\sqrt{1 + t^2}} \in \mathcal{U}, \qquad v = \frac{(-ts, n)}{\sqrt{1 + t^2}} \in \mathcal{U}^\perp.$$

Both spectral subspaces reduce the self-adjoint operator, so $\langle \mathcal{L}u, v \rangle = 0$. Expanding the four block terms and taking real parts gives

$$t\{\langle As, s \rangle - \langle Cn, n \rangle\} = (1 - t^2) \operatorname{Re} \langle s, Bn \rangle.$$

The case $t = 1$ is impossible because the left side is strictly positive. If $t > 1$, the same identity gives, for the corresponding unit vector $u = (s, tn)/\sqrt{1 + t^2} \in \mathcal{U}$,

$$\langle u, \mathcal{L}u \rangle = \langle Cn, n \rangle + \frac{\langle As, s \rangle - \langle Cn, n \rangle}{1 - t^2} < c,$$

contradicting that the upper cluster lies in $[a, \infty)$. Therefore $t < 1$, and $gt \le b(1 - t^2)$. Solving this quadratic inequality yields

$$\tan \theta_{\max}(\mathcal{U}, \mathcal{S}) = t \le \frac{2b}{g + \sqrt{g^2 + 4b^2}},$$

which is equivalent to (F.1).

For the perturbation claim, consider $\mathcal{L}_\rho = \mathcal{L} + \rho E$, $0 \le \rho \le 1$. Its two spectral clusters remain separated by at least $g - 2\rho e > 0$, so the upper projection varies continuously and keeps rank d. For a largest principal angle θ_ρ between its range and $\mathcal{U}$, choose associated unit vectors $p \in \mathcal{U}$ and $q \in \mathcal{U}^\perp$, and set

$$u_\rho = \cos\theta_\rho\, p + \sin\theta_\rho\, q, \qquad v_\rho = -\sin\theta_\rho\, p + \cos\theta_\rho\, q.$$

Then $u_\rho \in \mathcal{U}_\rho$ and $v_\rho \in \mathcal{U}_\rho^\perp$. Because $\mathcal{U}$ reduces $\mathcal{L}$, its two restrictions are separated by at least g, so

$$|\langle \mathcal{L}u_\rho, v_\rho\rangle| = \frac{\sin(2\theta_\rho)}{2}|\langle \mathcal{L}p, p\rangle - \langle \mathcal{L}q, q\rangle| \ge \frac{g}{2}\sin(2\theta_\rho).$$

Invariance of $\mathcal{U}_\rho$ under $\mathcal{L}_\rho = \mathcal{L} + \rho E$ gives

$$|\langle \mathcal{L}u_\rho, v_\rho\rangle| = \rho|\langle Eu_\rho, v_\rho\rangle| \le \rho e.$$

Continuity from $\theta_0 = 0$ selects the smaller branch throughout the path: it cannot cross $\pi/4$, because at such a crossing the left side would equal $g/2 > \rho e$. Therefore $2\theta_\rho \le \arcsin(2\rho e/g)$. Setting $\rho = 1$ proves (F.2). The scalar-block matrix $\left[\begin{smallmatrix} a & b \\ b & c \end{smallmatrix}\right]$ attains (F.1). For the perturbation bound, center $\mathcal{L} = \mathrm{diag}(g/2, -g/2)$, put $r_0 = g/2$, and take

$$E = \begin{bmatrix} -e^2/r_0 & e\sqrt{1 - e^2/r_0^2} \\ e\sqrt{1 - e^2/r_0^2} & e^2/r_0 \end{bmatrix}.$$

Then $\|E\|_{\mathrm{op}} = e$ and equality holds in (F.2); the diagonal perturbation $\mathrm{diag}(-e, e)$ attains the gap $g - 2e$. □

Application to the representation phase. For the exact operator

$$L_q = T_1(q) \oplus \bigoplus_{k\ge 2} T_k(q) \quad \text{on} \quad \mathcal{H}_1 \oplus \mathcal{H}_{\ge 2},$$

tail collapse gives $a = \lambda_{\min}(T_1)$ and $c = \|T_2\|_{\mathrm{op}}$. Thus, if $\Delta > 0$ and a self-adjoint approximate or empirical operator $\widetilde{L}$ obeys $\|\widetilde{L} - L_q\|_{\mathrm{op}} = e < \Delta/2$, its top-$d$ subspace is unique, has gap at least $\Delta - 2e$, and lies within angle $\frac{1}{2}\arcsin(2e/\Delta)$ of $\mathcal{H}_1$. If instead $\delta = \|T_2\|_{\mathrm{op}} - \lambda_{\min}(T_1) > 0$, replacing a weakest unit state direction by a strongest unit second-chaos direction improves the $\widetilde{L}$-trace objective by at least $\delta - 2e$. Consequently, the positive state-versus-tail ordering and the negative second-chaos advantage for the replicated-future operator survive operator error below half their respective margins. This statement preserves spectral ordering; it does not by itself extend the smooth invertible converse to a non-Gaussian model. Establishing an operator-norm approximation for a concrete observation or non-Gaussian model is a separate modeling step.

G Exact nonlinear dynamics with a shared Gaussian stationary law

The preceding proposition is useful only if a concrete nonlinear world can satisfy its operator-norm premise. This section supplies such a family while keeping the stationary distribution exactly fixed. It therefore changes one assumption at a time: the primitive conditional mean becomes nonlinear, but neither behavior law receives a different state marginal.

For an action a, let K_a denote the conditional-mean operator of the linear Gaussian channel

$$Y = A_a X + \xi_a, \qquad \xi_a \sim \mathcal{N}(0, I - A_a A_a^\top), \qquad \|A_a\|_{\mathrm{op}} < 1.$$

Let $\Phi_{a,t}$ be a family of Gaussian-measure-preserving bijections with $\Phi_{a,0} = I$, and write $U_{a,t} f = f \circ \Phi_{a,t}$. Replacing the primitive channel by

$$Y = A_a \Phi_{a,t}(X) + \xi_a \tag{G.1}$$

gives the conditional operator $P_{a,t} = U_{a,t} K_a$.

Proposition 16 (Strict phases persist under exact nonlinear stationary perturbations)**.** *Suppose the action set and word horizon are finite, every A_a is a strict contraction, and $U_{a,t} \to I$ strongly on $L^2(\gamma_d)$ as $t \to 0$. Let $P_{w,t}$ be the word operator obtained by composing the primitive operators in* (G.1)*, and let*

$$L_{q,t} = \sum_w q_w P_{w,t}^* P_{w,t}.$$

Then

$$\|L_{q,t} - L_{q,0}\|_{\mathrm{op}} \longrightarrow 0. \tag{G.2}$$

Consequently, a strict linear-recovery phase at $t = 0$ retains a unique top-d spectral subspace converging to $\mathcal{H}_1$. A strict nonlinear-shortcut phase retains a smooth, invertible, whitened coordinate system whose predictable energy is larger than that of the linear state coordinates.

In $d = 2$, for arbitrary signs $s_a \in \{-1, +1\}$, the explicit family

$$\Phi_{a,t}(z) = R_{s_a t \sin(\|z\|^2)} z \tag{G.3}$$

satisfies all assumptions. Each primitive channel in (G.1) is nonlinear for $t \neq 0$ and nevertheless preserves γ_2 exactly.

Proof. We first isolate where strict contraction is used. If $\sigma_1, \ldots, \sigma_d < 1$ are the singular values of A_a, Gaussian second quantization gives the squared Hilbert–Schmidt norm

$$\|K_a\|_{\mathrm{HS}}^2 = \sum_{\alpha \in \mathbb{N}^d} \prod_{j=1}^d \sigma_j^{2\alpha_j} = \prod_{j=1}^d (1 - \sigma_j^2)^{-1} < \infty.$$

Thus K_a is compact. The operators $U_{a,t}$ are unitary because the maps $\Phi_{a,t}$ preserve γ_d. Strong convergence to the identity is therefore uniform on the compact image of the unit ball under K_a, which yields

$$\|(U_{a,t} - I)K_a\|_{\mathrm{op}} \longrightarrow 0.$$

Equivalently, $\|P_{a,t} - K_a\|_{\mathrm{op}} \to 0$ for every primitive action.

For a word of fixed length, a telescoping expansion of the product of primitive operators now gives $\|P_{w,t} - P_{w,0}\|_{\mathrm{op}} \to 0$. All conditional-mean operators are contractions, so

$$\|P_{w,t}^* P_{w,t} - P_{w,0}^* P_{w,0}\|_{\mathrm{op}} \leq 2\|P_{w,t} - P_{w,0}\|_{\mathrm{op}}.$$

There are finitely many words; summing with weights q_w proves (G.2).

If $\Delta(q) > 0$ at $t = 0$, Proposition 15 applied with $E = L_{q,t} - L_{q,0}$ preserves the isolated top-d cluster and makes its spectral projection converge to that of $\mathcal{H}_1$. If $\Delta(q) < 0$, fix one smooth invertible improving coordinate from the negative-phase part of Theorem 1. Its advantage at $t = 0$ is strict. Operator-norm continuity makes the scores of this fixed coordinate system and of the linear coordinates continuous in t, so the strict advantage persists in a neighborhood of zero.

