SPECTRAL SIGNATURES AND EMERGENT EQUIVARIANCE IN LOOPED TRANSFORMERS FOR IN-CONTEXT DYNAMICAL-SYSTEM PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

We investigate the internal representations and computations learned by looped Transformers for in-context prediction of stochastic linear dynamical systems. Pretrained on diverse next-state prediction tasks, the models develop approximate spectral degeneracies and shared operator structure without explicit equivariance constraints. We explain these observations through a shared hidden representation of orthogonal coordinate transformations, in which compact core matrices organize the model's operators. Projecting onto this structure yields a computational skeleton that largely preserves the original model's outputs and predictive performance. With a specified first-loop routing rule, the fixed cores support prediction at new input dimensions and longer contexts without retraining. This compact representation makes the learned computation easier to interpret and analyze, connecting its hidden organization to its predictive function and reuse.

1 INTRODUCTION

In-context learning allows a trained Transformer to adapt its predictions to a new task from examples without updating its parameters (Brown et al., 2020; Garg et al., 2022). Understanding this ability involves both identifying the learning procedure and explaining how a trained network implements it. Studies of simple function classes connect in-context prediction to familiar learning algorithms (Akyürek et al., 2023; von Oswald et al., 2023a), while analyses of hidden representations examine how task information is encoded and used in subsequent computation (Hendel et al., 2023; Guo et al., 2024). Together, these perspectives motivate describing the learned computation in terms of its internal representations and operators. Such a description would allow us to test which components support prediction and whether they can be reused in new settings.

We study this organization through next-state prediction on trajectories from dynamical systems whose parameters are not provided to the model. Classical regression benchmarks supply explicit input–output examples; a trajectory instead presents transition examples implicitly through successive states. This setting shares the autoregressive prediction format studied in sequence modeling (Sander et al., 2024; von Oswald et al., 2023b). To predict a new system, the model must use the observed trajectory to adapt its predictions to that system's dynamics. Training across systems therefore provides a controlled setting for studying how shared model parameters support computation that depends on the context.

Our task family consists of stochastic linear systems motivated by Ornstein–Uhlenbeck dynamics, with varying dynamical properties. We train a compact attention-only looped Transformer (Yang et al., 2024) that repeatedly applies a shared block of causal softmax attention and residual updates. Training minimizes next-state prediction error; evaluation uses systems absent from the training-task set. We then examine the hidden states and parameter operators to identify structure shared across inputs, heads, and iterations.

Across trained models, we observe approximate spectral degeneracies and shared operator structure across heads. We explain these regularities through an approximately equivariant hidden representation of physical coordinate transformations. This structure is not enforced during training. A shared basis fitted from hidden activations reveals vector channels and predicts parameter structure not imposed during fitting. In this basis, compact core matrices separate learned coefficients from physical

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.

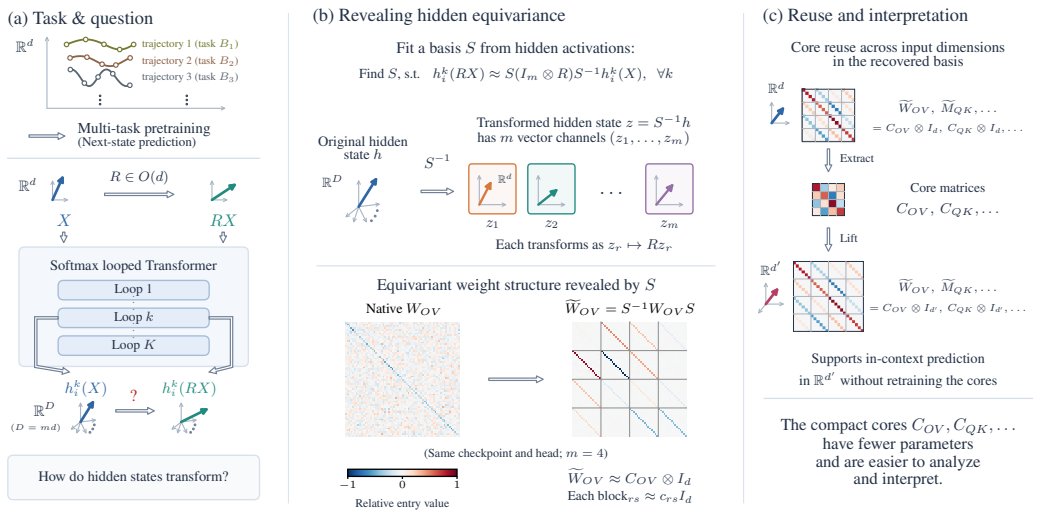

Figure 1: Overview of the study.

coordinates, yielding a computational skeleton that can be analyzed and instantiated at new input dimensions. Projection, transfer experiments, and interventions test the predictive role and reuse of this structure (Figure 1).

Our contributions are:

- **Structural identification and theoretical explanation.** We provide an equivariance-based explanation for the parameter structure and spectral signatures observed in trained models. A basis fitted from activations reveals the predicted structure across embedding, readout, value updates, and attention routing in trained models.
- **Functional validation and reuse.** Structural projection approximately preserves outputs and predictive accuracy. With a specified first-loop routing rule, fixed cores predict at new input dimensions and longer contexts without retraining within this task family.
- **Circuit interpretation.** The recovered representation exposes induction-like state-pair retrieval and common/contrast updates. Interventions establish their functional roles in the projected models.

1.1 Related Work

In-context algorithms and autoregressive prediction. Linear regression provides a tractable setting for studying in-context learning (Garg et al., 2022; Akyürek et al., 2023; Li et al., 2025; Bai et al., 2023). Analyses connect learned predictors to optimization procedures (von Oswald et al., 2023a; Ahn et al., 2023) and characterize training under linear-attention assumptions (Zhang et al., 2024). Bayesian perspectives describe inference over latent tasks (Xie et al., 2022; Zhang et al., 2025), while task-diversity experiments reveal departures from the finite-task Bayesian predictor (Raventós et al., 2023). For autoregressive sequences, studies connect attention trained on linear dynamics to in-context optimization (Sander et al., 2024; Zheng et al., 2024). For stochastic linear systems, prior work provides approximation and depth-separation bounds (Cole et al., 2025), analyzes learned attention and autoregressive optimization (Vlădărean et al., 2026), and studies filtering for unknown systems (Du et al., 2023). Looped models also admit analyses of multistep gradient descent under linear-attention restrictions (Gatmiry et al., 2024). We examine how a trained softmax predictor organizes its hidden states and operators.

Representations learned through prediction. Task and function vectors encode information from demonstrations in activations that can influence subsequent task execution (Hendel et al., 2023; Todd et al., 2024). In compositional regression, trained Transformers separate representation pro-

cessing from in-context linear learning (Guo et al., 2024). For sequences generated by hidden-state processes, Shai et al. (2024) identify belief-state geometry in the residual stream. We study how a shared transformation law for hidden states predicts structure across the operators acting on them.

Circuit mechanisms and computation extraction. Induction heads retrieve successors of matching tokens (Elhage et al., 2021; Olsson et al., 2022), and subsequent work studies their formation (Bietti et al., 2023; Reddy, 2024; Singh et al., 2024). Our induction-like interpretation concerns matching continuous states stored in historical pairs and retrieving their successors. For computation extraction, von Oswald et al. (2023b) recover block coefficients from linear-attention models and analyze softmax models through probes and linearization. Lutz et al. (2026) enforce layerwise feature and label permutation equivariance to identify update rules in softmax classification models. We recover a shared basis from activations and test its predictions for parameter structure, predictive behavior, and reuse of fixed cores.

Learned and imposed equivariant structure. Prior work learns equivariant latent structure from sequences (Miyato et al., 2022; Koyama et al., 2024), measures learned equivariance (Gruver et al., 2023), and studies its emergence and training dynamics (Domina et al., 2026; Shen et al., 2025). Invariant scalars and vector combinations (Villar et al., 2021), and explicitly rotation-equivariant attention (Assaad et al., 2023), provide established structural principles for vector-valued computation. We test whether these principles emerge in a trained predictor and support prediction and reuse across input dimensions.

2 Dynamical-System In-Context Learning

2.1 Task Family and Prediction

Task family. Each task τ is a d-dimensional stochastic linear system:

$$x_{t+1} = B_\tau x_t + \eta_t, \qquad B_\tau = \exp(-\Gamma_\tau \Delta t), \qquad \eta_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 2\Delta t I_d). \tag{1}$$

The initial state $x_1 \sim \mathcal{N}(0, I_d)$ is independent of the task and noise sequence. We parameterize the drift as

$$\Gamma_\tau = (I_d + J_\tau) H_\tau, \qquad H_\tau = H_\tau^\top \succ 0, \qquad J_\tau^\top = -J_\tau, \tag{2}$$

following a potential–circulation decomposition of nonequilibrium stochastic dynamics (Kwon et al., 2005; Lu & Tang, 2026). Here H_τ sets the confinement geometry, while J_τ introduces nonreversible circulation. Varying confinement scales, anisotropy, and circulation strength yields diverse relaxation and rotational dynamics, including fast and slow modes.

We sample K_{task} training-task matrices B_τ from the resulting distribution $\mathcal{P}$, which is invariant under orthogonal coordinate changes. Appendix A specifies the sampling procedure and distinguishes our discrete noise model from exact Ornstein–Uhlenbeck (OU) sampling. Main experiments use $d = 16$, $\Delta t = 0.005$, and $T_{\text{seq}} = 128$ states per trajectory.

In-context prediction. Each state is represented as a token. At position i, the model predicts x_{i+1} from $X_{\le i} = [x_1, \dots, x_i] \in \mathbb{R}^{d \times i}$; task parameters are not supplied. We write a context of length T as $X = [x_1, \dots, x_T]$. Training uses $T = T_{\text{seq}} - 1$ and minimizes next-state MSE, averaged over positions and coordinates:

$$\mathcal{L}(\theta; x_{1:T+1}) = \frac{1}{dT} \sum_{i=1}^{T} \| f_\theta(X_{\le i}) - x_{i+1} \|_2^2 . \tag{3}$$

On unseen tasks, we evaluate prediction error against both the noisy next state x_{i+1} and its conditional mean $B_\tau x_i$.

2.2 Orthogonal Symmetry

The population task distribution has no preferred coordinate system. For every $R \in O(d)$, a transition matrix $B_\tau \sim \mathcal{P}$ satisfies

$$B_\tau \overset{\text{d}}{=} R B_\tau R^\top . \tag{4}$$

Together with isotropic initial states and noise, this invariance implies $(X_{\le t}, x_{t+1}) \overset{\mathrm{d}}{=} (RX_{\le t}, Rx_{t+1})$, where the same R acts on every state in the context. Consequently, the Bayes-optimal predictor under squared error admits an equivariant version:

$$f^*(RX_{\le t}) = Rf^*(X_{\le t}), \qquad f^*(X_{\le t}) = \mathbb{E}[x_{t+1} \mid X_{\le t}], \tag{5}$$

with equality understood almost surely. This symmetry is a property of the task mixture, not generally of an individual fixed task or a finite training sample. It constrains optimal prediction without specifying how a trained network represents the symmetry internally.

2.3 Transformer Architecture

We use a causal, attention-only looped Transformer that repeatedly applies a shared block for K iterations. The main models use hidden dimension $D = 64$, two attention heads each of width $d_{\mathrm{head}} = D = 64$, and $K = 8$ iterations. We use this compact configuration to enable efficient training and analysis, rather than optimizing the architecture for predictive performance.

Each state is embedded by $W_E \in \mathbb{R}^{D\times d}$ and decoded by an independently learned readout $W_U \in \mathbb{R}^{d\times D}$. For head a, the learned value and output projections define $W^a_{OV} = W^a_O W^a_V \in \mathbb{R}^{D\times D}$. Using column vectors and starting from $h^0_i = W_E x_i$, the forward computation is

$$h^{k+1}_i = h^k_i + \sum_{a=1}^{n_{\mathrm{head}}} W^a_{OV} \sum_{j=1}^{i} \alpha^{a,k}_{ij} h^k_j, \quad k = 0, \dots, K-1, \tag{6}$$

$$\widehat{x}_{i+1} = f_\theta(X_{\le i}) = W_U h^K_i, \tag{7}$$

Here h^k_i is the hidden state at position i after k iterations; $\alpha^{a,k}_{ij}$ is head a's causal attention weight from query i to key j, recomputed from the hidden states at each iteration.

Queries and keys are RMS-normalized; values are unnormalized. Routing uses three forms:

$$M^a_{QK} = (W^a_Q)^\top W^a_K, \qquad M^a_{QQ} = (W^a_Q)^\top W^a_Q, \qquad M^a_{KK} = (W^a_K)^\top W^a_K. \tag{8}$$

The QK form determines the score numerator, while QQ and KK determine the query and key normalization factors. A learned relative-position bias is added to the attention logits only in the first iteration. Appendix B gives the full attention equations and architectural details.

3 Spectral Signatures and Shared Structure

We observe approximate spectral degeneracies in individual OV operators and a large subspace of linear maps that approximately commute with both heads. The analysis uses 25 reference models trained with $K_{\mathrm{task}} = 2000$, crossing five data seeds with five training seeds. Appendix C.1 gives the training details and the fixed representative model used for visualizations.

Eigenvalue clustering and singular-value steps. The representative model's OV eigenvalues form clusters during training (Figure 2(a)). Fitting $m = 4$ clusters of prescribed size $d = 16$ gives small within-cluster deviations relative to the overall eigenvalue spread across the reference models; Figure 2(b) tracks this relative error during training. The sorted singular values also show approximate steps in groups of d (Figure 2(c)). Appendix C.2 defines both diagnostics and reports aggregate errors.

Shared structure across heads. We next examine the two OV operators jointly through their *common commutant*: the linear maps $Y \in \mathbb{R}^{D\times D}$ satisfying $YW^a_{OV} = W^a_{OV}Y$ for both heads (Goodman & Wallach, 2009). To measure approximate commutation, we use the linear operator

$$\mathcal{A}(Y) = n^{-1/2}_{\mathrm{head}} (Y\widehat{W}^a_{OV} - \widehat{W}^a_{OV}Y)_a, \tag{9}$$

where $\widehat{W}^a_{OV}$ is the trace-centered, Frobenius-normalized OV matrix of head a. The kernel of $\mathcal{A}$ is the common commutant; small singular values identify matrix directions that approximately commute with both heads. Appendix C.2 gives its vectorized matrix representation in Equation 43. We order the singular values from smallest to largest; the first is trivially zero because $\mathcal{A}(I_D) = 0$.

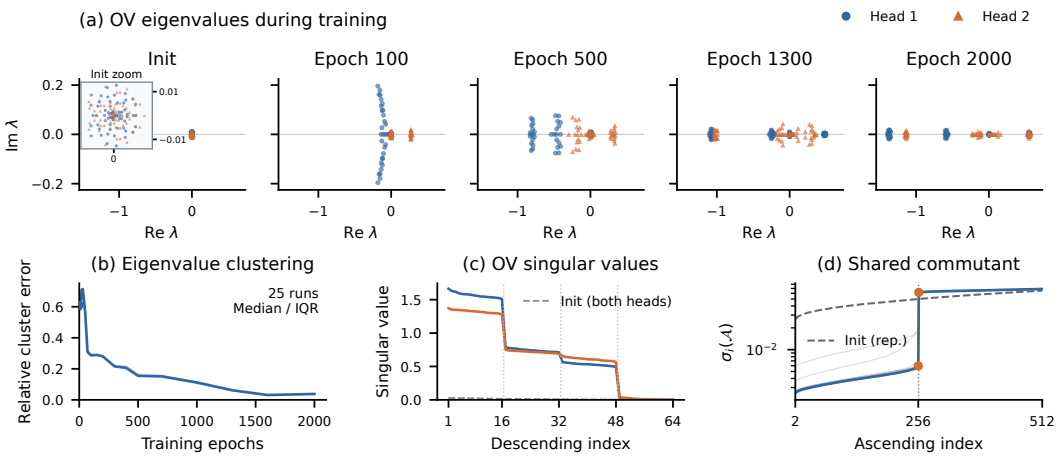

Figure 2: **Spectral structure emerges during training.** (a) OV eigenvalues of the fixed representative model, with shared axis limits across training snapshots; blue circles and orange triangles denote heads 1 and 2. The inset enlarges initialization. (b) Relative clustering error: median and interquartile range across 25 runs, after averaging the two heads per run. (c) Singular values at initialization (gray) and endpoint (colored); vertical guides mark groups of 16. (d) Common-commutator spectra: endpoints in gray, representative in blue, and its initialization dashed. Indices 2–512 of the 4096 singular values are shown; the trivial first mode is omitted. Markers indicate indices 256 and 257.

The bottom $d^2 = 256$ singular values are small and separated from the remaining spectrum (Figure 2(d)), indicating a large subspace of maps that approximately commute with both OV operators. Thus, the two heads exhibit shared structure beyond their individual spectral degeneracies. Appendix C.2 reports the gap and residual magnitudes. These spectral degeneracies and shared commuting directions emerge without explicit architectural constraints enforcing them. Section 4 develops an equivariant explanation of these observations.

4 Emergent Equivariant Structure and Transfer

We explain the spectral signatures in Section 3 through a shared equivariant representation of hidden states. Projection approximately preserves predictions and slightly improves accuracy. The extracted computation transfers across physical dimensions and context lengths without retraining.

4.1 Equivariant Representation and Predictions

We propose a shared hidden representation of physical rotations and reflections.

Hypothesis 1 (Shared hidden representation)**.** For $D = md$, there is an invertible S, shared across heads, positions, and iterations, such that hidden states approximately satisfy

$$h_i^k(RX) \approx \rho(R) h_i^k(X), \qquad \rho(R) = S(I_m \otimes R)S^{-1}, \quad R \in O(d). \tag{10}$$

Writing $h = Sz$, the coordinates z comprise m vectors $z_r \in \mathbb{R}^d$, each transforming according to the standard representation of $O(d)$ as $z_r \mapsto Rz_r$ (Hall, 2015); the reference model has $m = 4$ and $d = 16$. The next result gives the parameter structure when maps preserve these vector transformations and routing forms preserve attention scores.

Theorem 1 (Equivariant parameter structure)**.** *Fix $\rho(R) = S(I_m \otimes R)S^{-1}$ with S invertible and $D = md$. Suppose that, for every $R \in O(d)$ and every head a,*

$$\begin{aligned} \rho(R)W_E &= W_E R, & W_U\rho(R) &= RW_U, \\ W_{OV}^a\rho(R) &= \rho(R)W_{OV}^a, & \rho(R)^\top M_{AB}^a \rho(R) &= M_{AB}^a, \end{aligned} \qquad AB \in \{QQ, KK, QK\}. \tag{11}$$

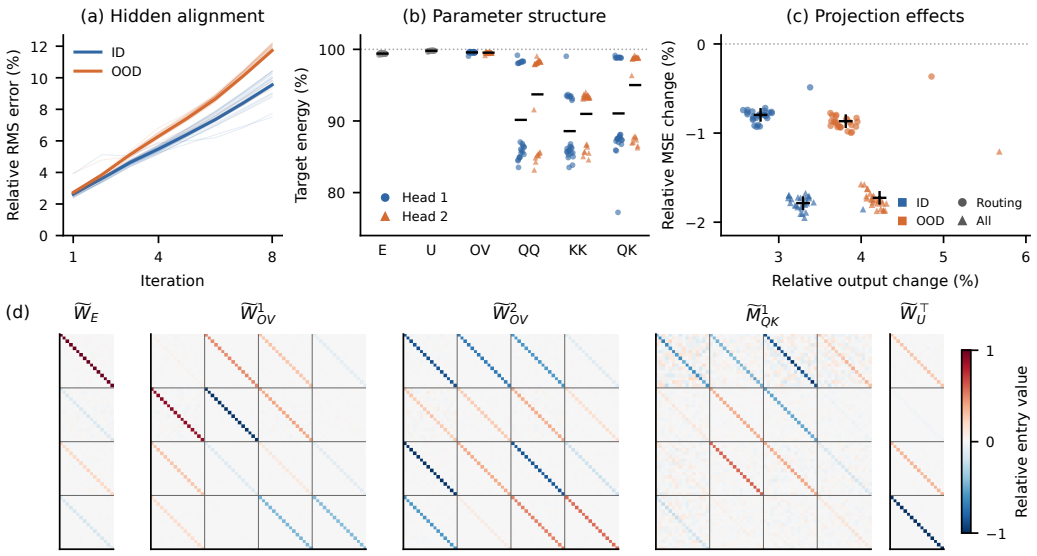

Figure 3: **Shared equivariant structure and its predictive role.** (a) Hidden-state alignment on ID/OOD inputs; thin lines show runs, thick lines their means. (b) Target energy in the fitted basis; heads 1/2 are offset left/right within each module, with black bars marking means. (c) Relative output and conditional-mean MSE changes after projecting routing forms only or all modules, without retraining. Each point pairs both metrics for one run on ID or OOD data (2000 trajectories each); black crosses mark means. Negative MSE change indicates improvement. All 25 runs are included in (a–c). (d) Representative transformed matrices before projection, each divided by its maximum absolute entry; grid lines separate 16×16 blocks.

Then, in the coordinates defined by S,

$$\begin{aligned} S^{-1}W_E &= e \otimes I_d, & W_U S &= u^\top \otimes I_d, \\ S^{-1}W^a_{OV}S &= C^a_{OV} \otimes I_d, & S^\top M^a_{AB} S &= C^a_{AB} \otimes I_d. \end{aligned} \tag{12}$$

Here $e, u \in \mathbb{R}^m$ and the core matrices are $m \times m$; C^a_{QQ} and C^a_{KK} are positive semidefinite. Under the architecture of Section 2.3, these conditions also imply $f_\theta(RX) = Rf_\theta(X)$.

The proof and spectral consequences are derived in Appendix D. Intuitively, applying an OV operator to $h = Sz$ is equivalent to applying $C^a_{OV} \otimes I_d$ to z and mapping the result back through S. Figure 3(d) illustrates this block pattern in the transformed weights of a trained model. The cores mix vector channels, with a parameterization whose size does not depend on physical dimension.

This structure explains two observations in Section 3. Each eigenvalue of an OV core is repeated d times in the full operator. Moreover, applying the same linear transformation to every vector copy commutes with all OV operators, yielding at least d^2 independent shared commuting directions. Native-coordinate singular-value repetition is also predicted when the basis is orthogonal, but does not follow for a general invertible S. We next test the representation and parameter structure in trained models and evaluate prediction under structural projection.

4.2 Empirical Validation and Structural Projection

We use the same reference checkpoints to evaluate unseen tasks from the training distribution (ID) and a joint task-parameter shift (OOD); Appendix C.1 specifies both sets.

For each model, we fit a shared basis S by aligning hidden states produced by paired inputs X and RX according to Equation 10. We first fit an orthogonal basis, then refine it within a constrained family of invertible transformations. The fitting objective uses only hidden-state alignment, without

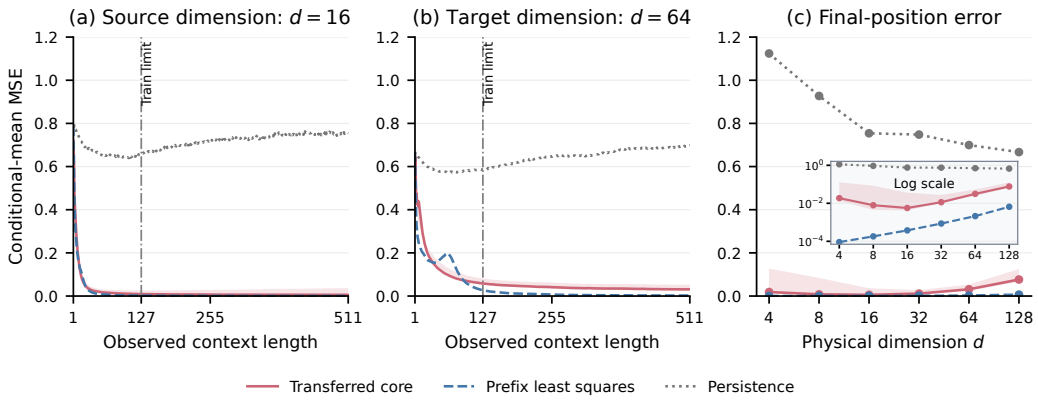

Figure 4: **Extracted computation transfers across dimensions and context lengths without retraining.** (a,b) Per-position MSE at $d = 16$ and 64; vertical lines mark the training context limit of 127 states. (c) Final-position MSE after 511 observed states across six physical dimensions; the inset shows the same results on a logarithmic MSE axis. Errors average 2000 test trajectories and physical coordinates against conditional means $B_\tau x_t$. The transferred-core curve shows the mean over 25 source models; shading spans their full range. Baselines use the same test data.

imposing the predicted parameter structure or optimizing prediction error. The basis is then fixed for all analyses (Appendix C.3).

The fitted representation captures hidden-state transformations on held-out inputs and orthogonal transformations (Figure 3(a)). Mean relative RMS alignment errors are approximately 6–8% on ID/OOD tasks, including all iterations and input positions.

Using the fitted basis S, we transform the model's maps and routing forms as in Theorem 1. Let G denote any resulting matrix, partitioned into $d \times d$ blocks G_{ij}. The representative model's transformed matrices show the predicted block pattern (Figure 3(d)).

Prediction under structural projection. To keep only the predicted equivariant structure, we replace each block by a scalar multiple of the identity, using its mean diagonal entry:

$$[\Pi_d(G)]_{ij} = \frac{\mathrm{tr}(G_{ij})}{d} I_d. \tag{13}$$

The target energy fraction $\|\Pi_d(G)\|_F^2 / \|G\|_F^2$ measures the fraction of squared Frobenius norm preserved by projection. Averaged across runs and, where applicable, heads, it exceeds 99% for each of W_E, W_U, and OV, compared with approximately 90%–93% for the routing forms (Figure 3(b)). These fractions describe structure in the activation-fitted basis; they do not establish the maximum attainable over admissible bases.

To test functional preservation, we either jointly project W_E, W_U, both OV operators, and all routing forms, or project only the routing forms, without retraining. Joint projection changes outputs by only a few percent on average and slightly reduces conditional-mean MSE on both evaluation sets (Figure 3(c)). Routing-only projection has a similarly small effect on outputs despite its lower target energy. Both interventions improve MSE in every evaluated model on both sets. Thus, deviations from the recovered structure contribute little to prediction on these tasks. Appendix C.4 reports exact changes and additional module-wise interventions.

4.3 Transferring the Extracted Computation

The Kronecker structure separates core coefficients from physical coordinates, allowing us to test whether the learned computation can be reused at new dimensions and context lengths. We extract e, u and the 4×4 cores by normalized block traces, then replace I_{16} by I_d while keeping these coefficients fixed. Predictors reconstructed in random orthogonal and standard hidden bases agree

up to numerical precision (Appendix E). We test all reference cores at six dimensions from $d = 4$ to 128, with contexts up to 511 states, roughly four times the training limit of 127.

In the source models' first iteration, one attention head attends almost exclusively to the previous token and the other to the current token. For transfer, we use fixed positional biases to strongly favor these attention patterns while keeping the corresponding head's core coefficients fixed (Appendix E.2). Later iterations use content-based routing without positional bias. Evaluation uses unseen tasks with the confinement scale, anisotropy, and circulation strength sampled from the same ranges as in pretraining. We compare conditional-mean MSE with persistence and causal prefix least squares (Appendix E).

The extracted cores support prediction across the tested dimensions and beyond the training context window without retraining (Figure 4). Across higher input dimensions, the ratio of mean final-position MSE to least-squares MSE remains comparable to that at the training dimension, while errors stay substantially below persistence (Figure 4(c)). At the training dimension, a matched-target comparison further shows that additional history improves prediction for nearly all source models (Appendix E.3). These results demonstrate the predictive utility of the extracted coefficients beyond the training dimension.

5 Interpreting the Extracted Computation

We use the recovered representation to identify familiar retrieval motifs and test their functional roles in all jointly projected reference models. The tests probe state-pair formation, matching, and two-head updates within these projected source models (Appendix G).

Forming and matching state pairs. Let $Z_j^k \in \mathbb{R}^{4\times d}$ denote the recovered vector channels at position j after k iterations, initialized as $Z_j^0 = ex_j^\top$. In the first iteration, the two heads attend primarily to the current and previous positions, forming state pairs as in early token-copying operations in autoregressive Transformers (von Oswald et al., 2023b). Grouping input sources gives, for $j > 1$, the exact decomposition

$$Z_j^1 = f_{c,j}x_j^\top + f_{p,j}x_{j-1}^\top + R_j, \tag{14}$$

where $f_{c,j} = e + \sum_a \alpha_{jj}^{a,0} C_{OV}^a e$ and $f_{p,j} = \sum_a \alpha_{j,j-1}^{a,0} C_{OV}^a e$ are the current- and previous-state coefficients in $\mathbb{R}^4$. The sum runs over both heads, and e accounts for the residual connection. The coefficients use actual attention weights and are not separately fitted. The remainder R_j collects contributions from earlier states x_ℓ with $\ell < j - 1$.

Earlier-state contributions have little effect on prediction: removing R from Z^1 and rerunning subsequent iterations changes mean MSE by less than 0.3% on either evaluation set.

In the second iteration, matching x_i to a stored predecessor x_{j-1} provides access to its successor x_j: a continuous-state analogue of induction-head retrieval (Olsson et al., 2022) (Figure 5(c)). The QK numerator contains the current–previous (CP) term

$$N_{ij,\mathrm{CP}}^a = (f_{c,i}^\top C_{QK}^a f_{p,j})(x_i^\top x_{j-1}). \tag{15}$$

In the head with the first-iteration self role, removing CP increases final MSE in every model, whereas removing current–current (CC) matching has little effect on most models. Keeping only CP leaves performance close to baseline for most models, but substantially worsens prediction in two cases (Figure 5(a)).

Differential retrieval across heads. The two-head update decomposes into common and contrast contributions:

$$\begin{aligned} \Delta Z_{i,\mathrm{common}}^k &= \frac{1}{2}\sum_{j\le i}(\alpha_{ij}^{1,k} + \alpha_{ij}^{2,k})(C_{OV}^1 + C_{OV}^2)Z_j^k, \\ \Delta Z_{i,\mathrm{contrast}}^k &= \frac{1}{2}\sum_{j\le i}(\alpha_{ij}^{1,k} - \alpha_{ij}^{2,k})(C_{OV}^1 - C_{OV}^2)Z_j^k. \end{aligned} \tag{16}$$

The residual stream update is $Z_i^{k+1} = Z_i^k + \Delta Z_{i,\mathrm{common}}^k + \Delta Z_{i,\mathrm{contrast}}^k$. The contrast term weights the shared OV difference by signed attention differences across positions. A single standard softmax head cannot directly implement this pattern on the same hidden states in one update, since its

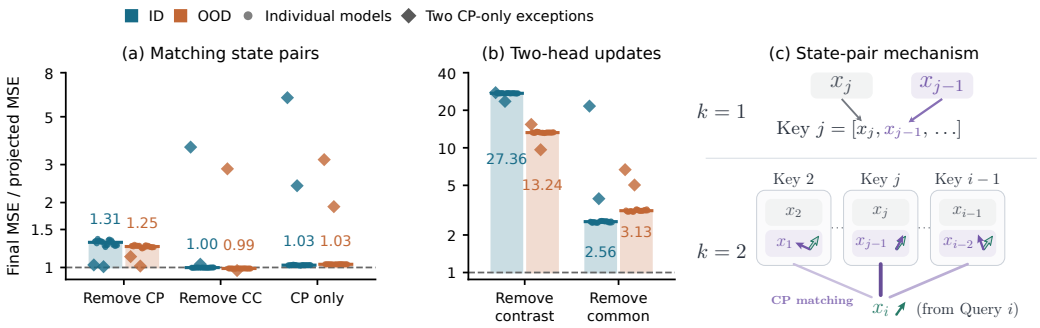

Figure 5: **Testing the extracted computation.** Figure 5(a,b) shows ID/OOD MSE ratios to the projected baseline across 25 models: (a) second-iteration numerator interventions in the self-role head; (b) common/contrast-only updates from iteration two onward. Bars span unity to the labeled medians; diamonds track the same two CP-only exceptions. Log axes have different ranges. Effect sizes are not directly comparable between (a) and (b) because intervention scopes differ. (c) Schematic of state-pair formation at $k = 1$ and CP matching at $k = 2$.

attention weights are nonnegative and its OV map is shared (Appendix G.3). Whereas Differential Transformer explicitly subtracts softmax maps (Ye et al., 2025), our contrast term arises from decomposing two ordinary heads with distinct OV cores. Removing contrast from iteration two onward substantially increases MSE in every model on both evaluation sets (Figure 5(b)). Removing the common term over the same iterations causes a smaller MSE increase, indicating that both components contribute, with greater sensitivity to contrast removal. Both interventions preserve the residual connection and recompute attention from the evolving states. Our contrast dependence is consistent with beneficial signed head combinations studied in shallow in-context regression (He et al., 2025; Cui et al., 2025). The early retrieval operations supply intermediate representations for further computation; subsequent loops process this information before the final readout.