It remains to verify the explicit family. In polar coordinates, (G.3) maps (r, θ) to $(r, \theta + s_a t \sin(r^2))$. It preserves the radial coordinate and has unit Jacobian in (r, θ), so it preserves every radially symmetric measure, including γ_2. Its inverse is $\Phi_{a,-t}$, and it is a smooth global diffeomorphism. Pointwise convergence to the identity, first for bounded continuous functions and then by density and unitarity for all of $L^2(\gamma_2)$, gives the required strong convergence. Finally, $A_a \Phi_{a,t}(x)$ is nonlinear for $t \neq 0$ in the explicit invertible action matrices used in the paper. Since $\Phi_{a,t}(X) \sim \gamma_2$, adding the stated Gaussian noise gives $A_a \Phi_{a,t}(X) + \xi_a \sim \gamma_2$, proving exact stationarity. □

Frozen finite-sample confirmation. We instantiate (G.3) with opposite action signs and $t = 0.2$, while keeping the exact length-five laws and primitive matrices fixed. Each of seven fresh seeds uses 65,536 current states per word and two conditionally independent future replicas. The empirical $\mathcal{H}_1 \oplus \mathcal{H}_2$ blocks retain opposite diagonal margins in all seeds. The mean margin is 0.02235 for q_+ (95% CI $[0.02022, 0.02448]$) and -0.01172 for q_- (95% CI $[-0.01327, -0.01016]$); the corresponding full five-dimensional eigengaps are positive in every seed. Regressing the conditional mean on the current state leaves a mean residual fraction of 0.01984 (95% CI $[0.01974, 0.01994]$), whereas the mean output-covariance operator error is 0.00776. Radius preservation and inversion are

checked to floating-point tolerance, but exact Gaussian stationarity follows from the polar-coordinate argument above, not from these diagnostics.

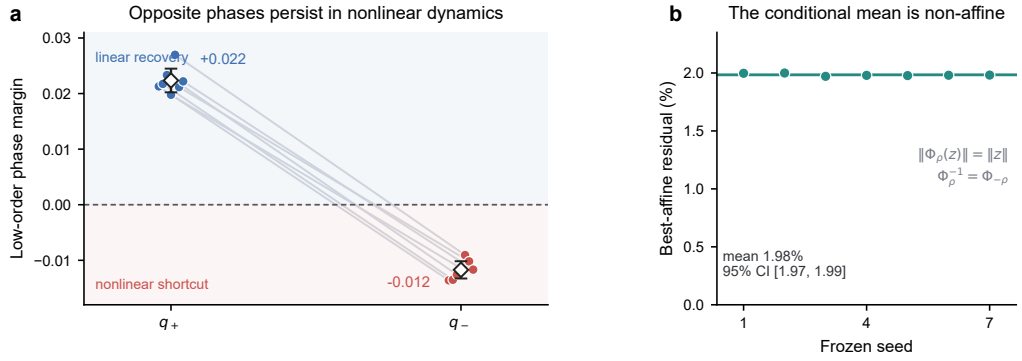

Figure 4: **Opposite action-word phases persist in an exact nonlinear Gaussian-stationary world. a,** Empirical low-order phase margins for the seven frozen seeds at $t = 0.2$. Gray lines join the two laws within a seed; diamonds and bars show means and seed-level Student-t 95% confidence intervals. **b,** Fraction of conditional-mean variance left unexplained by the best affine model for the same seeds. All observations are shown. Gaussian stationarity follows analytically from radius preservation and the explicit inverse, rather than from a finite-sample normality test.

H Exact Gaussianization does not guarantee identifiability

The negative-phase construction above assumes population whitening. A natural question raised by Gaussian-regularized JEPA objectives is whether the stronger constraint $h_{\#}\gamma_d = \gamma_d$ removes the nonlinear competitor. It does not in general. The following anisotropic-channel example is separate from the frozen length-five matched laws; it isolates the obstruction without an approximate regularizer or a learned encoder. It therefore proves that exact marginal Gaussianity is not a universal certificate, not that the same Gaussian-preserving flow improves the frozen negative-phase law.

Proposition 17 (A Gaussian-preserving nonlinear improvement)**.** *Let $X \sim \gamma_2$, and let L be the Gaussian second-quantization operator whose first-chaos eigenvalues are $\lambda_1 = 1/100$ and $\lambda_2 = 9/10$. For*

$$J(h) = \sum_{i=1}^{2} \langle h_i, Lh_i \rangle_{L^2(\gamma_2)},$$

there is a one-parameter group of smooth nonlinear diffeomorphisms $(\Phi_t)_{t\in\mathbb{R}}$ such that $(\Phi_t)_{\#}\gamma_2 = \gamma_2$, $\Phi_0 = I$, and

$$\frac{1}{2}\frac{\mathrm{d}^2}{\mathrm{d}t^2} J(\Phi_t)\Big|_{t=0} = \frac{709}{1000} > 0.$$

Consequently, $J(\Phi_t) > J(I)$ for every sufficiently small nonzero t. Thus exact standard-Gaussian output is not, by itself, sufficient to make the linear coordinates optimal in every anisotropic Gaussian world.

Proof. Consider the polynomial vector field

$$v(x_1, x_2) = (1 - x_2^2,\ x_1 x_2).$$

Its ordinary divergence and Gaussian radial term agree pointwise:

$$\operatorname{div} v = x_1 = x^\top v, \qquad \operatorname{div}_\gamma v := \operatorname{div} v - x^\top v = 0.$$

The field is complete. Indeed, for $x_2 \neq 0$,

$$H(x_1, x_2) = x_1^2 + x_2^2 - 2\log|x_2|$$

is constant along trajectories and is coercive on each open half-plane $\{x_2 > 0\}$ and $\{x_2 < 0\}$. Hence a trajectory that starts off the axis remains in a compact subset of its half-plane for every bounded time interval. On the invariant axis $x_2 = 0$, the solution is $x_1(t) = x_1(0) + t$. The flow Φ_t therefore exists for all $t \in \mathbb{R}$, is smooth, and has inverse Φ_{-t}. The stationary continuity equation $\nabla \cdot (v \, \mathrm{d}\gamma_2) = 0$ then gives $(\Phi_t)_{\#}\gamma_2 = \gamma_2$.

Let $A = v \cdot \nabla$ be the flow generator. The preceding divergence identity makes A skew-adjoint in $L^2(\gamma_2)$. Writing $u_i(t) = (\Phi_t)_i$, differentiation at zero gives $u_i'(0) = v_i$ and $u_i''(0) = A^2 x_i$. Since $Lx_i = \lambda_i x_i$, skew-adjointness yields

$$\frac{1}{2} J''(0) = \sum_{i=1}^{2} \left(\langle v_i, Lv_i \rangle - \lambda_i \|v_i\|_2^2 \right).$$

Now $v_1 = 1 - x_2^2$ is a second-chaos direction with $\|v_1\|_2^2 = 2$, while $v_2 = x_1 x_2$ has norm one. Second quantization therefore gives

$$\langle v_1, Lv_1 \rangle = 2\lambda_2^2, \qquad \langle v_2, Lv_2 \rangle = \lambda_1 \lambda_2.$$

Substitution produces

$$\frac{1}{2} J''(0) = 2(\lambda_2^2 - \lambda_1) + (\lambda_1 \lambda_2 - \lambda_2) = \frac{709}{1000}.$$

Also $J'(0) = 0$ by chaos orthogonality. Smoothness of the flow and of the displayed quadratic form finishes the local comparison. □

The same operator L has two relevant readings. It is the conditional mean operator for direct LeJEPA-style alignment when the OU channel has coordinate correlations (λ_1, λ_2); it is P^*P for profiled predictable energy when the channel correlations are $(\sqrt{\lambda_1}, \sqrt{\lambda_2})$. Because every Φ_t is invertible, the latter representation remains state-sufficient, so its profiled predictor may condition on the full current state. This spatial channel-anisotropy mechanism is separate from the temporal word-level dispersion that creates $\Delta(q_-) < 0$ for the matched witness; the vector field above is not claimed to be that witness's improving direction.

Figure 5 confirms both finite-sample consequences on disjoint frozen seed panels. The fitted quadratic coefficient is 0.70947 with 95% CI $[0.70481, 0.71414]$ for profiled predictable energy and 0.71045 with 95% CI $[0.70511, 0.71578]$ for direct alignment; both contain the exact value 0.709. These experiments evaluate the fixed analytic flow. They do not claim that an end-to-end learned encoder discovers it. For reproduction, the whitening-only control is

$$g_\eta(x_1, x_2) = \left(\sqrt{1 - \eta^2}\, x_1 + \eta \frac{x_2^2 - 1}{\sqrt{2}},\, x_2 \right), \qquad \eta = 0.35,$$

and the exact flow is integrated with fourth-order Runge–Kutta using maximum step size 0.002.