6 Discussion and Conclusion

In this work, we recover a compact computational structure from a looped Transformer trained for next-state prediction. A shared basis reveals vector channels and core matrices that separate learned coefficients from physical coordinates. Projection approximately preserves prediction, and fixed cores support transfer under a specified first-loop routing rule. This representation also exposes induction-like retrieval and differential updates whose functional roles can be tested by intervention.

These findings connect the organization of hidden representations to interpretable and reusable computation. The identified retrieval mechanisms produce intermediate representations for further processing, rather than a complete prediction algorithm. Equivariance enables reconstruction at new input dimensions, and our experiments establish the predictive utility of this transfer within the studied task family. Shared structure need not imply identical algorithms. How later loops process retrieved information, how training selects the cores, and how far their reuse extends beyond the studied task family remain open questions.

AI use statement

We used generative AI tools to assist with checking mathematical derivations, code implementation, and exploratory data analysis. We reviewed the AI-assisted code for experiments and analyses. AI tools also assisted with manuscript organization and drafting, with the initial draft developed under the authors' direction and with the content of each sentence specified by the authors. Experimental data were generated by dynamical-system simulations, not by generative AI. We take responsibility for the final content of this work.

REFERENCES

Kwangjun Ahn, Xiang Cheng, Hadi Daneshmand, and Suvrit Sra. Transformers learn to implement preconditioned gradient descent for in-context learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/8ed3d610ea4b68e7afb30ea7d01422c6-Abstract-Conference.html.

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=0g0X4H8yN4I.

Serge Assaad, Carlton Downey, Rami Al-Rfou, Nigamaa Nayakanti, and Ben Sapp. VN-Transformer: Rotation-equivariant attention for vector neurons. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=EiX2L4sDPG.

Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/b2e63e36c57e153b9015fece2352a9f9-Abstract-Conference.html.

Alberto Bietti, Vivien Cabannes, Diane Bouchacourt, Hervé Jégou, and Léon Bottou. Birth of a Transformer: A memory viewpoint. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://papers.nips.cc/paper_files/paper/2023/hash/0561738a239a995c8cd2ef0e50cfa4fd-Abstract-Conference.html.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

Frank Cole, Yuxuan Zhao, Yulong Lu, and Tianhao Zhang. In-context learning of linear dynamical systems with Transformers: Approximation bounds and depth-separation. In *Advances in Neural Information Processing Systems*, 2025. URL https://papers.nips.cc/paper_files/paper/2025/hash/f7ae4fe91d96f50abc2211f09b6a7e49-Abstract-Conference.html.

Yingqian Cui, Jie Ren, Pengfei He, Hui Liu, Jiliang Tang, and Yue Xing. Superiority of multi-head attention: A theoretical study in shallow Transformers in in-context linear regression. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 937–945. PMLR, 2025. URL https://proceedings.mlr.press/v258/cui25a.html.

M. Domina, J. W. Abbott, P. Pegolo, F. Bigi, and M. Ceriotti. How unconstrained machine-learning models learn physical symmetries. *PRX Intelligence*, 1:013013, 2026. doi: 10.1103/cnsd-p8nc. URL https://doi.org/10.1103/cnsd-p8nc.

Zhe Du, Haldun Balim, Samet Oymak, and Necmiye Ozay. Can Transformers learn optimal filtering for unknown systems? *IEEE Control Systems Letters*, 7:3525–3530, 2023. doi: 10.1109/LCSYS.2023.3335318. URL https://doi.org/10.1109/LCSYS.2023.3335318.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris

Olah. A mathematical framework for Transformer circuits. *Transformer Circuits Thread*, 2021. URL https://transformer-circuits.pub/2021/framework/index.html.

Shivam Garg, Dimitris Tsipras, Percy Liang, and Gregory Valiant. What can Transformers learn in-context? a case study of simple function classes. In *Advances in Neural Information Processing Systems*, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/c529dba08a146ea8d6cf715ae8930cbe-Abstract-Conference.html.

Khashayar Gatmiry, Nikunj Saunshi, Sashank J. Reddi, Stefanie Jegelka, and Sanjiv Kumar. Can looped Transformers learn to implement multi-step gradient descent for in-context learning? In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 15130–15152. PMLR, 2024. URL https://proceedings.mlr.press/v235/gatmiry24b.html.

Roe Goodman and Nolan R. Wallach. *Symmetry, Representations, and Invariants*, volume 255 of *Graduate Texts in Mathematics*. Springer, New York, 2009. doi: 10.1007/978-0-387-79852-3. URL https://doi.org/10.1007/978-0-387-79852-3.

Nate Gruver, Marc Finzi, Micah Goldblum, and Andrew Gordon Wilson. The Lie derivative for measuring learned equivariance. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=JL7Va5Vy15J.

Tianyu Guo, Wei Hu, Song Mei, Huan Wang, Caiming Xiong, Silvio Savarese, and Yu Bai. How do Transformers learn in-context beyond simple functions? a case study on learning with representations. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/c9694bf4f9bf3626f7d21158bab74f8e-Abstract-Conference.html.

Brian C. Hall. *Lie Groups, Lie Algebras, and Representations: An Elementary Introduction*, volume 222 of *Graduate Texts in Mathematics*. Springer, Cham, second edition, 2015. doi: 10.1007/978-3-319-13467-3. URL https://doi.org/10.1007/978-3-319-13467-3.

Jianliang He, Xintian Pan, Siyu Chen, and Zhuoran Yang. In-context linear regression demystified: Training dynamics and mechanistic interpretability of multi-head softmax attention. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 22686–22742. PMLR, 2025. URL https://proceedings.mlr.press/v267/he25q.html.

Roee Hendel, Mor Geva, and Amir Globerson. In-context learning creates task vectors. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 9318–9333. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-emnlp.624. URL https://aclanthology.org/2023.findings-emnlp.624/.

Masanori Koyama, Kenji Fukumizu, Kohei Hayashi, and Takeru Miyato. Neural Fourier transform: A general approach to equivariant representation learning. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/1aa1fde3661b23ba9b043082069fd144-Abstract-Conference.html.

Chulan Kwon, Ping Ao, and David J. Thouless. Structure of stochastic dynamics near fixed points. *Proceedings of the National Academy of Sciences*, 102(37):13029–13033, 2005. doi: 10.1073/pnas.0506347102. URL https://doi.org/10.1073/pnas.0506347102.

Yuhao Li, Ruoran Bai, and Haiping Huang. Spin-glass model of in-context learning. *Phys. Rev. E*, 112(1):L013301, Jul 2025. doi: 10.1103/5l5m-4nk5. URL https://link.aps.org/doi/10.1103/5l5m-4nk5.

Haiqi Lu and Ying Tang. Steering dynamical regimes of diffusion models by breaking detailed balance. *arXiv preprint arXiv:2602.15914*, 2026. URL https://arxiv.org/abs/2602.15914.

Patrick Lutz, Themistoklis Haris, Arjun Chandra, Aditya Gangrade, and Venkatesh Saligrama. Symmetry reveals layerwise dynamics: How Transformers perform in-context classification. In *Proceedings of the International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2604.11613v4.

Takeru Miyato, Masanori Koyama, and Kenji Fukumizu. Unsupervised learning of equivariant structure from sequences. In *Advances in Neural Information Processing Systems*, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/0503f5dce343a1d06d16ba103dd52db1-Abstract-Conference.html.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads. *Transformer Circuits Thread*, 2022. URL https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.

Allan Raventós, Mansheej Paul, Feng Chen, and Surya Ganguli. Pretraining task diversity and the emergence of non-Bayesian in-context learning for regression. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://papers.nips.cc/paper_files/paper/2023/hash/2e10b2c2e1aa4f8083c37dfe269873f8-Abstract-Conference.html.

Gautam Reddy. The mechanistic basis of data dependence and abrupt learning in an in-context classification task. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/540a3bf4863672cdaebfed69fa2ce335-Abstract-Conference.html.

Michael E. Sander, Raja Giryes, Taiji Suzuki, Mathieu Blondel, and Gabriel Peyré. How do Transformers perform in-context autoregressive learning? In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 43235–43254. PMLR, 2024. URL https://proceedings.mlr.press/v235/sander24a.html.

Adam S. Shai, Sarah E. Marzen, Lucas Teixeira, Alexander Gietelink Oldenziel, and Paul M. Riechers. Transformers represent belief state geometry in their residual stream. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://papers.nips.cc/paper_files/paper/2024/hash/8936fa1691764912d9519e1b5673ea66-Abstract-Conference.html.

Max W. Shen, Ewa Nowara, Michael Maser, and Kyunghyun Cho. Training dynamics of learning 3D-rotational equivariance. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=DLOIAW18W3.

Aaditya K. Singh, Ted Moskovitz, Felix Hill, Stephanie C. Y. Chan, and Andrew M. Saxe. What needs to go right for an induction head? A mechanistic study of in-context learning circuits and their formation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 45637–45662. PMLR, 2024. URL https://proceedings.mlr.press/v235/singh24c.html.

Eric Todd, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. Function vectors in large language models. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/4ae163cb8788970e53b4fd9578141139-Abstract-Conference.html.

Soledad Villar, David W. Hogg, Kate Storey-Fisher, Weichi Yao, and Ben Blum-Smith. Scalars are universal: Equivariant machine learning, structured like classical physics. In *Advances in Neural Information Processing Systems*, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/f1b0775946bc0329b35b823b86eeb5f5-Abstract.html.

Maria-Luiza Vlădărean, Xuhui Zhang, and Suvrit Sra. On learning linear dynamical systems in context with attention layers. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/0edcce16f918256cd3992c8a6f4afd79-Abstract-Conference.html.

Johannes von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *Proceedings of the International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 35151–35174, 2023a. URL https://proceedings.mlr.press/v202/von-oswald23a.html.

Johannes von Oswald, Maximilian Schlegel, Alexander Meulemans, Seijin Kobayashi, Eyvind Niklasson, Nicolas Zucchet, Nino Scherrer, Nolan Miller, Mark Sandler, Blaise Agüera y Arcas, Max Vladymyrov, Razvan Pascanu, and João Sacramento. Uncovering mesa-optimization algorithms in Transformers. *arXiv preprint arXiv:2309.05858*, 2023b. URL https://arxiv.org/abs/2309.05858v2. Version 2, revised October 2024.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit Bayesian inference. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=RdJVFCHjUMI.

Liu Yang, Kangwook Lee, Robert Nowak, and Dimitris Papailiopoulos. Looped Transformers are better at learning learning algorithms. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/b8402301e7f06bdc97a31bfaa653dc32-Abstract-Conference.html.

Tianzhu Ye, Li Dong, Yuqing Xia, Yutao Sun, Yi Zhu, Gao Huang, and Furu Wei. Differential Transformer. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/00b67df24009747e8bbed4c2c6f9c825-Abstract-Conference.html.

Ruiqi Zhang, Spencer Frei, and Peter L. Bartlett. Trained Transformers learn linear models in-context. *Journal of Machine Learning Research*, 25(49):1–55, 2024. URL https://www.jmlr.org/papers/v25/23-1042.html.

Yufeng Zhang, Fengzhuo Zhang, Zhuoran Yang, and Zhaoran Wang. What and how does in-context learning learn? Bayesian model averaging, parameterization, and generalization. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 1684–1692. PMLR, 2025. URL https://proceedings.mlr.press/v258/zhang25d.html.

Chenyu Zheng, Wei Huang, Rongzhen Wang, Guoqiang Wu, Jun Zhu, and Chongxuan Li. On mesa-optimization in autoregressively trained Transformers: Emergence and capability. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/581e1a06fa20f2c079dc5fb2db236335-Abstract-Conference.html.

A Task Distribution and Sampling

This appendix specifies the sampling procedure that induces the distribution $\mathcal{P}$ over transition matrices B_τ used in Section 2.1. This OU-based parameterization guarantees stable dynamics and a tractable Gaussian stationary distribution. With the isotropic noise used in our discrete process, each sampled transition admits a unique stationary covariance determined by a discrete Lyapunov equation (Appendix A.4). The construction applies to physical dimension $d \geq 2$; the main experiments use $d = 16$ and $\Delta t = 0.005$. For each task, we independently draw a length scale ℓ_τ, an anisotropy a_τ, and a dimensionless strength parameter r_τ. The first two are log-uniform on $[0.5, 2]$ and $[1, 4]$, respectively, and r_τ is uniform on $[0, 1]$. Here, log-uniform on $[u, v]$ means that the logarithm is uniform on $[\log u, \log v]$.

We first construct a symmetric positive-definite matrix H_τ, which determines the confinement geometry, from ℓ_τ, a_τ, and a random orthogonal eigenbasis (Appendix A.1). We then sample a Gaussian

skew-symmetric matrix and normalize it using H_τ and r_τ to obtain $J_\tau^\top = -J_\tau$, which introduces nonreversible dynamics (Appendix A.2). The resulting drift and discrete transition matrices are

$$\Gamma_\tau = (I_d + J_\tau)H_\tau, \qquad B_\tau = \exp(-\Gamma_\tau \Delta t).$$

This drift parameterization follows the potential–circulation decomposition of linear stochastic dynamics (Kwon et al., 2005), also used by Lu & Tang (2026) to study nonreversible diffusion models. Given B_τ, trajectories follow the discrete process in Equation 1, whose construction is detailed in Appendix A.3.

A.1 Confinement Geometry

To construct H_τ, we choose relative eigenvalues equally spaced between 1 and a_τ^2, then apply a common scaling to set the length scale to ℓ_τ. Define

$$s_i = 1 + \frac{i-1}{d-1}(a_\tau^2 - 1), \qquad c_\tau = \frac{1}{d}\sum_{i=1}^{d} s_i^{-1}, \qquad i = 1, \dots, d. \tag{17}$$

Set $\lambda_i = c_\tau s_i / \ell_\tau^2$ and sample a uniformly random orthogonal eigenbasis P_τ, independently of the scalar parameters. Then define

$$H_\tau = P_\tau \operatorname{diag}(\lambda_1, \dots, \lambda_d) P_\tau^\top. \tag{18}$$

The eigenbasis is generated by QR factorization of a matrix with independent standard Gaussian entries. Flipping the sign of any column of P_τ leaves H_τ unchanged. By construction,

$$\sqrt{\frac{\operatorname{tr}(H_\tau^{-1})}{d}} = \ell_\tau, \qquad \sqrt{\frac{\lambda_{\max}(H_\tau)}{\lambda_{\min}(H_\tau)}} = a_\tau. \tag{19}$$

Thus, ℓ_τ is the root-mean-square scale per dimension, and a_τ is the ratio of the largest to smallest principal standard deviations, for the continuous OU stationary distribution with covariance H_τ^{-1}. The discrete process is described in Appendix A.3.