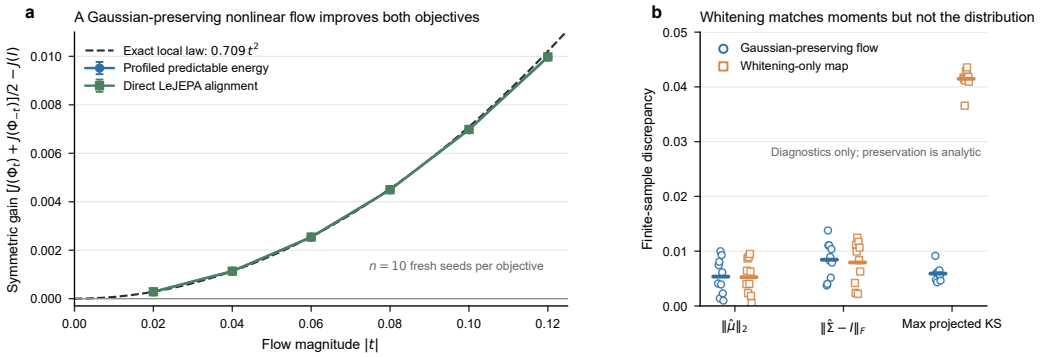

Figure 5: **Exact Gaussianization does not eliminate the nonlinear representation phase. a,** Symmetric objective gain along the Gaussian-preserving flow. Points show means over ten fresh seeds per objective and bars show seed-level Student-t 95% confidence intervals; each seed uses 100,000 samples. The dashed curve is the exact local prediction $0.709t^2$. The two objectives use disjoint seed groups. **b,** At $|t| = 0.12$, mean, covariance, and 24-direction projected-KS diagnostics for the exact flow and a nonlinear map that is only whitened. Each point is one profiled-energy seed; horizontal bars are means. Gaussian preservation of the flow follows analytically from $\mathrm{div}_\gamma\, v = 0$, not from these diagnostics.

Part II
Matched laws, statistical limits, and design

I MATCHED SUMMARIES, HORIZON SHARPNESS, AND OPEN FAMILIES

This section explains how two action-word laws can agree on every declared coverage summary and nevertheless occupy opposite representation phases. The linear part of the argument finds directions in the probability simplex that preserve the summaries. The spectral part chooses one such direction along which the largest quadratic eigenvalue crosses the fixed weakest linear eigenvalue. We give the abstract rank argument first, then the short length-four witness, and finally the exact length-five witness used in the experiments.

Matching T_1 is stronger than matching ordinary action frequencies, but it still constrains only the first Hermite block. The second block is a different linear image of the word law. The construction therefore has two logically separate tasks: find a probability direction that is invisible to every declared summary, and show that its image under T_2 crosses the relevant eigenvalue branch. The kernel identity below solves the first task; the exact witnesses and transverse-root argument solve the second.

For $w \in \{0,1\}^L$, let $\chi_S(w) = \prod_{i\in S}(-1)^{w_i}$ be a Walsh character. The linear map $C_{L,r}$ collects normalization, all Walsh moments with $|S| \le r$, and the independent coordinates of the complete T_1. We write $G_k q = \mathrm{svec}(T_k(q))$. Thus $\ker C_{L,r}$ contains exactly the signed changes of law that are invisible to every declared summary.

Theorem 18 (Neutral fibers and horizon sharpness)**.** *The information about T_k left after fixing the selected summaries is exactly*

$$\operatorname{rank}\begin{bmatrix} C_{L,r} \\ G_k \end{bmatrix} - \operatorname{rank}(C_{L,r}) = \operatorname{rank}\big(G_k|_{\ker C_{L,r}}\big)\,.$$

For the binary action construction and every $L \ge 4$, there are two full-support laws agreeing on all positional marginals through order $L-2$ and on the complete T_1, but differing in T_2. Within the fixed $d = 2$ binary construction and the degree-at-most-three positional-summary class, horizon four is the existential minimum for a phase separation, whereas horizon five is the minimum admitting a nonempty ambient-open action family. The $L = 4$ case is supplied by a separate commuting-shear witness; the all-horizon lift for $L \ge 5$ starts from the noncommuting length-five witness.

I.1 NEUTRAL DIRECTIONS ARE A KERNEL COMPUTATION

We first prove that this null space is not merely a dimension count: its image under G_k is exactly the higher-order operator information that remains free after the summaries have been fixed.

Proof of the rank identity in Theorem 18. Put $K = \ker C_{L,r}$. Two laws in the same summary fiber differ by an element of K, so the remaining variation in T_k is exactly the image of $G_k|_K$. Choose a complement $\mathbb{R}^{2^L} = K \oplus K'$ on which $C_{L,r}$ is injective. In this basis, the columns over K' contribute $\operatorname{rank}(C_{L,r})$ to the stacked map and the columns over K contribute $\operatorname{rank}(G_k|_K)$. Hence

$$\operatorname{rank}\begin{bmatrix} C_{L,r} \\ G_k \end{bmatrix} - \operatorname{rank}(C_{L,r}) = \operatorname{rank}(G_k|_K). \tag{A.29}$$

The summaries determine T_k exactly if and only if $K \subseteq \ker G_k$. Because the constant Walsh row belongs to $C_{L,r}$, every $\delta \in K$ has zero total mass. If q_0 is an interior law, then

$$0 < t < \min_{w:\delta_w \neq 0} \frac{q_{0,w}}{|\delta_w|}$$

makes $q_0 \pm t\delta$ distinct full-support laws with identical summaries. □

I.2 Lifting a separation to every longer horizon

Lemma 19 (Walsh lift)**.** *Suppose a signed length-L law δ_L annihilates all Boolean moments through degree r, satisfies $T_1(\delta_L) = 0$, and has $D_L = T_2(\delta_L) \neq 0$. Define $\delta_{L+1}(w, b) = b\delta_L(w)$ for $b \in \{-1, +1\}$. Then it annihilates every moment through degree $r + 1$, keeps T_1 fixed, and*

$$T_2(\delta_{L+1}) = B_+ D_L B_+^\top - B_- D_L B_-^\top, \qquad B_\pm = \mathrm{Sym}^2(A_\pm). \tag{A.30}$$

Proof. If a monomial does not contain the new bit, summing over b cancels its two copies. If it contains the new bit, multiplication by b removes that bit and leaves a monomial of degree at most r, which is annihilated by δ_L. Since $M_{(w,b)} = A_b M_w$, linearity gives

$$T_1(\delta_{L+1}) = A_+ T_1(\delta_L) A_+^\top - A_- T_1(\delta_L) A_-^\top = 0,$$

and functoriality of the symmetric square gives (A.30). For the certified noncommuting primitive actions of the length-five witness below, the linear contrast map

$$\mathcal{R}(D) = B_+ D B_+^\top - B_- D B_-^\top$$

on the six-dimensional symmetric-square operator space has strictly positive rational determinant (approximately 9.92234×10^{-5}). It therefore never maps a nonzero D_L to zero. Starting from the signed contrast of the length-five pair and iterating proves the statement for every $L \geq 5$, with moment matching through degree $L - 2$. The independent commuting-shear construction in the next subsection supplies $L = 4$. Together, the two base cases prove the declared existence statement for every $L \geq 4$. □

I.3 Why horizon four is the existential minimum

For the fixed degree-at-most-three summary class, $L \leq 3$ leaves no nonzero law contrast because those Walsh characters span every function on the word cube. The length-four commuting-shear witness supplies the existential separation. Explicitly, let

$$A_x = \frac{7}{10} I + \frac{x}{2} \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \qquad q_\theta(w) = \frac{1 + \theta \prod_{i=1}^4 w_i}{16}, \quad x, w_i \in \{-1, +1\}.$$

The two actions commute and $I - A_x A_x^\top \succ 0$, with Sylvester minors $13/50$ and $101/10000$. For $\theta_\pm = \mp 999/1000$, both laws have minimum atom $1/16000$; Walsh orthogonality makes all degree-at-most-three positional moments equal. Direct rational calculation also gives the common complete operator

$$T_1 = \mathrm{diag}\left(\frac{17529701}{10^8}, \frac{5764801}{10^8} \right).$$

For $\theta = -999/1000$, rational Sylvester minors certify $\lambda_{\min}(T_1) - \lambda_{\max}(T_2) > 1/50$. For $\theta = 999/1000$, the first quadratic coordinate has Rayleigh advantage greater than $1/100$. Thus the two laws have opposite phases while matching all declared summaries. This witness is existentially sharp but nongeneric: its action tuple lies in a proper algebraic subset. The degree-three summary matrix has rank 15, while an explicit nonsymmetric tuple makes the summary-plus-T_1 matrix rank 16. Thus no ambient-open length-four family can satisfy the required incidence constraint.

I.4 The exact length-five witness

At $L = 5, d = 2, r = 3$, the constraint matrix has 32 columns and

$$1 + \binom{5}{1} + \binom{5}{2} + \binom{5}{3} + 3 = 29$$

rows. Index words lexicographically and multiply actions in temporal order, $M_w = A_{w_5} \cdots A_{w_1}$. The primitive actions are

$$A_0 = \begin{pmatrix} -7/400 & -3917/5000 \\ 1753/2500 & -57/5000 \end{pmatrix}, \quad A_1 = \begin{pmatrix} 2643/5000 & 397/2000 \\ 173/625 & 4299/5000 \end{pmatrix}. \tag{A.31}$$

The Sylvester minors of $I - A_i A_i^\top$ are

$$\left(\frac{38597819}{10^8}, \frac{490346397008441}{2500000000000000}\right), \quad \left(\frac{68117979}{10^8}, \frac{15588812867769}{625000000000000}\right),$$

so both actions are strict contractions.