A.2 Nonreversible Dynamics

Independently of P_τ and the scalar parameters, sample a skew-symmetric matrix J_0 with independent standard Gaussian entries above the diagonal. Set

$$J_\tau = \frac{r_\tau \pi}{\Delta t \, \|H_\tau^{1/2} J_0 H_\tau^{1/2}\|_2} J_0, \tag{20}$$

where $\|\cdot\|_2$ is the matrix spectral norm. The denominator is nonzero almost surely. This normalization gives

$$\Delta t \, \|H_\tau^{1/2} J_\tau H_\tau^{1/2}\|_2 = r_\tau \pi. \tag{21}$$

To interpret this choice, in coordinates $y = H_\tau^{1/2} x$ the drift matrix becomes

$$H_\tau^{1/2} \Gamma_\tau H_\tau^{-1/2} = H_\tau + H_\tau^{1/2} J_\tau H_\tau^{1/2}.$$

The two terms on the right are its symmetric and skew-symmetric parts, respectively. Thus $r_\tau \pi$ sets the spectral norm of the skew-symmetric part multiplied by the time step; it is not generally the rotation angle of the full discrete transition. Since the normalization of J_τ depends on H_τ, the two matrices are generally statistically dependent.

A.3 From Continuous OU Dynamics to Discrete Trajectories

For a task $\tau = (H_\tau, J_\tau)$, the underlying continuous OU process is

$$\mathrm{d}x(s) = -\Gamma_\tau x(s)\,\mathrm{d}s + \sqrt{2}\,\mathrm{d}W(s), \qquad \Gamma_\tau = (I_d + J_\tau)H_\tau, \tag{22}$$

where $W(s)$ is standard Brownian motion in $\mathbb{R}^d$. The experiments instead use the discrete update

$$x_{t+1} = B_\tau x_t + \sqrt{2\Delta t}\,\varepsilon_t, \qquad \varepsilon_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d), \tag{23}$$

with $B_\tau = \exp(-\Gamma_\tau \Delta t)$ and initial state $x_1 \sim \mathcal{N}(0, I_d)$ independent of the task and noise sequence. This is Equation 1 with $\eta_t = \sqrt{2\Delta t}\,\varepsilon_t$. It can be constructed by operator splitting: for each time step of size Δt, compose the exact deterministic drift flow $x \mapsto B_\tau x$ with the pure diffusion update $x \mapsto x + \sqrt{2\Delta t}\,\varepsilon_t$. Each component flow uses step size Δt; their composition defines one step of the split scheme. We now compare this update with the exact transition of the continuous OU system in Equation 22.

Fix a task τ and let $s_t = (t-1)\Delta t$ denote the sampling times. Multiplying Equation 22 by the deterministic integrating factor $\exp(\Gamma_\tau s)$ gives

$$\mathrm{d}\big[\exp(\Gamma_\tau s)x(s)\big] = \sqrt{2}\exp(\Gamma_\tau s)\,\mathrm{d}W(s). \tag{24}$$

Integrating over $[s_t, s_{t+1}]$ and multiplying by $\exp(-\Gamma_\tau s_{t+1})$ yields the exact transition

$$x(s_{t+1}) = B_\tau x(s_t) + \xi_t, \qquad \xi_t = \sqrt{2}\int_{s_t}^{s_{t+1}} \exp\big[-\Gamma_\tau(s_{t+1}-v)\big]\,\mathrm{d}W(v), \tag{25}$$

where $B_\tau = \exp(-\Gamma_\tau \Delta t)$. Conditional on τ, the integrand is deterministic, so ξ_t is Gaussian with zero mean. Independent Brownian increments make these noise terms independent across intervals and independent of the state history up to s_t. The covariance follows from the Itô isometry, with $u = s_{t+1} - v$:

$$\Omega_\tau^{\mathrm{exact}} = \mathbb{E}[\xi_t \xi_t^\top \mid \tau] = 2\int_0^{\Delta t} \exp(-\Gamma_\tau u)\exp(-\Gamma_\tau^\top u)\,\mathrm{d}u. \tag{26}$$

Thus exact sampling of the continuous OU process uses $\xi_t \mid \tau \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Omega_\tau^{\mathrm{exact}})$.

For fixed Γ_τ, expanding the exact noise covariance gives

$$\Omega_\tau^{\mathrm{exact}} = 2\Delta t I_d - (\Gamma_\tau + \Gamma_\tau^\top)(\Delta t)^2 + O((\Delta t)^3). \tag{27}$$

The split scheme uses the leading term $2\Delta t I_d$ while retaining the exact deterministic transition B_τ. This fixed-drift expansion does not establish a small approximation error throughout the sampled task family: the normalization in Equation 20 itself depends on Δt. We use the isotropic covariance as part of the definition of the experimental discrete process. Its trajectories are not exact samples of the continuous OU process, and their stationary covariance generally differs from H_τ^{-1}, as derived below.

A.4 Stationary Covariance of the Split Scheme

For a fixed task, both the continuous process and the split scheme have unique centered Gaussian stationary distributions. To see stability, note that Γ_τ is similar to $H_\tau^{1/2}(I_d + J_\tau)H_\tau^{1/2}$, whose symmetric part is $H_\tau \succ 0$. Thus all eigenvalues of Γ_τ have positive real parts and B_τ has spectral radius less than one.

Setting the covariance time derivative to zero in the continuous OU process gives the continuous Lyapunov equation, with noise covariance rate $2I_d$. Since $\Gamma_\tau = (I_d + J_\tau)H_\tau$ and $J_\tau^\top = -J_\tau$, the matrix H_τ^{-1} satisfies this equation:

$$\Gamma_\tau H_\tau^{-1} + H_\tau^{-1}\Gamma_\tau^\top = (I_d + J_\tau) + (I_d - J_\tau) = 2I_d. \tag{28}$$

Stability ensures uniqueness, so the continuous stationary covariance is H_τ^{-1}. The split scheme instead has stationary covariance Σ_τ given by the discrete Lyapunov equation and its convergent series solution:

$$\Sigma_\tau = B_\tau \Sigma_\tau B_\tau^\top + 2\Delta t I_d, \qquad \Sigma_\tau = 2\Delta t \sum_{n=0}^{\infty} B_\tau^n (B_\tau^\top)^n. \tag{29}$$

Since exact OU sampling preserves H_τ^{-1}, $H_\tau^{-1} = B_\tau H_\tau^{-1} B_\tau^\top + \Omega_\tau^{\mathrm{exact}}$. Subtracting this identity from Equation 29 gives the exact correction

$$\Sigma_\tau = H_\tau^{-1} + E_\tau, \qquad E_\tau = \sum_{n=0}^{\infty} B_\tau^n (2\Delta t I_d - \Omega_\tau^{\mathrm{exact}})(B_\tau^\top)^n. \tag{30}$$

This correction describes the change in stationary covariance caused by replacing the exact OU noise with isotropic noise.

Its leading dependence on the step size can be made explicit while holding Γ_τ fixed. Define the invertible Lyapunov operator $\mathcal{G}_\tau(A) = \Gamma_\tau A + A\Gamma_\tau^\top$, so that $\exp(-\Delta t\mathcal{G}_\tau)A = B_\tau A B_\tau^\top$. Writing Id for the identity operator on matrices, the scalar series $z/(1-e^{-z}) = 1 + z/2 + z^2/12 + O(z^4)$ yields

$$\begin{aligned}\Sigma_\tau &= 2\Delta t\big(\mathrm{Id} - \exp(-\Delta t\mathcal{G}_\tau)\big)^{-1}(I_d) \\ &= H_\tau^{-1} + \Delta t I_d + \frac{(\Delta t)^2}{6}(\Gamma_\tau + \Gamma_\tau^\top) + O((\Delta t)^4). \end{aligned} \tag{31}$$

Here we used $2\mathcal{G}_\tau^{-1}(I_d) = H_\tau^{-1}$ from Equation 28. In particular, the leading covariance correction is $\Delta t I_d$. This is a fixed-drift expansion: the sampler in Equation 20 itself depends on Δt, so the expansion does not describe a limit in which tasks are resampled as Δt changes.

The experiments initialize $x_1 \sim \mathcal{N}(0, I_d)$ independently of the task, rather than sampling from each task's stationary distribution.

A.5 Orthogonal Invariance of the Sampler

For any fixed $R \in O(d)$, the random eigenbasis construction ensures $H_\tau \overset{\mathrm{d}}{=} RH_\tau R^\top$. The Gaussian law on skew-symmetric matrices is also invariant under $J_0 \mapsto RJ_0R^\top$: its density depends only on the Frobenius norm, which orthogonal conjugation preserves. Finally, the spectral norm in Equation 20 is invariant under simultaneous conjugation of H_τ and J_0. The entire sampling map therefore respects the transformation, so the joint law of (H_τ, J_τ) is invariant under simultaneous orthogonal conjugation. Since the matrix exponential satisfies $\exp(-R\Gamma_\tau R^\top \Delta t) = R\exp(-\Gamma_\tau \Delta t)R^\top$, the induced distribution $\mathcal{P}$ of B_τ satisfies Equation 4.

With isotropic initial states and independent isotropic noise, the discrete trajectory law transforms accordingly. This statement concerns the population distribution; a finite draw of training tasks need not be closed under rotations. In the diversity experiments, the scalar sampling ranges and trajectory-generation procedure remain fixed as the number of training tasks changes.

B Transformer Architecture Details

This appendix specifies the attention computation in Section 2.3. States are column vectors; a indexes heads, k indexes iterations, and i, j index query and key positions.

Linear maps and model configuration. Let d and D denote the physical and hidden dimensions, n_{head} the number of heads, and d_{head} the width of each head. The independently learned embedding and readout have shapes

$$W_E \in \mathbb{R}^{D\times d}, \qquad W_U \in \mathbb{R}^{d\times D}. \tag{32}$$

For each head a, the linear maps have matrix shapes

$$W_Q^a, W_K^a, W_V^a \in \mathbb{R}^{d_{\mathrm{head}}\times D}, \qquad W_O^a \in \mathbb{R}^{D\times d_{\mathrm{head}}}. \tag{33}$$

The main configuration uses $d = 16$, $D = 64$, $d_{\mathrm{head}} = D = 64$, $n_{\mathrm{head}} = 2$, and $K = 8$. Each head has the full hidden width. All attention-head linear maps are shared across the K iterations. The task count K_{task} is distinct from the iteration count K.

Embedding, readout, and Q/K/V weights are initialized with independent $\mathcal{N}(0, 0.02^2)$ entries. Output projections W_O^a use standard deviation $0.02/\sqrt{2n_{\mathrm{layer}}n_{\mathrm{head}}} = 0.01$, where $n_{\mathrm{layer}} = 1$ counts the shared block, not its iterations. Relative-position biases start at zero; linear maps have no biases and dropout is zero. Training enables bfloat16 autocast on supported GPUs (otherwise float32) and permits TF32 matrix multiplication.

Normalized causal attention. Queries and keys use RMS normalization without learned affine parameters:

$$\mathrm{RMSNorm}_\epsilon(v) = \frac{v}{\sqrt{\|v\|_2^2/d_{\mathrm{head}} + \epsilon}}, \qquad \epsilon > 0. \tag{34}$$

The implementation uses the default $\epsilon = \texttt{torch.finfo}(v.\texttt{dtype}).\texttt{eps}$; for float32 inputs this is $2^{-23} \approx 1.19209 \times 10^{-7}$. The value pathway is unnormalized: neither the input to W_V^a nor the resulting values $W_V^a h_j^k$ are normalized. At iteration $k = 0, \ldots, K-1$, the content score is the scaled dot product of the normalized queries and keys. Substituting Equation 34 and the routing forms in Equation 8 gives

$$\begin{aligned} s_{ij}^{a,k} &= \frac{\mathrm{RMSNorm}_\epsilon(W_Q^a h_i^k)^\top \mathrm{RMSNorm}_\epsilon(W_K^a h_j^k)}{\sqrt{d_{\text{head}}}} \\ &= \frac{(h_i^k)^\top M_{QK}^a h_j^k}{\sqrt{d_{\text{head}}}\sqrt{(h_i^k)^\top M_{QQ}^a h_i^k / d_{\text{head}} + \epsilon}\sqrt{(h_j^k)^\top M_{KK}^a h_j^k / d_{\text{head}} + \epsilon}}, \end{aligned} \tag{35}$$

The logit and causal attention weight are then

$$\ell_{ij}^{a,k} = s_{ij}^{a,k} + \mathbf{1}_{k=0} b^a(i-j), \tag{36}$$

$$\alpha_{ij}^{a,k} = \frac{\exp(\ell_{ij}^{a,k})}{\sum_{r=1}^{i} \exp(\ell_{ir}^{a,k})}, \qquad 1 \le j \le i. \tag{37}$$

Thus $s_{ij}^{a,k}$ excludes positional bias, while $\ell_{ij}^{a,k}$ includes it. The learned relative-position bias $b^a(i-j)$ is used only at $k = 0$; later iterations use content scores alone. The causal mask sets $\alpha_{ij}^{a,k} = 0$ for $j > i$.

The attention-weighted hidden state is

$$\mu_i^{a,k} = \sum_{j=1}^{i} \alpha_{ij}^{a,k} h_j^k. \tag{38}$$

Each head contributes $W_{OV}^a \mu_i^{a,k}$ to the residual update in Equation 6. Hidden states and attention weights are recomputed at every iteration, with $h_i^0 = W_E x_i$ and final prediction $W_U h_i^K$.

The model has no feed-forward sublayers, residual-stream normalization, or final normalization. No explicit physical $O(d)$ equivariance constraints are imposed on its parameters.

Scale dependence. We normalize queries and keys while leaving the value pathway unnormalized to preserve state magnitudes. Neglecting the RMSNorm stabilizer ϵ, uniform input rescaling by $\gamma > 0$ leaves attention weights unchanged, while values, residual updates, and the linear readout scale by γ. Thus $f_\theta(\gamma X) = \gamma f_\theta(X)$, provided the normalized queries and keys are nonzero. With finite ϵ, this relation is approximate when squared RMS magnitudes dominate ϵ.

C Experimental Protocols and Structural Analysis

This appendix documents the models, evaluation data, and analysis methods used for the spectral evidence in Section 3 and the structural validation in Section 4.2. The shared-basis protocol also supports the transfer and diversity experiments; additional module-wise projection results are reported below.

C.1 Models and Evaluation Data

Sections 3 and 4 use the same 25 models with $K_{\text{task}} = 2000$: five data seeds (2025–2029), each crossed with five training seeds (2025–2029). Each model is trained on 8000 trajectories, four per task, without orthogonal augmentation. We analyze the final checkpoint after 2000 training epochs. The architecture is that of Appendix B. Training minimizes the next-state MSE in Equation 3 using AdamW with learning rate 10^{-4}, batch size 100, weight decay 0.01, $(\beta_1, \beta_2) = (0.9, 0.999)$, and no learning-rate decay.

The model with data seed 2028 and training seed 2028 is used consistently for the reference-model matrix and spectral visualizations. Aggregate analyses include all 25 runs.

The ID evaluation set contains 500 unseen tasks from the training population, with four trajectories per task. It is shared by the five training seeds for a given data seed. The OOD evaluation set is shared by all runs and contains 2000 unseen tasks with one trajectory per task. It changes the sampling ranges in Appendix A to $\ell_\tau \in [16, 32]$, $a_\tau \in [4, 16]$, and $r_\tau \in [1, 2]$. Both evaluation sets contain 2000 trajectories of 128 states. Both ID and OOD evaluations use unseen tasks. The distinction is distributional: ID tasks follow the training parameter distribution, whereas OOD tasks are drawn from shifted parameter ranges.