Let $C \in \mathbb{Q}^{29\times 32}$ be the constraint matrix just defined. To make the rational construction unambiguous, order its rows as follows: the constant row; Walsh subsets in increasing cardinality and, within each cardinality, lexicographic subset order; and finally $\mathrm{svec}(T_1) = (T_{1,11}, T_{1,12}, T_{1,22})$. Exact elimination gives rank 29, pivot columns $1, \ldots, 29$, and free columns $30, 31, 32$. Partition $C = [C_p\ C_f]$, where $C_p = C_{:,1:29}$ and $C_f = C_{:,30:32}$. The selected rational minor is nonzero, hence C_p is invertible, and the unique RREF kernel basis is explicitly

$$N = \begin{bmatrix} -C_p^{-1} C_f \\ I_3 \end{bmatrix}. \tag{A.31a}$$

Equivalently, it is the unique $N \in \mathbb{Q}^{32\times 3}$ satisfying

$$CN = 0, \qquad N_{\{30,31,32\},:} = I_3. \tag{A.32}$$

The positive law is $q_+ = 10^{-12} a$, where

$$\begin{aligned} a = (&653750000, 50451527881, 653750000, 5290073496,\\ &653750000, 5290073496, 55087851377, 653750000,\\ &653750000, 54265434611, 104063212491, 653750000,\\ &104063212491, 653750000, 653750000, 163133637365,\\ &112011103911, 653750000, 653750000, 52940679037,\\ &653750000, 52940679037, 3142901157, 2800922340,\\ &653750000, 3965317923, 653750000, 51776283454,\\ &653750000, 51776283454, 101574061335, 16274245144). \end{aligned} \tag{A.33}$$

Set

$$\zeta = \left(\frac{25561266727}{500000000000}, \frac{20184062267}{200000000000}, -\frac{105556634831}{1000000000000}\right)^\top, \qquad q_- = q_+ - N\zeta. \tag{A.34}$$

Equations (A.31)–(A.34) specify the witness using rational numbers only. The constant row, Walsh rows, and final three rows of $CN = 0$ prove that the laws have unit mass, match every degree-at-most-three moment, and have identical complete T_1, respectively.

Exact coordinate comparison gives

$$\min_w q_{+,w} = \frac{523}{800000} > \frac{1}{2000}, \qquad \min_w q_{-,w} > \frac{1}{2000}. \tag{A.35}$$

All word channels are strict contractions because

$$\max_w \|M_w\|_F^2 = \frac{32900809100733328095055231911343258769}{39062500000000000000000000000000000000} < 1.$$

For the positive endpoint, rational Sylvester tests give

$$\lambda_{\min}(T_1) > \frac{1501}{20000}, \qquad \lambda_{\max}(T_2(q_+)) < \frac{1101}{20000},$$

and hence $\Delta(q_+) > 1/50$. For the negative endpoint, a rational upper bound on the common weakest linear eigenvalue and a rational quadratic Rayleigh certificate give

$$\lambda_{\min}(T_1) < \frac{7513}{100000}, \qquad \lambda_{\max}(T_2(q_-)) > \frac{381}{4000},$$

so

$$\lambda_{\max}(T_2(q_-)) - \lambda_{\min}(T_1) > \frac{503}{25000} > \frac{1}{50}. \tag{A.36}$$

Table 3: Exact checks supporting the length-five witness and its open neighborhood.

Object	Exact argument	Domain	Strict conclusion
Neutral fiber	rational elimination and selected minor	fixed actions	$\operatorname{rank} C = 29$, $\dim \ker C = 3$
Full support	coordinatewise rational comparison	q_+, q_-	$\min_w q_{\pm,w} > 1/2000$
Positive phase	Sylvester inertia bounds	q_+	$\lambda_{\min}(T_1) > 1501/20000$, $\lambda_{\max}(T_2) < 1101/20000$
Negative phase	rational Rayleigh and inertia bounds	q_-	$\lambda_{\max}(T_2) - \lambda_{\min}(T_1) > 503/25000$
Transverse root	resultant, Sturm sequence, and inertia	$t \in [0.47, 0.54]$	unique $t_0 \in (0.5024, 0.5025)$, $\Delta'(t) < -3/100$
Ambient-open family	Taylor, resolvent, Weyl, and support bounds	block-Frobenius radius 10^{-5}	all six lower bounds in (A.41) hold on the full closed ball

Numerically the two margins are $0.0203409658\ldots$ and $-0.0201450577\ldots$, but their signs and buffers follow from the displayed rational inequalities.

Table 3 summarizes the six mathematical checks that connect the rational witness to the open-family conclusion. The underlying rational objects and replay instructions are listed separately in Appendix M; the proof here records the argument and the strict inequalities it uses.

Along $q(t) = (1-t)q_+ + tq_-$, T_1 is constant and $T_2(t)$ is affine. Exact root isolation on $[0.47, 0.54]$ shows that the top eigenvalue remains simple, $\Delta'(t) < -3/100$, and there is a unique root

$$t_0 \in (0.5024, 0.5025). \tag{A.37}$$

For its normalized top eigenvector $v(t)$, the standard simple-eigenvalue formula gives $\Delta'(t) = -v(t)^\top (T_2(q_-) - T_2(q_+))v(t)$. The derivative bound therefore proves a transverse crossing, not a tangency.

Here is the exact algebra behind the isolation step. Let $p_1(s) = \det(T_1 - sI)$ and $p_2(t,s) = \det(T_2(t) - sI)$. A phase crossing must be a common zero of these two rational polynomials. Eliminating s gives the univariate resultant $R(t) = \operatorname{Res}_s(p_1(s), p_2(t,s))$. A Sturm sequence for R has exactly one sign-count drop between 0.5024 and 0.5025, and none on the remainder of $[0.47, 0.54]$. Rational inertia tests at the interval endpoints identify the common root with the weakest eigenvalue of T_1 and the largest eigenvalue of $T_2(t)$, rather than with another eigenvalue branch. The same signed-minor calculation keeps the top eigenvalue simple on the entire interval. Thus the polynomial certificate proves precisely the spectral crossing used by the theorem.

I.5 An ambient-open family of witnesses

Fixed laws do not preserve T_1-neutrality after perturbing the actions, so the laws must move with the action tuple. Let $\mathcal{W} \in \mathbb{R}^{32\times 6}$ contain the five degree-four Walsh characters and the degree-five character, normalized by $1/\sqrt{32}$. These columns annihilate all degree-at-most-three moments. Let $G(A)$ map a law to the three independent entries of T_1, and put

$$R(A) = G(A)\mathcal{W} \in \mathbb{R}^{3\times 6}. \tag{A.38}$$

At (A.31), columns $1, 2, 6$ form an invertible pivot. Freeze the other three Walsh coefficients and solve the pivot system $R(A)z(A) = 0$. This defines a continuous neutral direction $d(A) = \mathcal{W}z(A)$ anchored at $q_- - q_+$. With u uniform and $q_{\mathrm{mid}} = (q_+ + q_-)/2$, define

$$q_A(t) = 0.9\{q_{\mathrm{mid}} + (t - \tfrac{1}{2})d(A)\} + 0.1u, \qquad t \in [\tfrac{1}{4}, \tfrac{3}{4}]. \tag{A.39}$$

The uniform buffer stabilizes support; the moving direction keeps both low-order moments and T_1 equal between the two endpoints.

On the complete closed block-Frobenius ball

$$\|A_0 - \bar{A}_0\|_F^2 + \|A_1 - \bar{A}_1\|_F^2 \le 10^{-10}, \tag{A.40}$$

product-rule bounds for all 32 word matrices and an explicit Taylor remainder for the inverse pivot block give uniformly

$$\begin{array}{rcl@{\qquad}rcl} \sigma_{\min}(R_{\mathrm{pivot}}(A)) & > & 0.00890, & \min_{w,t} q_A(t)_w & > & 0.00357, \\ \Delta(q_A(1/4)) & > & 0.00972, & -\Delta(q_A(3/4)) & > & 0.00572, \\ \text{top eigengap at the crossing} & > & 0.0427, & |\Delta'(t_0(A))| & > & 0.0194. \end{array} \tag{A.41}$$

For clarity, we spell out how these uniform estimates are assembled. Write $E_i = A_i - \bar{A}_i$ and $\rho = \max_i \|\bar{A}_i\|_{\mathrm{op}}$. Repeatedly adding and subtracting one perturbed factor gives the telescoping identity

$$M_w(A) - M_w(\bar{A}) = \sum_{j=1}^{5} A_{w_5} \cdots A_{w_{j+1}} E_{w_j} \bar{A}_{w_{j-1}} \cdots \bar{A}_{w_1}. \tag{A.42}$$

Consequently, on a radius-r block ball,

$$\|M_w(A) - M_w(\bar{A})\|_F \le \sum_{j=1}^{5} (\rho + r)^{5-j} \rho^{j-1} \max_i \|E_i\|_F. \tag{A.43}$$

The covariance perturbation follows from

$$MM^\top - \bar{M}\bar{M}^\top = (M - \bar{M})M^\top + \bar{M}(M - \bar{M})^\top. \tag{A.44}$$

Summing these wordwise bounds controls both $G(A) - G(\bar{A})$ and the first- and second-chaos operators.

Let $P(A)$ be the 3×3 pivot block of $R(A)$. Weyl's singular- value inequality gives

$$\sigma_{\min}(P(A)) \ge \sigma_{\min}(P(\bar{A})) - \|P(A) - P(\bar{A})\|_{\mathrm{op}} > 0.00890. \tag{A.45}$$

Let $z_0 = \mathcal{W}^\top(q_- - q_+)$ be the base null vector, and let $z_f = (z_{0,3}, z_{0,4}, z_{0,5})^\top$ collect its frozen nonpivot coordinates. If $F(A)$ is the remaining block, the moving pivot coordinates are exactly $z_p(A) = -P(A)^{-1}F(A)z_f$. The resolvent identity

$$P(A)^{-1} - P(\bar{A})^{-1} = -P(A)^{-1}\{P(A) - P(\bar{A})\}P(\bar{A})^{-1} \tag{A.46}$$

then bounds $z(A) - z(\bar{A})$, hence every coordinate of the moving law. This yields the support floor in (A.41). Finally, Weyl's eigenvalue inequality applied to the operator perturbations obtained from (A.42)–(A.44) preserves the two endpoint signs and the top eigengap. Differentiating the affine T_2 path and using the simple-eigenvalue formula controls the crossing derivative. Substituting $r = 10^{-5}$ into these expressions gives the six numerical lower bounds in (A.41), with outward rounding.

The same estimates keep primitive actions and all word channels strictly contractive. These are uniform bounds on the whole ball, not random or corner checks. The implicit function theorem and the final two inequalities yield a unique simple crossing throughout its interior. This proves the nonempty ambient-open family. The common summary value may move with A; equality of the two endpoint summaries is what is preserved.

J Local testing lower bound

The phase is statistically regular near the transverse crossing: changing the margin by γ requires moving the categorical word law by order γ, so distinguishing the two signs requires order γ^{-2} samples. The lower bound allows the learner to observe the entire replicated Gaussian rollout; it is not caused by hiding states or futures.

Corollary 20 (Local phase-testing hardness)**.** *For each $\gamma \in (0, 10^{-3}]$, the matched segment contains full-support laws q_γ^I, q_γ^N with margins $+\gamma$ and $-\gamma$. When the channel and primitive actions are known but the active word law is one of these two laws, any classifier based on $n \le 9/(320000\gamma^2)$ independent observations of (W, X, Y_1, Y_2) has worst-case error at least $1/4$.*

Proof. **Step 1: locate the two phase margins.** Along $q(t) = q_+ + t(q_- - q_+)$, write $v = q_- - q_+$. On $t \in [0.47, 0.54]$, the exact certificate gives one simple root $t_0 \in (0.5024, 0.5025)$,

$$\Delta'(t) < -\frac{3}{100}, \qquad \Delta(0.47) > \frac{1}{1000}, \qquad \Delta(0.54) < -\frac{1}{1000},$$

$$\sup_t \sum_w \frac{v_w^2}{q_w(t)} < 4. \tag{A.47}$$

For each $0 < \gamma \le 10^{-3}$, the intermediate value theorem and (A.47) give $t_I < t_0 < t_N$ with phase margins $+\gamma, -\gamma$ and

$$|t_N - t_I| \le \frac{200}{3}\gamma. \tag{A.48}$$

The separation bound is the mean-value theorem. On each side of the root, a margin change of γ needs at most $100\gamma/3$ displacement in t, so the two laws are at most $200\gamma/3$ apart.

Step 2: reduce the complete observation to the word law. The word-conditioned Gaussian kernels are identical under the two hypotheses, even when the classifier observes the complete tuple (W, X, Y_1, Y_2). The joint one-sample KL divergence is therefore the categorical word-law KL.

Step 3: integrate categorical Fisher information. The pathwise Fisher information is $I(t) = \sum_w v_w^2/q_w(t) < 4$ by (A.47). Integrating this bound along the segment yields

$$\mathrm{KL}(q(t_I)\|q(t_N)) \le 4|t_N - t_I|^2 \le \frac{160000}{9}\gamma^2. \tag{A.49}$$

For n independent tuples the KL is multiplied by n.

Step 4: convert information to error. If $n \le 9/(320000\gamma^2)$, (A.49) is at most $1/2$. Pinsker's inequality gives total variation at most $1/2$, and Le Cam's two-point inequality gives worst-case testing error at least $(1 - \mathrm{TV})/2 \ge 1/4$. This is an unknown-active-law result; it does not apply when the generating law is supplied to the classifier. □

K Action-word design as a semidefinite program

The next proposition separates an exact mathematical reformulation from the quality of a particular numerical solver output.

Proposition 21 (Exact SDP representation of the phase margin)**.** *Let $\mathcal{Q}$ be a nonempty compact set of word laws described by affine constraints and lower support bounds. Then*

$$\max_{q\in\mathcal{Q}} \Delta(q) = \max_{q,\alpha,\beta} \{\alpha - \beta : T_1(q) \succeq \alpha I_d,\ T_2(q) \preceq \beta I_{p_2},\ q \in \mathcal{Q}\}. \tag{A.50}$$

Proof. Both $T_1(q) = \sum_w q_w D_w$ and $T_2(q) = \sum_w q_w D_w^{\odot 2}$ are affine in q. Therefore all constraints on the right side of (A.50) are affine or semidefinite.

Fix q. The first matrix inequality is equivalent to $\alpha \le \lambda_{\min}(T_1(q))$, while the second is equivalent to $\beta \ge \lambda_{\max}(T_2(q))$. Every feasible triple hence obeys

$$\alpha - \beta \le \lambda_{\min}(T_1(q)) - \lambda_{\max}(T_2(q)) = \Delta(q).$$

Equality is attained for this same q by choosing the two extremal eigenvalues as α and β. Maximizing over q proves (A.50), and compactness yields attainment. The KKT residuals reported for our numerical program diagnose solver accuracy; they are not used in this equivalence and do not claim that the exact witness q_+ is itself the matched-fiber optimizer. □

L Conditional estimation and data-driven word-law design

The estimator mirrors the population proof: one head estimates predictable linear energy and a second estimates predictable quadratic energy. Replicas make both heads unbiased cross moments. Since these Gaussian-polynomial matrices are unbounded, we use a robust spectral mean estimator rather than a bounded matrix Bernstein inequality.

Corollary 22 (Exact-state Replica–Hermite certificate)**.** *Assume access to literal whitened state coordinates or an independently certified exact representation $Z = QX$, $Q \in O(d)$. Given n i.i.d.*

replicated tuples and $\delta \in (0,1)$, for conditionally independent future replicas Y_1, Y_2, set $Z_r = QY_r$, $r \in \{1,2\}$, and define

$$\Phi_1(z) = z, \qquad \Phi_2(z) = \operatorname{svec}\left(\frac{zz^\top - I}{\sqrt{2}}\right).$$

The symmetrized cross outer products have expectations T_1, T_2 in state coordinates and orthogonal conjugates under $Z = QX$. Let

$$\widehat{\Delta} := \lambda_{\min}(\widehat{T}_1) - \lambda_{\max}(\widehat{T}_2).$$

Minsker's preliminary spectral Catoni estimates $\widehat{T}_1, \widehat{T}_2$ satisfy, with probability at least $1-\delta$,

$$|\widehat{\Delta} - \Delta| \le b_1(n,\delta) + b_2(n,\delta), \qquad b_k(n,\delta) = \sqrt{\frac{2v_k\ell_k}{n}},$$

where $\ell_k = \log(4p_k/\delta)$ and

$$p_1 = d, \quad p_2 = \frac{d(d+1)}{2}, \quad v_1 = d+2, \qquad v_2 = \frac{\sqrt{15}}{2}\sqrt{d^4 + 10d^3 + 25d^2 + 24d}.$$

Write

$$S = \sqrt{v_1\ell_1} + \sqrt{v_2\ell_2}.$$

Hence, for fixed d, δ, a nonzero phase margin is certified with $n = O(|\Delta|^{-2})$.

Proof. **Step 1: construct unbiased linear and quadratic heads.** For $k = 1, 2$, let Φ_k be as in Corollary 22 and define the self-adjoint sample operator

$$\mathcal{Z}_k = \frac{1}{2}\{\Phi_k(Z_1)\Phi_k(Z_2)^\top + \Phi_k(Z_2)\Phi_k(Z_1)^\top\}. \tag{A.51}$$

In literal state coordinates,

$$\mathbb{E}[\mathcal{Z}_1 \mid W = w] = M_w M_w^\top.$$

For the second chaos,

$$\mathbb{E}[\Phi_2(Y) \mid X, W = w] = \rho_2(M_w)\Phi_2(X), \qquad \rho_2(M) : A \mapsto MAM^\top.$$

Replica independence and $\operatorname{Cov}(\Phi_2(X)) = I$ then give

$$\mathbb{E}[\mathcal{Z}_2 \mid W = w] = \rho_2(M_w)\rho_2(M_w)^* = (M_w M_w^\top)^{\odot 2}. \tag{A.52}$$

Under $Z = QX$, these expectations are orthogonal conjugates, so their eigenvalues and Δ are unchanged.