Prediction metrics average over all 127 next-state positions and all 16 coordinates. Conditional-mean targets are computed by applying each task's transition to the observed, noisy current state, $B_\tau x_t$; they are not recursively generated noise-free rollouts.

C.2 Spectral and Shared-Structure Diagnostics

Eigenvalue clustering. For a $D \times D$ matrix with $D = md$, the eigenvalue clustering diagnostic fits m complex-valued centers with exactly d eigenvalues assigned to each center. Here $D = 64$ and $d = 16$, giving four clusters of 16. Let λ_i be the eigenvalues, $g(i)$ their fitted cluster assignments, and c_r the mean of the d eigenvalues assigned to cluster r. Then

$$\epsilon_{\mathrm{eig}} = \left(\frac{\sum_{i=1}^{D} |\lambda_i - c_{g(i)}|^2}{\sum_{i=1}^{D} |\lambda_i - \bar{\lambda}|^2} \right)^{1/2}, \qquad \bar{\lambda} = \frac{1}{D} \sum_{i=1}^{D} \lambda_i. \tag{39}$$

The fit alternates between a minimum-cost assignment for fixed centers, with d slots per center, and updates of the centers to their assigned means. Assignment costs are squared distances in the complex plane. We use 32 initializations: one farthest-point initialization and 31 random sets of m eigenvalues sampled without replacement, with random seed 2025. Each fit stops when consecutive cluster assignments are identical or after 100 iterations. We retain the smallest within-cluster sum of squares across the 32 initializations; this procedure does not guarantee a global optimum. Eigenvalues and clustering are computed in double precision. The cardinality constraint is part of the diagnostic, not an independently observed count. For numerically constant eigenvalue spectra, we set the relative clustering error to zero; a small error alone does not establish distinct cluster centers.

Singular-value steps. The singular-value step diagnostic groups the sorted values into $m = D/d$ consecutive groups of d and replaces each group by its mean. The grouping is fixed by the ordering and requires no iterative clustering. For descending singular values $s_1 \geq \cdots \geq s_D$, define $b(i) = \lceil i/d \rceil$ and $\mu_r = d^{-1} \sum_{i=(r-1)d+1}^{rd} s_i$. The centered relative step error is

$$\epsilon_{\mathrm{sv}} = \left(\frac{\sum_{i=1}^{D} (s_i - \mu_{b(i)})^2}{\sum_{i=1}^{D} (s_i - \bar{s})^2} \right)^{1/2}, \qquad \bar{s} = \frac{1}{D} \sum_{i=1}^{D} s_i. \tag{40}$$

For numerically flat singular-value spectra, we set the relative fit error to zero. A small error therefore indicates agreement with the prescribed grouping, but does not by itself establish distinct spectral steps.

Common-commutant diagnostics. To measure shared approximately commuting directions, we center and Frobenius-normalize the OV operators across heads:

$$W_0^a = W_{OV}^a - \frac{\mathrm{tr}(W_{OV}^a)}{D} I_D, \qquad \widehat{W}_{OV}^a = \frac{W_0^a}{\|W_0^a\|_F}. \tag{41}$$

Centering removes a scalar-identity component that contributes nothing to commutators; normalization removes each remaining component's scale without changing the exact common commutant. We define the commutator operator $\mathcal{A}(Y) = n_{\mathrm{head}}^{-1/2} (Y \widehat{W}_{OV}^a - \widehat{W}_{OV}^a Y)_a$. Its kernel is the common commutant, and small singular values identify approximately commuting matrix directions. Singular values use the Frobenius inner product on the domain and the sum of Frobenius inner products on the output tuple. To obtain a matrix representation, let vec stack the columns of a matrix. Then

$$\mathrm{vec}(Y \widehat{W}_{OV}^a - \widehat{W}_{OV}^a Y) = L_a \,\mathrm{vec}(Y), \qquad L_a = (\widehat{W}_{OV}^a)^\top \otimes I_D - I_D \otimes \widehat{W}_{OV}^a. \tag{42}$$

Stacking the vectorized outputs over heads gives

$$\begin{bmatrix} \mathrm{vec}(\mathcal{A}(Y)_1) \\ \vdots \\ \mathrm{vec}(\mathcal{A}(Y)_{n_{\mathrm{head}}}) \end{bmatrix} = A\,\mathrm{vec}(Y), \qquad A = \frac{1}{\sqrt{n_{\mathrm{head}}}} \begin{bmatrix} L_1 \\ \vdots \\ L_{n_{\mathrm{head}}} \end{bmatrix} \in \mathbb{R}^{n_{\mathrm{head}}D^2 \times D^2}. \tag{43}$$

For two heads and $D = 64$, A has size 8192×4096. Column vectorization preserves the specified inner products, so $\sigma_j(\mathcal{A}) = \sigma_j(A)$. Ordering them as $\sigma_1 \le \sigma_2 \le \cdots$, we define the boundary ratio and bottom-spectrum RMS by

$$r_{\mathrm{comm}} = \frac{\sigma_{d^2}(\mathcal{A})}{\sigma_{d^2+1}(\mathcal{A})}, \qquad q_{\mathrm{comm}} = \left(\frac{1}{d^2} \sum_{j=1}^{d^2} \sigma_j(\mathcal{A})^2 \right)^{1/2}. \tag{44}$$

A small q_{comm} indicates approximate commutation in the bottom d^2 directions, while a small r_{comm} indicates separation from the remaining spectrum. The RMS includes the trivial zero direction supplied by the identity. Appendix D.4 derives the reference dimension d^2 and explains why additional degeneracies can shift the gap beyond this boundary.

Numerical implementation. For $d = 16$, the metric calculations retain the smallest 260 singular values. For Figure 2(d), we extend the computation to the smallest 512 values for all 25 final models and the representative model at initialization. We obtain the singular values of A by computing the lowest eigenvalues of the symmetrized matrix $A^\top A$ in float64 and taking their square roots, clipping small negative eigenvalues to zero. OV products are formed in float32 before conversion to float64. Numerically scalar operators are excluded before normalization; both heads are retained in all 25 reference runs. We set $r_{\mathrm{comm}} = 1$ when its denominator is numerically zero.

Aggregate spectral measurements. Across the 50 OV head operators from the 25 reference models, the relative eigenvalue clustering error averages 0.0401 (range: 0.0162–0.0957), and the relative singular-value step error averages 0.0462 (range: 0.0358–0.0875). Across models, the common-commutant boundary ratio r_{comm} averages 0.114 (range: 0.091–0.300), and the bottom-spectrum RMS q_{comm} averages 0.00454.

C.3 Shared-Basis Fitting and Evaluation

Each model uses one fitted basis S, shared across all iterations, to align hidden states from paired inputs X and RX under the physical action in Equation 10. Held-out tasks and transformations test whether this action extends beyond the fitting pairs. Calibration and validation each use 32 trajectories from 32 distinct tasks, drawn from an independently generated dataset with the training-distribution parameter ranges in Appendix A. The two banks contain disjoint tasks and each uses four independently sampled orthogonal transformations. The same banks are used for all models.

Alignment objective. The fit uses outputs of all eight iterations, at one-based input positions $\mathcal{I} = \{2, 4, 8, 16, 32, 64, 96, 127\}$. For a bank $\mathcal{B}$ of sequences and a set $\mathcal{R}$ of orthogonal transformations, with $m = D/d$, the objective is

$$\mathcal{L}_{\mathrm{align}}(S) = \frac{1}{K|\mathcal{R}|} \sum_{k=1}^{K} \sum_{R \in \mathcal{R}} \frac{\sum_{X \in \mathcal{B}} \sum_{i \in \mathcal{I}} \left\| h_i^k(RX) - S(I_m \otimes R)S^{-1} h_i^k(X) \right\|_2^2}{\sum_{X \in \mathcal{B}} \sum_{i \in \mathcal{I}} \| h_i^k(X) \|_2^2}. \tag{45}$$

The sequence and position sums are taken before normalization; iterations and transformations receive equal weight. The objective uses only activation alignment, without whitening, a parameter-structure penalty, or prediction-error optimization.

Initialization and optimization. The initializer is the polar factor of a frame formed from W_E and products of up to four OV operators applied to W_E. Candidate blocks are normalized to unit Frobenius norm. Starting with the embedding block, we greedily select blocks with the largest squared Frobenius norm outside the column span of the selected blocks, until D/d blocks have been selected.

Optimization first uses an orthogonal parameterization $S = S_0 \exp(A - A^\top)$, with A initialized to zero and S_0 the initializer, for 800 Adam steps at learning rate 0.025. Invertible refinements start from the fitted orthogonal basis and use 600 steps at learning rate 0.008, with singular values constrained to $[c^{-1/2}, c^{1/2}]$ for $c \in \{1, 3, 10\}$, bounding the condition number by c. These refinements optimize S directly and, after each Adam update, reconstruct it from its SVD with singular values clipped to the stated interval. Each optimization retains its best calibration iterate, including the initial iterate. We select the basis with the lowest validation alignment error among the orthogonal fit and the three constrained refinements, and fix it for subsequent measurements. All computations use float64. The fit is not guaranteed to attain the global minimum of the alignment objective. Moreover, basis selection minimizes validation alignment error rather than maximizing structural energy or minimizing output differences under projection. Other admissible bases may yield higher target energy fractions or better preserve the original model's outputs, but these improvements need not occur together.

Held-out evaluation. Independent test evaluation samples 128 trajectories without replacement from each ID/OOD set in Appendix C.1, using a fixed random seed of 2718. It uses eight orthogonal transformations (four proper rotations and four with a reflection component) and all 127 input positions at all eight iterations. The reported RMS is the square root of the mean of the normalized squared errors, rather than the mean of per-call RMS values. No test activations enter fitting or selection. Across the 25 models, mean relative RMS alignment errors are 6.41% on ID tasks and 7.57% on OOD tasks.

C.4 Structural Decomposition and Projection

We decompose each $d \times d$ block G_{ij} of a transformed matrix G into three mutually Frobenius-orthogonal components:

$$\begin{aligned} \mathsf{T}_{ij} &= \frac{\mathrm{tr}(G_{ij})}{d} I_d, \\ \mathsf{V}_{ij} &= \mathrm{diag}(G_{ij}) - \mathsf{T}_{ij}, \\ \mathsf{B}_{ij} &= G_{ij} - \mathrm{diag}(G_{ij}). \end{aligned} \tag{46}$$

Here $\mathrm{diag}(G_{ij})$ retains the diagonal entries as a diagonal matrix: T is the scalar-identity component, V the traceless diagonal component, and B the off-diagonal component. T includes the trace of every block, including blocks between different representation copies. Collecting these blocks gives the Frobenius-orthogonal projection $\Pi_d(G)$ in Equation 13, including for the rectangular embedding and readout maps. Write $G = \Pi_d(G) + \Delta_G$. For nonzero G, the target energy fraction and relative structural residual satisfy

$$e_{\mathrm{target}}(G) = \frac{\|\Pi_d(G)\|_F^2}{\|G\|_F^2}, \qquad \varepsilon_{\mathrm{str}}(G) = \frac{\|\Delta_G\|_F}{\|G\|_F} = \sqrt{1 - e_{\mathrm{target}}(G)}. \tag{47}$$

For nonorthogonal S, these quantities use the selected coordinate metric and are not invariant native-coordinate energies.

Projection implementation. Projection retains T and maps it back to native coordinates:

$$\begin{aligned} W_{OV}^a &\mapsto S\Pi_d(S^{-1}W_{OV}^a S)S^{-1}, \\ M_{AB}^a &\mapsto S^{-\top}\Pi_d(S^\top M_{AB}^a S)S^{-1}, \qquad AB \in \{QQ, KK, QK\}, \\ W_E &\mapsto S\Pi_d(S^{-1}W_E), \qquad W_U \mapsto \Pi_d(W_U S)S^{-1}. \end{aligned} \tag{48}$$

For the full-width heads, the desired OV product is implemented by setting its value factor to the identity and its output factor to the projected operator. Routing forms are substituted directly into the normalized bilinear score, keeping the RMS stabilization constants unchanged. This defines a computational projection without requiring a common Q/K factorization at the original head width. All iterations of the shared block use the same replacement; the original model state is restored between conditions. These interventions keep the original physical dimension; Appendix E uses the extracted cores to construct predictors at new dimensions.

As an implementation check, retaining all three components $\mathsf{T} + \mathsf{V} + \mathsf{B}$ through the intervention pipeline reproduces the original predictions on four ID and four OOD trajectories from one reference model, with relative output errors of 4.99×10^{-15} and 6.86×10^{-15}, respectively, in float64.

Intervention conditions and results. Figure 6 shows the following sets of projected modules:

Panel and condition	Projected modules
(a) `M_T`	QQ, KK, and QK
(b) `U_T`	Readout
(c) `all_T`	Embedding, readout, OV, and QQ/KK/QK
(d) `E_T`	Embedding
(e) `OV_T`	OV
(f) `E_OV_T`	Embedding and OV

Figure 6 shows all conditions on the same 25 models. For each run and evaluation set, let $\widehat{Y}$ and $\widehat{Y}_{\mathrm{proj}}$ collect the original and projected model predictions over all trajectories and all 127 input positions. We compute

$$\delta_{\mathrm{out}} = \frac{\|\widehat{Y}_{\mathrm{proj}} - \widehat{Y}\|_F}{\|\widehat{Y}\|_F}, \qquad \delta_{\mathrm{MSE}} = \frac{\mathrm{MSE}_{\mathrm{proj}}}{\mathrm{MSE}_{\mathrm{original}}} - 1, \tag{49}$$

where both MSEs use the conditional-mean targets $B_\tau x_t$. These relative changes are averaged across runs and reported as percentages. Runs with the same data seed share training data; the reported means summarize the 25 runs.

Structural projection removes deviations from the predicted equivariant form, yet jointly projecting all modules largely preserves prediction performance. It changes outputs by 3.30% on ID tasks and 4.22% on OOD tasks, while reducing MSE by 1.79% and 1.73%, respectively. Routing-only projection also slightly improves MSE (by 0.80% on ID and 0.87% on OOD), while readout-only projection has small effects (-0.11% and $+0.08\%$, respectively).

The E and OV interventions highlight the importance of coordination. Projecting E or OV alone increases MSE by 12.33% and 11.46% on ID tasks, and by 9.04% and 9.13% on OOD tasks. Jointly projecting E and OV largely preserves prediction performance, with mean MSE increases of only 0.055% on ID and 0.167% on OOD and relative output changes of 2.21% and 3.02%, respectively. In every run on both evaluation sets, this joint intervention yields lower MSE than either intervention alone. Based on this observation, we hypothesize that deviations from the target structure in E and OV partially compensate for one another in the original model.

D Equivariant Representation: Conditions and Derivations

We prove Theorem 1 by first deriving the equivariant structure for a single OV operator. The same block argument then gives the embedding, readout, and routing structures in Section 4.1. We conclude with the resulting spectral degeneracies and common-commutant structure. Empirical diagnostics, basis fitting, and projection experiments are documented in Appendix C.

D.1 From a Shared Representation to the OV Structure

The representation assumption. Suppose the hidden space consists of m copies of the physical vector space $\mathbb{R}^d$. In coordinates adapted to these copies, a hidden state is $z = (z_1^\top, \dots, z_m^\top)^\top$, and a physical rotation or reflection $R \in O(d)$ acts on every vector in the same way:

$$(z_1, \dots, z_m) \longmapsto (Rz_1, \dots, Rz_m). \tag{50}$$

Thus the action on z is $I_m \otimes R$. The model's native hidden coordinates may mix these copies through an invertible matrix S. Writing $h = Sz$, the same action becomes

$$\rho(R) = S(I_m \otimes R)S^{-1}. \tag{51}$$

We assume one such representation across heads, positions, and iterations. The reference model has $m = 4$ and $d = 16$.