Step 2: bound the matrix second moments. The operators in (A.51) are unbounded Gaussian polynomials, so a bounded matrix Bernstein inequality is not applicable. For the moment bounds, return to literal state coordinates. For $k = 1$, condition first on $W = w$ and write $C_w = \mathbb{E}[Z_1 Z_2^\top \mid W = w] = M_w M_w^\top$, where $0 \preceq C_w \preceq I$. Isserlis' formula gives

$$\mathbb{E}[\mathcal{Z}_1^2 \mid W = w] = \frac{d+1}{2} I + \frac{3}{2} C_w^2 + \frac{\operatorname{tr}(C_w)}{2} C_w \preceq (d+2) I.$$

Average this conditional inequality over w.

For $a = \Phi_2(Z_1)$, $b = \Phi_2(Z_2)$, the inequality $\|a\langle b,u\rangle + b\langle a,u\rangle\|^2 \le 2\|a\|^2\langle b,u\rangle^2 + 2\|b\|^2\langle a,u\rangle^2$ and Cauchy–Schwarz bound the raw second moment of their symmetrized outer product using

$$\mathbb{E}\|\Phi_2(G)\|^4 = \frac{d^4 + 10d^3 + 25d^2 + 24d}{4}, \qquad \sup_{\|u\|=1} \mathbb{E}\langle u, \Phi_2(G)\rangle^4 \le 15,$$

which gives the displayed value of v_2. For completeness, if $R = \|G\|^2 \sim \chi^2_d$, then

$$\|\Phi_2(G)\|^2 = \frac{1}{2}\operatorname{tr}\{(GG^\top - I)^2\} = \frac{1}{2}(R^2 - 2R + d).$$

Using the first four moments of χ^2_d and expanding the square gives the polynomial displayed above. For a unit symmetric-square direction, the scalar projection is a variance-one element of the second Gaussian chaos; diagonalization and Isserlis' formula give the sharp uniform fourth-moment bound 15. Finally, for any unit u,

$$\|\mathcal{Z}_2 u\|^2 \le \frac{1}{2}\|a\|^2\langle b,u\rangle^2 + \frac{1}{2}\|b\|^2\langle a,u\rangle^2.$$

Cauchy–Schwarz and symmetry therefore yield

$$\|\mathbb{E}\mathcal{Z}_2^2\|_{\mathrm{op}} \le \frac{\sqrt{15}}{2}\sqrt{d^4+10d^3+25d^2+24d} = v_2.$$

Step 3: apply the robust spectral mean bound. Use Minsker's preliminary spectral Catoni estimator (Minsker, 2018, Eq. (3.5), Theorem 3.1 and Remark 2). It applies to i.i.d. self-adjoint matrices and any deterministic $v_k \ge \|\mathbb{E}\mathcal{Z}_k^2\|_{\mathrm{op}}$. Its spectral function ψ satisfies

$$-\log(1-x+x^2/2) \le \psi(x) \le \log(1+x+x^2/2).$$

For p_k, v_k, ℓ_k stated in Corollary 22, take

$$\theta_k = \sqrt{\frac{2\ell_k}{nv_k}}, \qquad \widehat{T}_k = \frac{1}{n\theta_k}\sum_{i=1}^n \psi(\theta_k \mathcal{Z}_{k,i}).$$

In the notation of Minsker's Remark 2, set $s_k = 2\ell_k$ and $\sigma_k^2 = v_k$. The failure probability for head k is then at most $2p_k\exp(-s_k/2) = \delta/2$. A union bound therefore implies, with probability at least $1-\delta$,

$$\|\widehat{T}_k - T_k\|_{\mathrm{op}} \le \sqrt{\frac{2v_k\ell_k}{n}}, \qquad k=1,2. \tag{A.53}$$

Step 4: convert operator error to a phase certificate. Weyl's inequality applied to $\widehat{\Delta} = \lambda_{\min}(\widehat{T}_1) - \lambda_{\max}(\widehat{T}_2)$ proves the displayed confidence radius. If $|\Delta| \ge \gamma$ and $n > 8S^2/\gamma^2$, that radius is strictly smaller than $\gamma/2$, so the confidence interval excludes zero with the correct sign. This proof uses literal state or independently certified exact orthogonal state coordinates; ordinary L^2 closeness of a learned encoder is insufficient to control the quadratic feature Φ_2. □

The same estimates can be used to choose a behavior law rather than only to certify a fixed one. The next result formalizes the resulting plug-in rule.

Corollary 23 (Finite-sample plug-in word-law design)**.** *Let $\mathcal{W}$ contain m words and let $\mathcal{Q}$ be any nonempty compact subset of the probability simplex. For every word w, suppose self-adjoint estimates $\widehat{D}_{1,w}$ and $\widehat{D}_{2,w}$ obey*

$$\max_w \|\widehat{D}_{1,w} - D_w\|_{\mathrm{op}} \le e_1, \qquad \max_w \|\widehat{D}_{2,w} - D_w^{\odot 2}\|_{\mathrm{op}} \le e_2. \tag{A.54}$$

Define

$$\widehat{\Delta}(q) = \lambda_{\min}\left(\sum_w q_w \widehat{D}_{1,w}\right) - \lambda_{\max}\left(\sum_w q_w \widehat{D}_{2,w}\right).$$

If $\widehat{q} \in \mathcal{Q}$ is ξ-optimal for this empirical margin, then

$$\Delta(\widehat{q}) \ge \max_{q\in\mathcal{Q}} \Delta(q) - 2(e_1+e_2) - \xi. \tag{A.55}$$

In particular, take n independent replicated pilot rollouts for each word and estimate both blocks using Corollary 22. With probability at least $1-\delta$, (A.54) holds with

$$e_k = \sqrt{\frac{2v_k\ell_{k,m}}{n}}, \qquad \ell_{k,m} = \log\left(\frac{4p_k m}{\delta}\right), \qquad k\in\{1,2\}. \tag{A.56}$$

Consequently, for fixed dimension and word set, the true design regret is $O(n^{-1/2}) + \xi$. More explicitly, if $0 \le \xi < \rho$, then

$$n \ge \frac{8\left(\sqrt{v_1 \ell_{1,m}} + \sqrt{v_2 \ell_{2,m}}\right)^2}{(\rho - \xi)^2}$$

is sufficient for true margin regret at most ρ. Here n is the number of replica pairs per word, so the total pilot budget is

$$N_{\mathrm{total}} = mn.$$

For $m = |\mathcal{A}|^L$ and fixed feature dimension, the displayed guarantee therefore scales as

$$N_{\mathrm{total}} = O\left(|\mathcal{A}|^L (\rho - \xi)^{-2} \log \frac{|\mathcal{A}|^L}{\delta}\right).$$

It is intended for moderate finite macro horizons, not as a horizon-free sampling guarantee.

Proof. For any q in the probability simplex, convexity of the operator norm and (A.54) give

$$\left\| \sum_w q_w (\widehat{D}_{1,w} - D_w) \right\|_{\mathrm{op}} \le e_1, \qquad \left\| \sum_w q_w (\widehat{D}_{2,w} - D_w^{\odot 2}) \right\|_{\mathrm{op}} \le e_2.$$

Weyl's inequality therefore yields the uniform deviation bound

$$\sup_{q \in \mathcal{Q}} |\widehat{\Delta}(q) - \Delta(q)| \le e_1 + e_2 =: e. \tag{A.57}$$

Let $q^\star$ maximize the true margin over $\mathcal{Q}$. Empirical ξ-optimality and (A.57) imply

$$\Delta(\widehat{q}) \ge \widehat{\Delta}(\widehat{q}) - e \ge \widehat{\Delta}(q^\star) - \xi - e \ge \Delta(q^\star) - 2e - \xi,$$

which proves (A.55). Apply Corollary 22 separately to each word with failure probability δ/m, then take a union bound, to obtain (A.56). Substitution into (A.55) gives the displayed sample requirement. □

Part III
Reproducibility and machine-checked scope

M Reproducibility and compute details

The exact-length-five spectral learner uses one orthogonal observation mix fixed before evaluation and the 20 seeds

$$117, 143, 171, 197, 223, 251, 277, 307, 337, 367,$$
$$397, 431, 463, 499, 541, 577, 613, 653, 691, 733.$$

It uses nested sample sizes $2^9, \dots, 2^{16}$, common random numbers across five laws, and 32,768 held-out probe states per seed. The full experiment ran in 11.614 seconds on one RTX 4060 Ti. For the four SDP programs, the supplement records primal feasibility, dual feasibility, complementarity, stationarity, and native SCS residuals. The worst stationarity residual is 1.96×10^{-8}; for the matched-fiber H_2 program it is 3.99×10^{-10}, with native primal–dual gap 1.12×10^{-10}. These are numerical KKT diagnostics, not an exact rational dual certificate. The exact-state Replica–Hermite validation uses $d = 2$, $\delta = 0.1$, $\gamma = 0.2$, $n = 52{,}622$, and 50 repetitions per phase.

The accompanying anonymous supplement includes the rational action/law data, exact verification scripts, the Lean project, fixed protocols, raw JSON results, and replay commands.

Table 4: Protocol for the ten-seed pixel experiment.

Component	Prespecified setting
Observation	Deterministic invertible-shear 64×64 RGB renderer.
Encoder	Four-layer stride-2 CNN; 689,586 parameters; two-dimensional output; no predictor head.
Optimization	AdamW, learning rate 3×10^{-4}, 20,000 updates, 512 replica pairs and 1,024 whitening states per update; final iterate only.
Candidate bank	Four independently initialized restarts per law and outer seed, with paired base randomness across the anonymous laws.
Selection	Largest mean held-out F_q over 32 independent blocks; no state, Hermite, phase, or semantic law label is available before selection.
Confirmation	Ten fresh paired seeds, no exclusions; independent calibration, test-A, and test-B panels evaluated after checkpoint selection.
Controls	Restart zero, a prespecified random restart, and a post-hoc oracle used only as a diagnostic ceiling.
Development boundary	The four-restart rule was chosen using earlier pilot runs and held fixed for the ten confirmation seeds.
Compute	Three RTX 2080 Ti GPUs; orchestrated wall time 10 hours 30 minutes.

The ten outer seeds are

$$2199143922, 731265263, 1835842072, 1863697692, 1591028803,$$
$$2420910389, 193429475, 2558417431, 2141736174, 1403969431.$$

Each selection block uses 32,768 independent q-free whitening states and 4,096 replica pairs for every one of the 32 words. After selection is frozen, three disjoint Monte Carlo panels of 2^{20} states each fit the population normalizer and evaluate state alignment and Hermite fractions. The supplement contains the training, selection, and evaluation records for the full 80-checkpoint bank and the source needed to regenerate it. Checkpoint banks are excluded from the submission archive; the selected models can be reconstructed from the frozen records and source. All prespecified validity and scientific criteria are satisfied, and the positive- and negative-phase criteria each hold in 10/10 seeds. Figure 2 is generated from the complete ten-seed cohort rather than a selected subset.

Table 5: Protocol for joint predictor learning after replica-based discovery.

Component	Prespecified setting
Source representation	Objective-selected pixel checkpoint for each law and outer seed; no state or Hermite label is used for source selection.
Predictors	(i) a 38,976-parameter ordered-action-word expert bank; and (ii) the released LeWorldModel `Embedder` and AdaLN-zero `ARPredictor`, totaling 127,570 action-encoder and predictor parameters. Both are randomly initialized for the joint stage.
Joint objective	Predictor minimizes single-future JEPA risk; the encoder minimizes negative replicated-future energy plus 0.1 times that risk.
Optimization	AdamW; 1,000 updates; encoder and predictor learning rates 3×10^{-4} and 5×10^{-4}.
Selection	Best held-out combined predictive objective; representation labels are revealed only after selection.
Development and confirmation	Three disclosed development seeds followed by seven held-out pixel-confirmation pairs, with no exclusions.
Required controls	Zero action-word condition, strictly positive selected update, and nonzero encoder and predictor parameter changes.

For the LeWorldModel family, five binary primitive actions are converted to one-hot action tokens; the current two-dimensional representation is projected to width 32 and repeated across the five token positions, and the final autoregressive token predicts the macro-step representation. The released `Embedder`, `ARPredictor`, and `ConditionalBlock` modules are imported without modification; only this input/output wrapper is new.

The same seven held-out outer seeds are used for both predictor families:

$$1835842072, 1863697692, 1591028803, 2420910389, 193429475, 2558417431, 2141736174.$$

Every seed satisfies the prespecified scientific and validity criteria for both families. For the expert bank, the mean state-R^2 and $\mathcal{H}_2$-fraction gaps are 0.4917 and 0.4512; the mean word conditioning gains are 0.1026 for q_+ and 0.1056 for q_-. For LeWorldModel, the corresponding quantities are 0.4880, 0.4513, 0.0977, and 0.0930. Its total trainable model has 817,156 parameters, including the 689,586-parameter CNN. Complete per-seed metrics and the training, selection, and evaluation records are included in the supplement.

Table 6: Protocol for the learned-predictor decision experiment.

Component	Prespecified setting
Representations	The seven confirmed LeWorldModel-stage CNN pairs; the encoders remain frozen throughout this experiment.
Prediction model	A fresh copy of the released LeWorldModel action embedder and AdaLN autoregressive predictor for each law. Paired laws use bitwise-identical initial weights.
Balanced fitting	512 independent transition pairs per word for training, sampled in exactly balanced mini-batches; 1,000 AdamW updates.
Label-free selection	256 independent pairs per word; checkpoint selection uses only held-out JEPA risk. State, goal, and Hermite labels remain hidden.
Decision panel	4,096 independent start–goal tasks. The learned predictor ranks all 32 words; exact state-space cost is evaluated only after the word is selected.
Primary outcomes	State-optimal action agreement, normalized oracle regret, and the fraction with normalized regret at most 1%.
Confirmation	Seven seed pairs, no exclusions; all prespecified per-seed criteria hold, and the three aggregate 95% confidence intervals exclude zero in the required direction.

Across the seven pairs, action agreement is 0.6543 (95% CI $[0.6297, 0.6789]$) under q_+ and 0.3735 ($[0.3535, 0.3935]$) under q_-. Mean normalized regret is 0.0727 ($[0.0571, 0.0883]$) versus 0.3658 ($[0.3229, 0.4088]$); the paired gap is 0.2931 ($[0.2502, 0.3360]$). The corresponding 1%-oracle rates

are 0.6975 and 0.3975. Independent residual panels verify nontrivial word conditioning for every fitted head. Training, selection, residual calibration, and decision panels are mutually disjoint, and no representation label enters fitting or checkpoint selection.

Table 7: Protocol for the nonlinear Gaussian-stationary experiment.

Component	Prespecified setting
Nonlinear channel	$Y = A_a R_{s_a t \sin(\|X\|^2)} X + \xi_a$, with $t = 0.2$ and opposite action signs; the rotation preserves γ_2 exactly.
Behavior laws	The exact full-support length-five pair q_+, q_-; primitive matrices, words, and laws are unchanged from the main construction.
Estimator	65,536 current states per word and two conditionally independent future replicas; empirical blocks on $\mathcal{H}_1 \oplus \mathcal{H}_2$.
Confirmation	Seven seeds, no exclusions; all prespecified validity, spectral, nonlinearity, and stationarity criteria hold, and the aggregate 95% confidence intervals exclude zero in the required directions.
Primary outcomes	Positive/negative diagonal phase margins, full top-two eigengaps, and the best-affine conditional-mean residual fraction.

The seven seeds are $7800101, \ldots, 7800107$. Their mean positive and negative margins are 0.02235 and -0.01172; the 95% confidence intervals exclude zero in opposite directions. The conditional mean's best-affine residual fraction is 0.01984 (95% CI $[0.01974, 0.01994]$). The supplement reports every seed and the complete aggregation rule.

The finite-pilot design study uses 20 prespecified seeds, $43301, \ldots, 43320$, and seven nested budgets of 64, 128, 256, 512, 1,024, 2,048, and 4,096 replica pairs per word. Its primary feasible set imposes the support floor $q_w \geq 1/2000$ and fixes total mass together with every degree-one, degree-two, and degree-three positional Walsh moment to its value under q_-. The 26-row constraint matrix has rank 26 and is independent of the estimated channels. For each word, preliminary spectral Catoni estimates of the first- and second-Hermite replicated-future cross moments are inserted into the phase-margin SDP; a T_1-only design and the unconstrained simplex are comparators. No seed, budget, or solver output is excluded, and the free simplex cannot replace a failure on the moment fiber. The prespecified endpoint is 4,096 pairs per word. The formal run contains $20 \times 7 \times 2 = 280$ cells and takes 51.1 seconds in the recorded environment. All prespecified validity and scientific criteria hold. The complete results and protocol are included in the anonymous supplement.

The open-loop word-ranking experiment is a separate controlled consequence cohort and does not use the pixel-confirmation checkpoints. It freezes the ten final direct-observation checkpoints from five paired seeds $\{3258276500, 1015767769, 3134186937, 298601391, 3977294129\}$. A fixed panel, generated with PCG64 seed 1472626054835472922, contains 512 independent start–goal pairs from $\mathcal{N}(0, I_2)$; every encoder ranks the same 32 length-five words. The state oracle uses the analytic Gaussian endpoint cost, while encoder costs use tensor-product Gauss–Hermite conditional expectation. Orders 48, 64, and 96 give identical word selections, regret, and 1%-oracle rates for every checkpoint; between orders 64 and 96, the worst relative RMS cost discrepancy is 3.83×10^{-8}, with no unresolved near ties. The reported Student-t intervals use the five paired neural-training seeds as independent units and are conditional on this fixed task panel. This evaluation performs no training and ran in 280.3 seconds on one RTX 4060 Ti. An earlier Monte Carlo evaluation did not meet its numerical-convergence criterion; the reported deterministic quadrature evaluation was specified after that diagnostic and is not presented as an outcome-blind replication.