An OV operator must respect this action. Fix one head and write $W = W_{OV}^a$ to simplify notation. For this map to be equivariant, transforming a hidden vector before applying W must agree with transforming its output:

$$W\rho(R)h = \rho(R)Wh \quad \text{for every } h, \qquad \text{or equivalently} \qquad W\rho(R) = \rho(R)W. \tag{52}$$

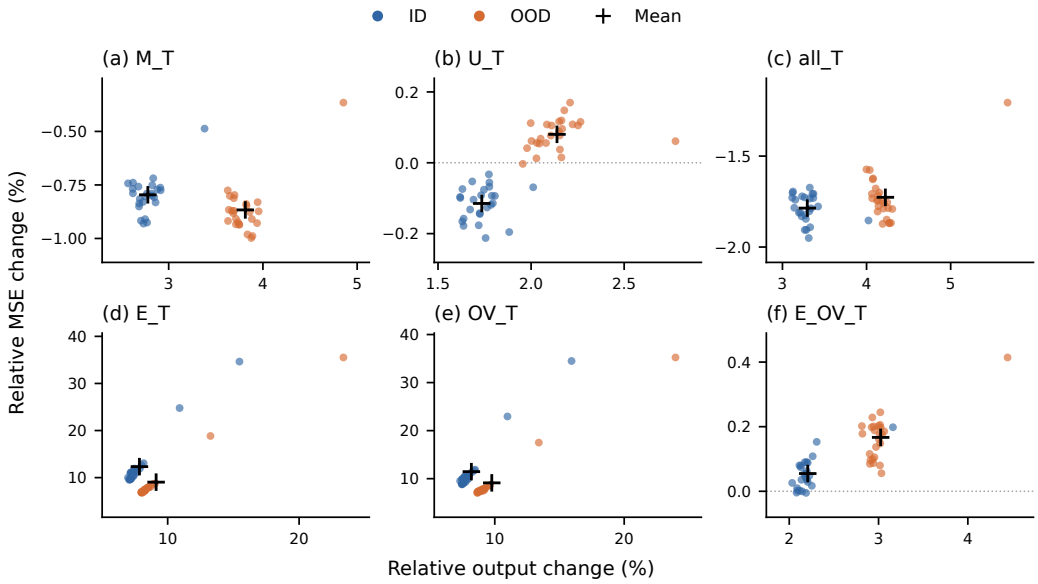

Figure 6: **Additional structural projection conditions.** Points show 25 runs under each intervention, colored by ID (blue) or OOD (orange); black plus signs mark the means. Panels use independent linear axis ranges. Negative MSE changes indicate improvement.

Let $\widetilde{W} = S^{-1}WS$ denote the operator in the copy coordinates. Substituting the expression for $\rho(R)$ and multiplying on the left by S^{-1} and on the right by S gives

$$\widetilde{W}(I_m \otimes R) = (I_m \otimes R)\widetilde{W}, \qquad R \in O(d). \tag{53}$$

Each block must be a scalar multiple of the identity. Write $\widetilde{W}$ as an $m \times m$ array of $d \times d$ blocks G_{rs}. The block G_{rs} maps the sth vector copy into the rth. Because $I_m \otimes R$ is block diagonal with the same R in every block, Equation 53 reads

$$G_{rs}R = RG_{rs} \qquad \text{for every } R \in O(d). \tag{54}$$

To see what this forces, consider any one block G. First choose a coordinate sign flip $F = \mathrm{diag}(\eta_1, \ldots, \eta_d)$, where $\eta_p \in \{-1, 1\}$. The (p, q) entry of $GF = FG$ gives

$$(\eta_q - \eta_p)G_{pq} = 0. \tag{55}$$

For $p \neq q$, we can choose opposite signs, so $G_{pq} = 0$. Thus G is diagonal. Next choose a permutation that exchanges any two coordinates. Commuting with this permutation forces the corresponding diagonal entries to be equal. Therefore $G = cI_d$.

Applying this argument to every block gives $G_{rs} = c_{rs}I_d$. Collecting these scalars into $C = (c_{rs}) \in \mathbb{R}^{m\times m}$, we obtain

$$\widetilde{W} = C \otimes I_d, \qquad W = S(C \otimes I_d)S^{-1}. \tag{56}$$

The meaning is simple: the output in copy r is $\sum_s c_{rs}z_s$. The operator can mix whole vectors, but it must use the same mixing coefficients at every physical coordinate. Conversely, this form commutes with $\rho(R)$ for every R. Since every head shares the same representation, the same S works for all OV operators, with a separate core C^a_{OV} for each head.

D.2 The Other Modules and the Full Predictor

Embedding and readout. The embedding maps a physical vector into the hidden representation, so its compatibility condition is $\rho(R)W_E = W_E R$. In adapted coordinates, $\widetilde{W}_E = S^{-1}W_E$ therefore satisfies

$$(I_m \otimes R)\widetilde{W}_E = \widetilde{W}_E R. \tag{57}$$

It has m block rows of size $d \times d$. Each block obeys the same commutation relation just proved for OV, and hence equals $e_r I_d$. Thus $\widetilde{W}_E = e \otimes I_d$ for $e \in \mathbb{R}^m$. Readout reverses the direction: $W_U \rho(R) = R W_U$. The identical argument applied to the block columns of $\widetilde{W}_U = W_U S$ gives $\widetilde{W}_U = u^\top \otimes I_d$, where $u \in \mathbb{R}^m$ is independent of e. In particular,

$$W_U W_E = (u^\top e) I_d. \tag{58}$$

Routing forms. Routing compares two hidden vectors through a bilinear form $h^\top M h'$. Under $h = Sz$ and $h' = Sz'$, this becomes $z^\top \widetilde{M} z'$ with $\widetilde{M} = S^\top M S$. Thus a form changes by congruence, whereas OV changes by similarity. For the bilinear value to be unchanged by a physical transformation, we require

$$\rho(R)^\top M \rho(R) = M. \tag{59}$$

In adapted coordinates this is

$$(I_m \otimes R)^\top \widetilde{M} (I_m \otimes R) = \widetilde{M}. \tag{60}$$

Each block G consequently satisfies $R^\top G R = G$. Since R is orthogonal, multiplying by R on the left gives $GR = RG$. The OV block argument now applies directly: $\widetilde{M} = C \otimes I_d$. Applying it to QQ, KK, and QK gives the three routing cores $C^a_{QQ}, C^a_{KK}, C^a_{QK}$. The QQ and KK cores are symmetric positive semidefinite because $\widetilde{M}^a_{QQ} = (W^a_Q S)^\top (W^a_Q S) \succeq 0$ and similarly for KK. Indeed, for any $v \in \mathbb{R}^m$ and unit $w \in \mathbb{R}^d$, $(v \otimes w)^\top (C \otimes I_d)(v \otimes w) = v^\top C v$. There is no symmetry requirement on the QK or OV cores.

Together, these results establish Equation 12. For the specified representation, these forms are equivalent to the respective module compatibility conditions in Equation 11.

Why the full predictor is equivariant. The embedding initializes states satisfying $h^0_i(RX) = \rho(R) h^0_i(X)$. Suppose this relation holds at iteration k. Invariance of QK preserves the numerator of the attention score; invariance of QQ and KK preserves its two RMS normalization factors, including the stabilization constant. Positional biases and the causal mask do not depend on physical coordinates, so the attention weights are unchanged. Their weighted sum of hidden states therefore transforms by $\rho(R)$, as does its OV update because each OV commutes with $\rho(R)$. Adding the residual preserves this transformation rule. Induction gives $h^K_i(RX) = \rho(R) h^K_i(X)$, and readout gives

$$f_\theta(R X_{\le i}) = W_U \rho(R) h^K_i(X) = R W_U h^K_i(X) = R f_\theta(X_{\le i}). \tag{61}$$

These are sufficient module-level conditions. The shared representation and module compatibility are the structural hypothesis tested in the paper; they are not inferred from output equivariance alone.

D.3 Spectral Degeneracies

For $W^a_{OV} = S(C^a_{OV} \otimes I_d) S^{-1}$, reordering the tensor-product basis converts $C^a_{OV} \otimes I_d$ into $I_d \otimes C^a_{OV}$, a block diagonal matrix with d identical copies of C^a_{OV}. Similarity preserves the characteristic polynomial, so

$$\det(\lambda I_D - W^a_{OV}) = \det(\lambda I_m - C^a_{OV})^d. \tag{62}$$

Each eigenvalue therefore has algebraic multiplicity divisible by d, including any additional multiplicity within C^a_{OV}.

For singular values, let $C^a_{OV} = U_C \Sigma_C V_C^\top$ be a singular-value decomposition. Then

$$C^a_{OV} \otimes I_d = (U_C \otimes I_d)(\Sigma_C \otimes I_d)(V_C^\top \otimes I_d). \tag{63}$$

The two outer factors are orthogonal, and each diagonal entry of Σ_C appears d times in the middle factor. Thus the singular values repeat d times in representation coordinates. This repetition also holds for the native W^a_{OV} when S is orthogonal, since orthogonal similarity preserves singular values; it does not follow for a general invertible S.

Repeated spectra alone do not identify a physical representation or a basis shared across heads. The corresponding empirical diagnostics are defined in Appendix C.2.

D.4 Common-Commutant Structure

Define the space of operators commuting with every OV matrix as

$$\mathcal{C} = \{Y \in \mathbb{R}^{D\times D} : YW_{OV}^a = W_{OV}^a Y \text{ for every } a\}. \tag{64}$$

The shared structure in Equation 12 implies

$$S\big(I_m \otimes \mathbb{R}^{d\times d}\big)S^{-1} \subseteq \mathcal{C}, \qquad \dim \mathcal{C} \geq d^2. \tag{65}$$

If the C_{OV}^a generate the full matrix algebra $\mathbb{R}^{m\times m}$, this inclusion is an equality. For the inclusion, any $A \in \mathbb{R}^{d\times d}$ satisfies

$$(I_m \otimes A)(C_{OV}^a \otimes I_d) = C_{OV}^a \otimes A = (C_{OV}^a \otimes I_d)(I_m \otimes A). \tag{66}$$

Conjugating by S gives a common commuting map, and the injective linear map $A \mapsto S(I_m \otimes A)S^{-1}$ supplies d^2 independent directions. For equality, let $\widetilde{Y} = S^{-1}YS$ commute with all $C_{OV}^a \otimes I_d$. It then commutes with every product and linear combination of these matrices. If the cores generate the full matrix algebra, it therefore commutes with every $F_{rs} \otimes I_d$, where $F_{rs} \in \mathbb{R}^{m\times m}$ is the matrix unit with a single nonzero entry at (r, s). Commuting with $F_{rr} \otimes I_d$ makes $\widetilde{Y}$ block diagonal; commuting with $F_{rs} \otimes I_d$ for $r \neq s$ makes all its diagonal blocks equal. Hence $\widetilde{Y} = I_m \otimes A$, proving equality and dimension d^2. The predicted minimal dimension is therefore 256 for $d = 16$; additional degeneracies can enlarge it. The heads themselves need not commute, since the C_{OV}^a may implement different, noncommuting operations on the vector copies.

Even an exact commutant dimension does not by itself identify the action of physical rotations on hidden states. The numerical test for approximate commutation and its spectral summaries are specified in Appendix C.2; the activation-based test of the physical action is described in Appendix C.3.

E Core Transfer: Construction and Evaluation

E.1 Source Models and Predictor Construction

We use all 25 reference models at $K_{\text{task}} = 2000$, $d = 16$, $D = d_{\text{head}} = 64$, and $K = 8$, after 2000 training epochs. These cross data seeds 2025–2029 with training seeds 2025–2029. Each source uses its own fitted basis obtained using the protocol in Appendix C.3. All sources are evaluated at every target dimension $d \in \{4, 8, 16, 32, 64, 128\}$, without selection by prediction performance.

Following the structural decomposition in Appendix C.4, we extract the coefficient of the scalar-identity component T from each block in the fitted basis: for a 16×16 block G_{ij}, this coefficient is $\mathrm{tr}(G_{ij})/16$. Embedding and readout yield $e, u \in \mathbb{R}^4$; the two OV maps and six routing forms yield eight 4×4 core matrices. These coefficients are identical at every target dimension. With a shared orthogonal $S_d \in O(4d)$, the target maps are

$$\begin{aligned} W_E &= S_d(e \otimes I_d), & W_U &= (u^\top \otimes I_d)S_d^\top, \\ W_{OV}^a &= S_d(C_{OV}^a \otimes I_d)S_d^\top, & M_{AB}^a &= S_d(C_{AB}^a \otimes I_d)S_d^\top, \end{aligned} \tag{67}$$

where $AB \in \{QK, QQ, KK\}$. Similarity and congruence take the same form here because the new basis is orthogonal, even though the original fitted basis need not be. The constructed predictors use these maps and the positional biases specified below, without further optimization. The hidden-to-physical dimension ratio is four, with two full-width heads, one shared block, and eight iterations. The native $d_{\text{head}} = 4d$ attention scaling and RMS normalization are unchanged; there is no additional content-temperature adjustment.

Attention scores are computed directly from the routing forms: M_{QK}^a determines the numerator, while M_{QQ}^a and M_{KK}^a determine the RMS normalization terms. This is the equivalent bilinear-form expression of the attention computation in Equation 35. Consistently changing the shared orthogonal basis S_d only changes hidden coordinates: attention scores remain unchanged, while hidden updates and readout transform compatibly, leaving predictions unchanged. Numerical checks comparing random and identity bases on four full trajectories per dimension give relative output discrepancies below 8×10^{-15}.

Test set	Head role	Mean target mass (%)	Range across runs (%)
ID	Previous	99.77	98.01–99.97
ID	Self	99.34	97.77–99.48
OOD	Previous	99.70	95.22–99.97
OOD	Self	99.41	99.03–99.48

Table 1: **Original-model first-iteration attention concentrates on the learned positional roles.** Each run averages 2000 trajectories and positions 2–127; means and ranges include all 25 source models. Head roles are assigned from learned biases.

E.2 Learned Positional Roles and Their Idealization

Positional roles in the source models. All 25 source models have one head's learned bias maximum at lag one and the other head's maximum at lag zero, allowing head permutation. We compute maxima over reachable lags 0–126, excluding the unused lag 127 in models trained with 127 input positions. These maxima identify positional preferences; the actual attention probabilities are examined below.

For transfer, we label the head whose learned bias peaks at lag one as the previous head ($a = 1$), and the head whose bias peaks at lag zero as the self head ($a = 2$). We reorder each head's OV, QK, QQ, and KK cores accordingly.

Actual first-iteration attention. We measure the original models' attention probabilities after softmax, including the learned content scores, Q/K normalization, and positional biases, without projection or intervention. Each of the 25 source models is evaluated on the ID and OOD sets of Appendix C.1, with 2000 trajectories per set and 127 input positions. We exclude position $i = 1$, where causality forces both heads to attend to the sole available token. For the remaining positions, the target attention mass is $\alpha^{a,0}_{i,i-1}$ for the previous head and $\alpha^{a,0}_{i,i}$ for the self head. We average over trajectories and positions $i = 2, \ldots, 127$ within each run, then report the mean and range across runs in Table 1.

Across all 25 source models and both test sets, each head assigns its highest attention weight to its designated position on every trajectory and at every noninitial input position. This confirms that the previous/self roles are reflected in the actual first-iteration attention patterns, not just in the learned positional biases.

Observations with other positional encodings. The self/previous division also appears in trained models with sinusoidal or learned additive positional embeddings (Figure 7). This first-iteration self/previous division recurs across multiple positional encoding schemes, model configurations, and training seeds, although some models do not exhibit a distinct pairing.

Positional rule for transfer. After aligning heads, the construction sets $b^1(1) = 100$ and $b^2(0) = 100$, with every other lag bias zero. These biases enter only at $k = 0$; the content score is still present. At the first input position, causality leaves only the self key available. The observed attention concentration motivates this fixed rule. The finite bias idealizes the learned attention pattern without imposing an exact hard attention mask. This rule also extends naturally to sequences longer than those used in training.

E.3 Evaluation and Complete Results

At each dimension, we evaluate on 2000 distinct test tasks, with one trajectory of length $T_{\mathrm{seq}} = 512$ per task. The datasets use seed 2025, the scalar parameter ranges and split OU discretization of Appendix A, and $\Delta t = 0.005$. At each of the 511 prediction positions, squared error against the conditional mean $B_\tau x_t$ is averaged over all physical coordinates and test trajectories. All-position MSE additionally averages these 511 values; final-position MSE uses only the last prediction position.

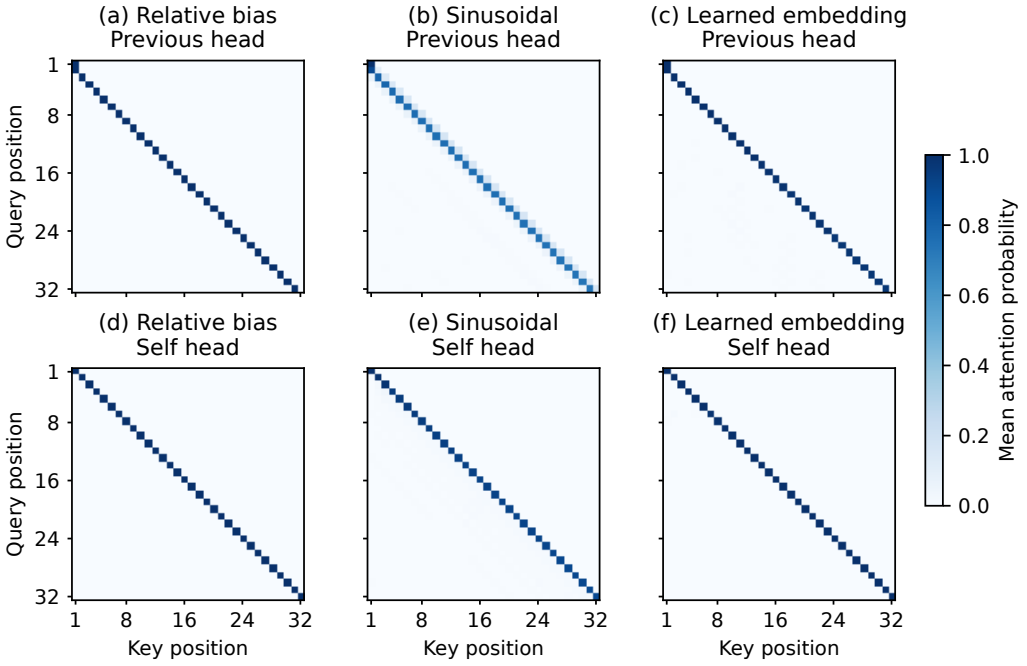

Figure 7: **First-iteration previous/self attention across positional encoding schemes.** Rows show the previous and self heads; columns show the reference model with both seeds 2028, a sinusoidal-embedding model, and a learned-embedding model. Each panel averages the first 32 test trajectories from its model's training dataset, displaying positions 1–32 from inference on all 127 input positions. All panels share a probability scale. The historical additive-embedding examples use training seed 2025, width 128, and eight iterations: the sinusoidal example uses Q/K normalization, while the learned example uses residual RMS normalization. Their datasets and normalization settings differ. These examples illustrate similar positional roles across different model configurations.

At position $i \geq 2$, prefix least squares predicts $\widehat{x}_{i+1} = \widehat{B}_i x_i$, where

$$\widehat{B}_i = \arg\min_{B \in \mathbb{R}^{d \times d}} \sum_{t=1}^{i-1} \|x_{t+1} - B x_t\|_2^2 + 0.01 \|B\|_F^2. \tag{68}$$

The implementation uses the pseudoinverse of the regularized Gram matrix. At $i = 1$ it uses persistence. Model and least-squares evaluations use float64 and float32, respectively. Baselines are shared across sources on the same test set at each dimension.

Table 2 reports all-position and final-position MSE, and Figure 8 shows the complete per-position curves across all six dimensions. Persistence and prefix least squares serve as a simple baseline and a strong task-specific baseline, respectively, contextualizing the prediction errors of the transferred cores. These results show that cores extracted at $d = 16$ support prediction across dimensions without changing their coefficients or retraining.

Benefit of additional context. To test whether history beyond the training window improves prediction, we compare the full input $(x_1, \ldots, x_{511})$ with its most recent 127 states $(x_{385}, \ldots, x_{511})$ at $d = 16$. Both conditions predict the same conditional mean $B_\tau x_{511}$ on the same 2000 test trajectories, for all 25 source models. The truncated input is processed from its new starting position through all eight iterations, using the same extracted cores, normalization, attention scaling, and fixed positional rule. No parameters are retrained.

Table 3 shows a 45.24% reduction in mean MSE with the full context, with improvements in 24 of 25 sources. The exception, with both data and training seeds 2025, increases MSE from 0.029542 to 0.036071 (22.10%). This source is already a high-error outlier among the original models on both

	All positions			Final position		
d	Core	LS	Persist.	Core	LS	Persist.
4	0.029064	0.006356	1.122772	0.018600	0.000091	1.123961
8	0.016989	0.007393	0.865470	0.007917	0.000181	0.927283
16	0.017975	0.011458	0.711381	0.005650	0.000375	0.754189
32	0.030684	0.020463	0.667783	0.011377	0.000860	0.748175
64	0.065079	0.039993	0.638062	0.031248	0.002114	0.699590
128	0.130781	0.078360	0.609043	0.076540	0.006540	0.667005

Table 2: **Conditional-mean MSE for core transfer.** Core is the constructed predictor, LS is causal prefix least squares with ridge 0.01, and Persist. predicts the current state. All-position MSE averages 511 predictions; final-position MSE uses only the last one. Each source error averages 2000 trajectories and all physical coordinates. Core entries additionally average all 25 source models; baselines are shared across sources.

d	Recent 127 MSE	Full 511 MSE	Reduction	Sources improved
16	0.010317	0.005650	45.24%	24/25

Table 3: **Additional history improves fixed-target prediction.** Each MSE averages physical coordinates and the same 2000 tasks, then all 25 source models, at the final prediction position. Reduction is one minus the ratio of the two aggregate MSEs, computed before rounding, rather than an average of per-source percentage reductions.

ID and OOD tasks, and its extracted cores have the highest all-position transfer MSE at every tested dimension. It remains included in all aggregates. These results support a predictive benefit from additional history at this dimension and prediction position.

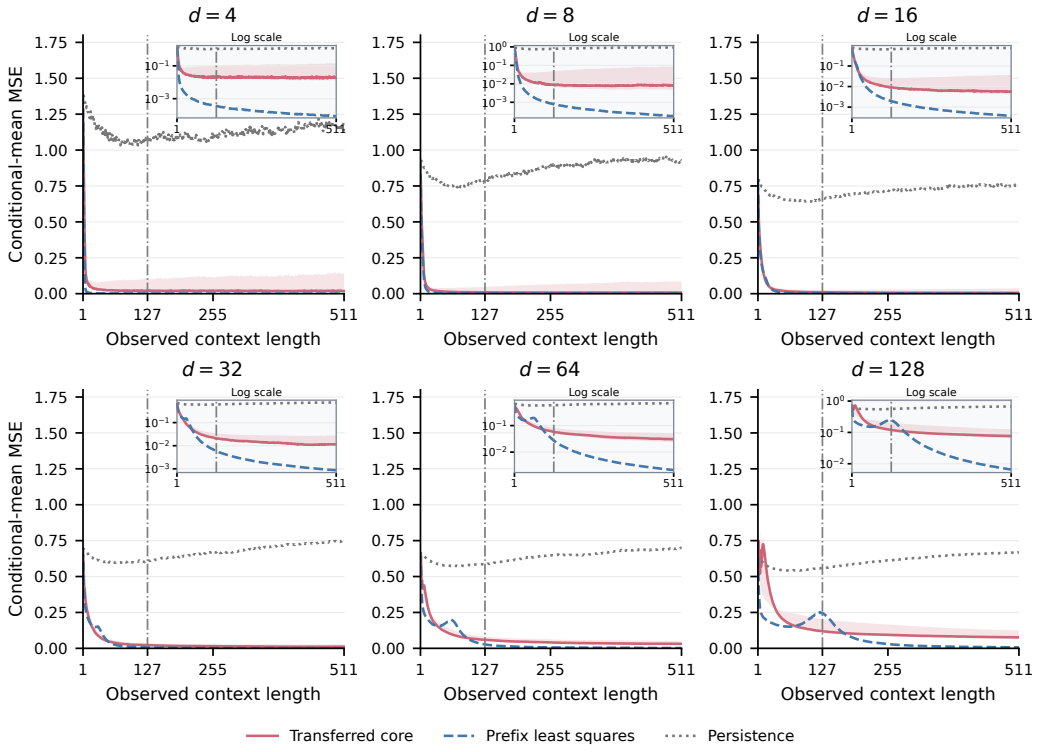

Figure 8: **Complete transfer prediction curves.** Per-position conditional-mean MSE at all six dimensions, with shared linear vertical limits and logarithmic-scale insets showing the same curves. Inset vertical limits are adjusted separately for each dimension. The vertical dash-dotted line marks 127 observed states, the maximum context during source training. Each source uses all 2000 test trajectories and keeps its own coefficients fixed across dimensions. The model line averages all 25 sources, with minimum–maximum shading including every source.

F Additional Diversity Results

F.1 Training and Evaluation Protocol

The sweep varies the base-task count $K_{\text{task}} \in \{4, 20, 100, 500, 2000\}$ and the presence of orthogonal augmentation at fixed dimension $d = 16$. Each condition crosses five data seeds with five model-training seeds (both 2025–2029), giving 250 models in total. All use the architecture in Section 2.3, 8000 training trajectories, and 2000 epochs with batch size 100, learning rate 10^{-4}, and weight decay 0.01, matching training examples and updates across conditions. Unless stated otherwise, analysis uses the final checkpoint and each run's basis selected by the protocol in Appendix C.3. Task sampling follows Appendix A.

Augmentation applies an independent Haar-distributed $R \in O(16)$ to all states of each training sequence, conjugating its dynamics by R. These rotations are sampled once at dataset construction and remain fixed during training. They preserve the base dynamics' orthogonal invariants and the number of trajectories, while adding coordinate representations. Thus K_{task} counts base tasks, not distinct oriented transition matrices. Orthogonal augmentation is applied only to the training set; paired augmented and unaugmented conditions share the same ID test trajectories.

Prediction MSE is computed against the conditional mean $B_\tau x_t$ and averaged over all 127 prediction positions and 16 coordinates. Each ID set contains 2000 trajectories drawn from $K_{\text{task}}/4$ distinct tasks, with task identities varying across task counts and data seeds. OOD evaluation instead uses a common set of 2000 tasks across all conditions, with jointly shifted task-parameter ranges relative to training.

The causal ridge least-squares baseline estimates a transition map using only observed transition pairs in the current prefix, with regularization parameter 0.01. At the first prediction position, the baseline uses the initial state as its next-state prediction, $\hat{x}_2 = x_1$.

F.2 Structural Diagnostics across Task Diversity

Table 4 supplements the OV measurements with the other analyzed modules. Fractions are measured in each run's fitted basis and averaged over runs, with equal weighting of the two heads where applicable.

K_{task}	Aug.	W_E	W_U	W_{OV}	M_{QQ}	M_{KK}	M_{QK}
4	No	34.97	50.52	16.28	46.73	49.96	22.05
4	Yes	98.83	99.79	96.75	91.37	88.92	93.55
20	No	35.28	76.93	66.55	70.57	52.39	43.55
20	Yes	99.01	99.90	98.55	93.00	89.03	94.63
100	No	77.40	96.63	92.63	83.49	75.53	77.17
100	Yes	99.48	99.90	99.60	92.99	91.80	93.63
500	No	98.96	99.53	99.27	90.44	88.40	91.00
500	Yes	99.48	99.90	99.61	92.53	91.02	93.57
2000	No	99.41	99.81	99.55	91.93	89.78	93.03
2000	Yes	99.53	99.90	99.63	92.34	90.46	93.95

Table 4: Mean target energy fractions (%) across the diversity sweep. Aug. indicates fixed per-sequence orthogonal augmentation. Each row averages 25 runs; head-specific quantities are also averaged over heads.

Figure 9 brings together final-checkpoint eigenvalue clustering, common-commutant separation, singular-value steps, and fitted OV target energy. The spectral metrics are defined in Appendix C.2. Singular-value step errors decrease with task diversity, with augmentation most helpful at low diversity. Unlike eigenvalue multiplicity, this repetition is an empirical signature, not implied by similarity to $C_{OV}^a \otimes I_{16}$ under a general invertible change of basis.

F.3 Prediction and Projection Effects

Figure 10(a,b) compares ID and OOD prediction across training conditions. Together with the structural diagnostics in Figure 9, it shows that augmentation promotes OV structural agreement, while

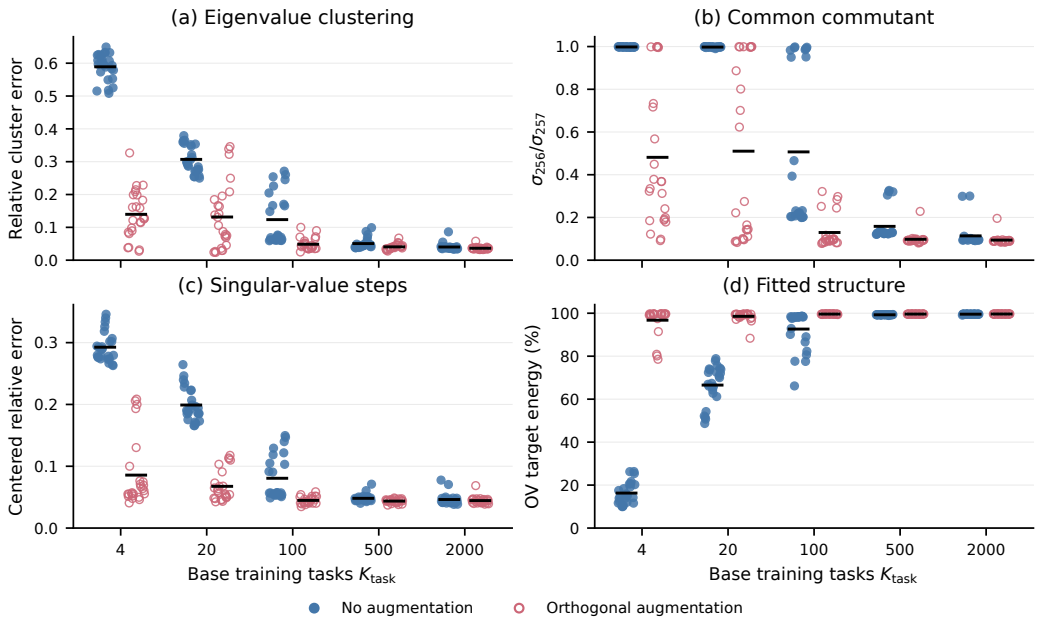

Figure 9: **Structural diagnostics across task diversity.** (a) OV eigenvalue clustering error; (b) common-commutant boundary ratio; (c) centered OV singular-value step error; (d) fitted OV target energy. Spectral metrics use native coordinates; target energy uses each run's fitted basis. Points are final-checkpoint runs, averaging heads in (a,c,d); bars are 25-run means. Colors follow the legend.

broader dynamical diversity remains important for OOD prediction. Figure 10(c,d) distinguishes the effect of joint structural projection on outputs from its effect on prediction error. At low unaugmented task counts, projection can improve prediction while substantially changing outputs; improved MSE alone therefore does not establish preservation of the original computation. At high task counts, both prediction and output changes are small.