For the label-free crossing experiment, the estimator observes a fixed hidden 5×5 orthogonal mixture of two normalized linear and three normalized quadratic coordinates. It receives two conditionally independent futures from the same latent reset and complete word, but no state or Hermite labels. Each of the ten formal seeds uses two independent panels, 32 words, and $2{,}097{,}152$ replica pairs per word and panel. Panel A selects the minimum eigengap on a 101-point grid followed by a 1,001-point local refinement; panel B evaluates the selected location independently. The exact root interval is used only after estimation. All ten seeds satisfy the prespecified geometry, cross-panel agreement, and independent-gap checks. The pooled estimates have mean 0.502102, 95% bootstrap CI $[0.499086, 0.504612]$, mean absolute root error 0.003368, and maximum absolute error 0.01145. The complete run took 19.7 seconds on one RTX 4060 Ti. This experiment establishes

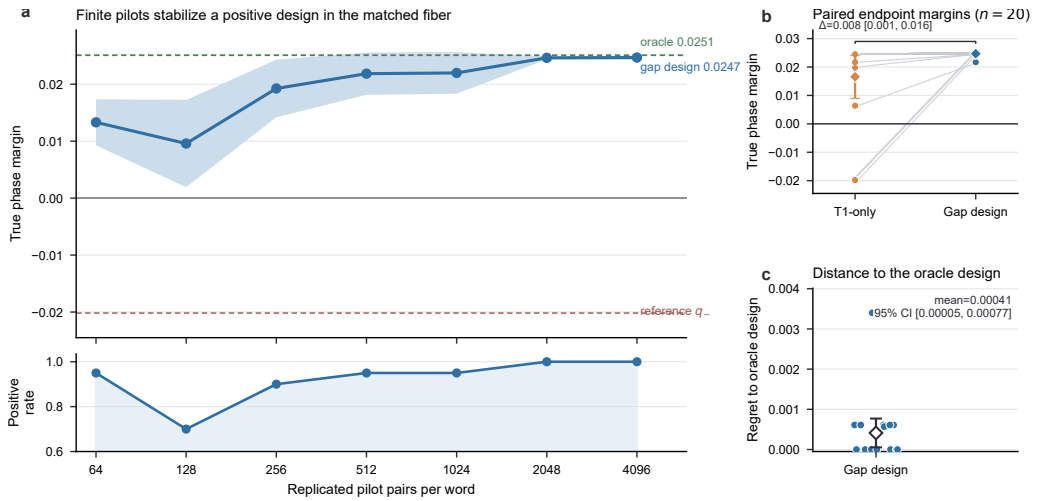

Figure 6: **Finite pilots turn the spectral criterion into behavior design. a**, True population margin and the fraction of positive-margin laws as the number of replica pairs per word increases; bands are seed-level Student-t 95% confidence intervals. **b**, Paired comparison at 4,096 pairs per word between full-margin and T_1-only design. **c**, Regret of the selected law relative to the population-oracle design. Diamonds and intervals denote means and seed-level 95% confidence intervals; small points denote the 20 independent runs. Every law matches the reference's positional action moments through degree three.

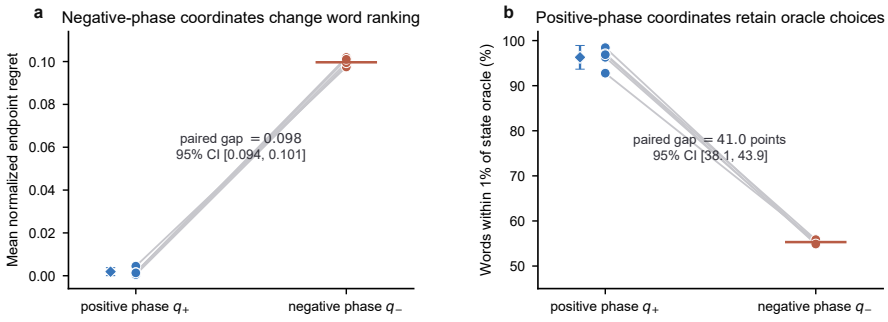

Figure 7: **In the controlled simulator, the learned coordinates alter open-loop action-word selection. a**, Normalized state-space regret (lower is better). **b**, Tasks whose oracle excess is at most 1% of the uniform-word excess baseline (higher is better). Lines join five paired neural-training seeds; intervals are across seed aggregates, conditional on the fixed task panel. This direct-observation, quadrature cohort is separate from the seven-pair pixel–LeWorldModel cohort in Figure 2c and Table 6.

observation-side detectability under reset access; it does not estimate the latent state or train an encoder.

N LEAN VERIFICATION

We formalize the main proof components in Lean 4.32.1, using Mathlib for standard analysis, probability, and linear-algebraic foundations (The mathlib Community, 2020). The replica/Hermite expansion and the all-order tensor monotonicity result are formalized end to end. For the remaining results, Lean checks the paper-specific derivation from the analytic, spectral, or finite-certificate inputs stated explicitly in the corresponding theorem interface. Table 8 gives the paper-to-code map.

Scope and methodology. We invoke a Mathlib theorem directly whenever its statement matches the needed mathematics. Results that are standard but not presently available in the required interface—for example, the particular infinite-dimensional Davis–Kahan reduction used in Appendix F—appear as named hypotheses of a paper-facing theorem. They are not introduced as Lean axioms. Thus *verified from stated premises* means that every subsequent inference is kernel checked, while the named premise remains justified by the paper proof and its cited source.

N.1 WHAT IS VERIFIED

The formalization follows the proof order rather than introducing a second mathematical narrative. P1 and P2 establish the Gaussian-chaos representation and the tail ordering used by the phase theorem. P3–P7 assemble the positive and negative phases, the sharp horizon construction, and its open continuation. P8–P14 cover the testing, estimation, design, stability, and predictor-capacity consequences. An auxiliary boundary target checks the Gaussian-divergence identity, the finite second-variation algebra, the exact rational gain, and inversion from the all-time flow law used in Proposition 17. Global completeness and Gaussian-measure preservation remain analytic steps proved in Appendix H; they are not presented as closed Lean declarations. Exact rational computations are represented by checked finite certificates; external analytic results remain visible in the theorem types.

Build information. The audited paper-facing source tree contains no project-defined `axiom`, `sorry`, or `admit`; the paper-facing declarations do not use `native_decide`. The recorded `#print axioms` audit reports only `propext`, `Classical.choice`, and `Quot.sound`, the standard foundations used by Lean and Mathlib. The anonymous supplement includes the exact theorem interfaces, a machine-readable coverage map, and the commands `lake build` and `lake env lean FntrAnalysis/AxiomsAudit.lean`. The boundary module also has the standalone replay target `lake env lean FntrAnalysis/T38/AxiomsAudit.lean`.

Component	Representative Lean declaration	Status
P1: replica identity and Hermite expansion	`paperP1_replica_and_fullHermiteExpansion`	VERIFIED
P2: all-order tail monotonicity	`wordLawChaosOperator_norm_succ_le`	VERIFIED
P3: positive representation phase	`paperP3_positivePhase`	VERIFIED*
P4: smooth invertible negative-phase competitor	`paperP4_negativePhase_with SeparatedDependencies`	VERIFIED*
P5: horizon sharpness and word-law lifting	`paperP5_lengthFour_exact_witness`; `paperP5_ambientOpen_minimum_is_five`	VERIFIED*
P6: exact matched pair and phase crossing	`paperP6_frozenPair_withSeparated SpectralDependency`	VERIFIED*
P7: ambient-open continuation	`paperP7_fixedPivot_wordPerturbation ambientOpen`	VERIFIED*
P8: local testing lower bound	`paperP8_localPhaseTestingHardness`	VERIFIED*
P9: robust phase estimation	`paperP9_replicaHermiteCatoni`	VERIFIED*
P10: exact semidefinite design form	`paperP10_sdp_exact`	VERIFIED*
P11: approximate whitening	`paperP11_allPhases_with SeparatedDependencies`	VERIFIED*
P12: approximate block stability	`paperP12_approximateBlockClosure Stability`	VERIFIED*
P13: predictor-capacity reversal and selection regret	`p13_full_frozen_theorem`; `normalized_selection_regret_le_two _mul_uniform_error_div`	VERIFIED*
P14: plug-in design regret	`paperP14_pluginDesignRegret`	VERIFIED*
Auxiliary Gaussian-preserving boundary	`exact_gaussianDivergence_zero`; `exact_diagonal_witness_secondVariation _positive`	VERIFIED*

Table 8: Lean 4 verification inventory. VERIFIED denotes a result closed in Lean using Mathlib. VERIFIED* denotes a kernel-checked implication from explicit analytic, spectral, or finite-certificate premises. The supplement records each such premise and its paper proof or exact certificate.