F.4 Structure–Prediction Association and Seed Variation

Across the pooled runs, higher OV target energy generally accompanies lower OOD error, but near-unit target energy still spans a range of errors (Figure 11(a)). This association mixes task counts and augmentation conditions. Figure 11(b) fixes $K_{\text{task}} = 100$ to separate out task-count variation.

For the unaugmented $K_{\text{task}} = 100$ runs, we use 95% OV target energy as a descriptive separator chosen after observing a gap between two groups; it is not a theoretical threshold. All runs remain in the plots and aggregate statistics. Table 5 crosses the same five model-training seeds with each data seed. Both within-row and between-row variation indicate sensitivity to training data and model-training seed; this design does not provide 25 independent datasets or establish a sharp transition.

Data seed	Higher target fraction	Lower target fraction
2025	3	2
2026	3	2
2027	5	0
2028	5	0
2029	0	5

Table 5: Unaugmented $K_{\text{task}} = 100$ runs by training-data seed and observed OV target energy. The 95% separator lies within the observed gap.

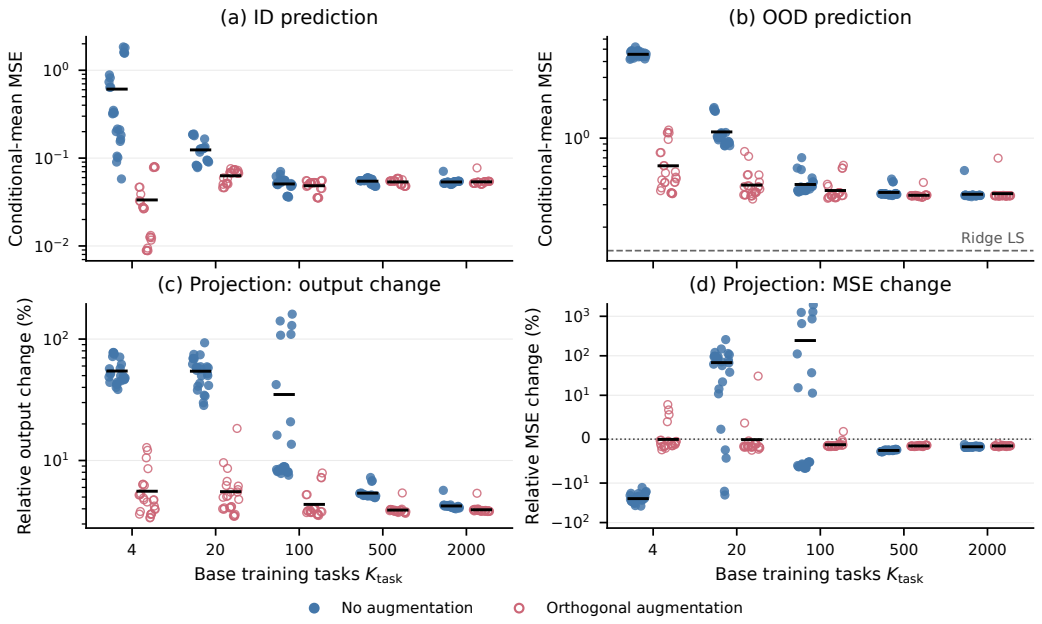

Figure 10: **Prediction and structural projection across task diversity.** Original-model MSE on (a) condition-specific ID sets and (b) the common OOD set; dashed line: causal ridge least squares. OOD projection effects: (c) relative output change and (d) relative MSE change (negative means improvement). Figure 10(a–c) uses log scales; (d) is linear within $\pm 10\%$ and logarithmic outside. Points are runs; bars are 25-run means. Evaluation follows Appendix F.1; colors follow the legend.

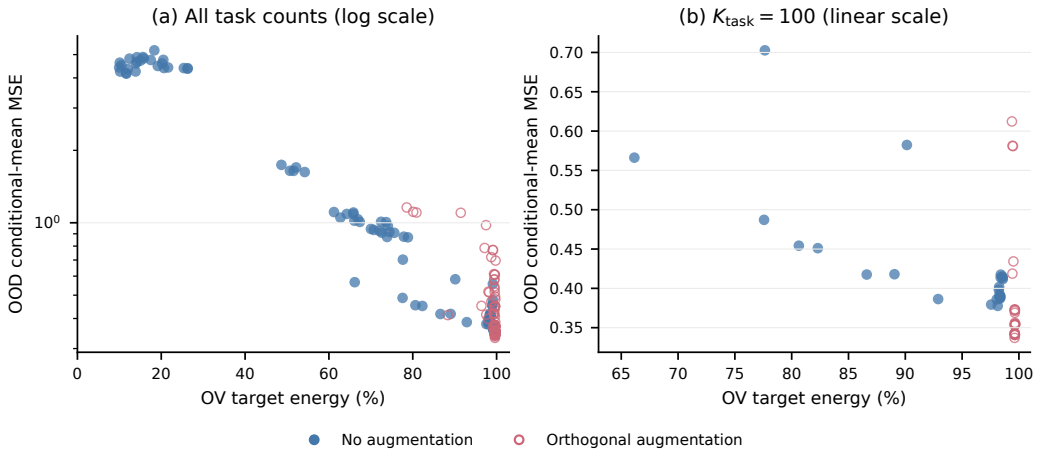

Figure 11: **Structure–prediction association across and within task counts.** OV target energy versus original-model conditional-mean OOD MSE for (a) all 250 runs and (b) all 50 $K_{\text{task}} = 100$ runs. The vertical axis is logarithmic in (a) and linear in (b); the horizontal range in (b) focuses on the fixed-task-count models. Both panels use the common 2000-task OOD set and all 127 positions. Each point is one run. Blue filled and red hollow markers denote training without and with orthogonal augmentation.

At $K_{\text{task}} = 100$, projection effects vary widely across unaugmented runs, with severe prediction degradation in some cases (Figure 10(c,d)). Figure 11(b) further shows that models with similar OV target energy can have different OOD prediction errors.

F.5 Development during Training

Figure 12 follows saved checkpoints for all five task counts, separating training without and with orthogonal augmentation. Each condition summarizes 25 runs by their median and interquartile range. Across task counts and augmentation conditions, eigenvalue clustering error typically decreases early in training, while the common-commutant boundary ratio remains near one. A later decrease in the latter is observed for unaugmented $K_{\text{task}} \geq 100$ and all augmented conditions.

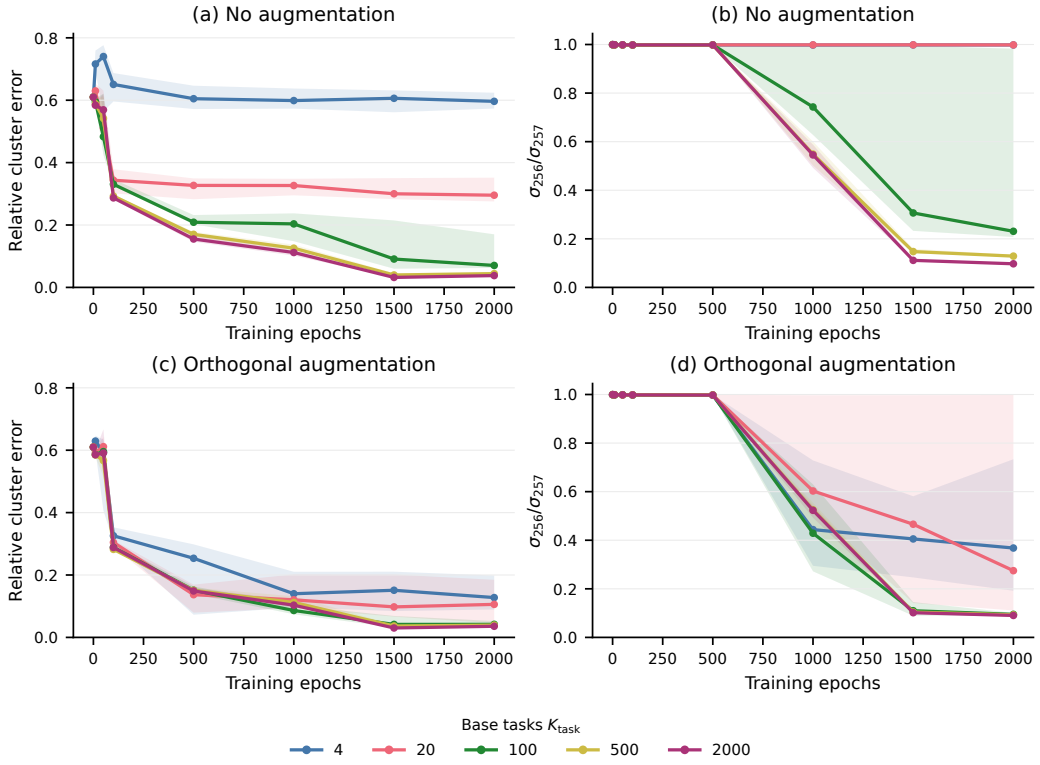

Figure 12: **Spectral structure during training.** Native-coordinate OV eigenvalue clustering error (left) and common-commutant boundary ratio (right), without augmentation in (a,b) and with orthogonal augmentation in (c,d). Colors distinguish the five base-task counts. Eigenvalue errors are averaged over heads before aggregation. Lines show medians and shaded bands show interquartile ranges over 25 runs per condition, not confidence intervals. Epoch zero denotes initialization; lines only connect sampled checkpoints.

Figure 13 shows module heatmaps for the representative unaugmented $K_{\text{task}} = 2000$ model (both seeds 2028). All eight checkpoints use its basis fitted at epoch 2000, with the transformations in Appendix C.4. The heatmaps thus track structure relative to the final representation, not the best alignment achievable by refitting at each stage.

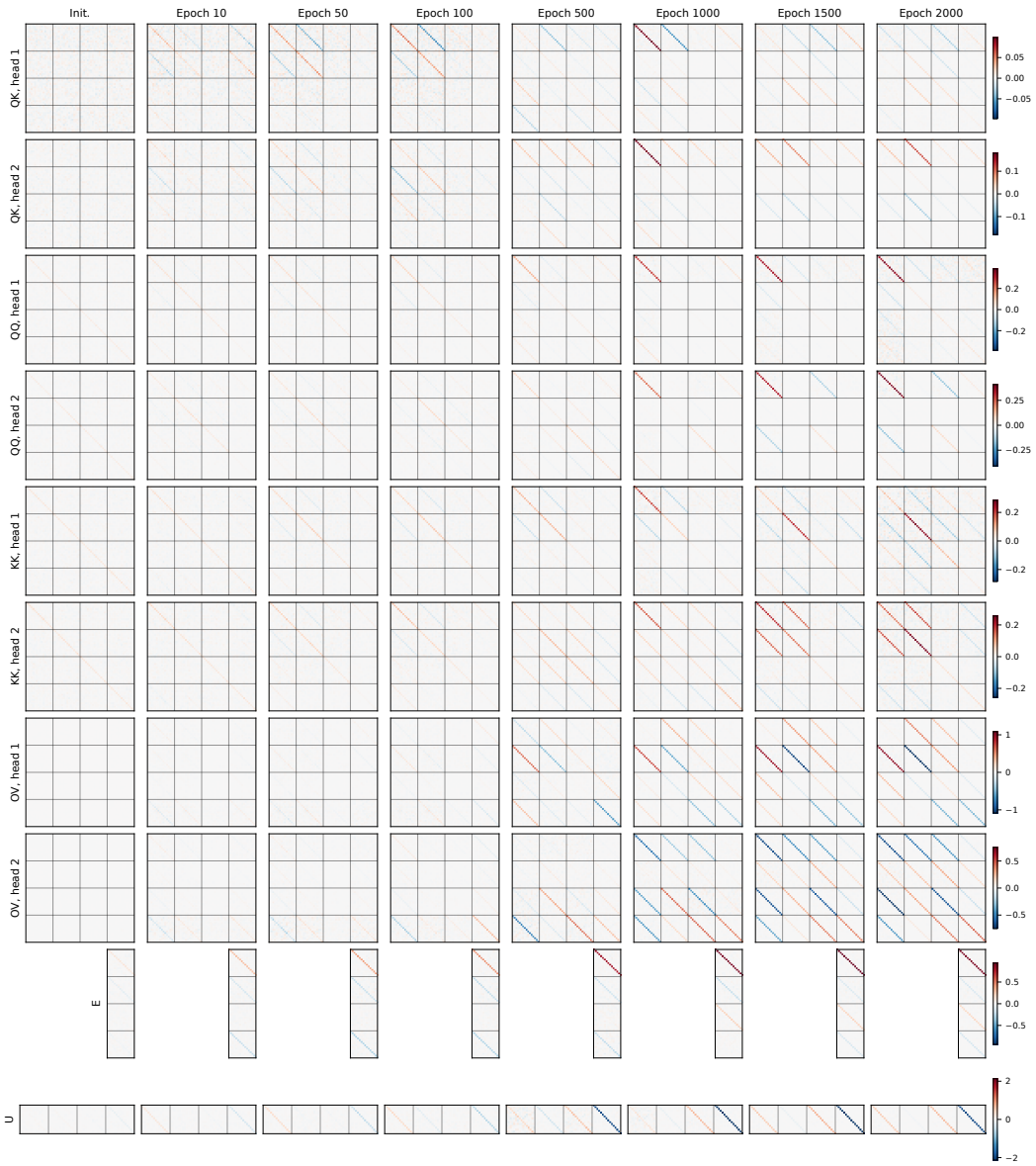

Figure 13: **Module structure during training in a fixed final basis.** Rows show $S^\top M^a_{QK} S$, $S^\top M^a_{QQ} S$, $S^\top M^a_{KK} S$, and $S^{-1} W^a_{OV} S$ for both heads, followed by $S^{-1} W_E$ and $W_U S$. Columns show initialization and 10, 50, 100, 500, 1000, 1500, and 2000 completed training epochs. Grid lines delimit 16×16 blocks. Each row uses a symmetric color scale fixed across checkpoints by its maximum absolute entry; scales differ between rows. All panels use the same basis fitted at epoch 2000.

G Functional Analysis of the Projected Models

G.1 Models, Evaluation, and Numerical Checks

The analysis in Section 5 uses all 25 unaugmented $K_{\text{task}} = 2000$ reference models, crossing five training-data seeds with five model-training seeds (2025–2029). We use their final checkpoints after 2000 epochs and the fitted bases of Appendix C.3. Embedding, readout, both OV operators, and all QQ, KK, and QK forms are jointly projected as in Section 4.2. The source dimension $d = 16$, head width 64, eight iterations, and learned first-iteration positional biases are unchanged. No parameters are retrained or fitted for these tests. The head names *self* and *previous* refer to first-iteration roles only: the assignment maximizes the sum of the corresponding lag-zero and lag-one learned biases over the two head permutations. It is fixed before evaluating interventions.

Both evaluation banks contain 128 distinct tasks with one trajectory per task and are shared across all models. The ID bank uses the training task distribution; the OOD bank uses the joint parameter shift of Appendix C.1. Tasks used in earlier circuit probes are excluded from both banks, and basis-calibration and validation tasks are also excluded from the ID bank. These are fresh subsets relative to those probes, not previously untouched datasets. Each forward pass uses all 127 input positions. The primary metric averages squared error against $B_\tau x_i$ over coordinates, positions 17–127, and tasks, after the eighth iteration. Absolute errors are therefore not directly comparable with the larger test sets and full-position averages in Section 4.

The explicit core implementation runs in CPU float64 with the original float32 RMSNorm epsilon unchanged. We compare predictions computed directly from the extracted cores with those computed by the native Transformer implementation using the jointly projected weights. This comparison uses four trajectories per model from each evaluation bank and covers every input position after each of the eight iterations; the maximum relative error is 4.42×10^{-15}. This check validates the implementation of the projected model, not equality to the original dense model. All reported results use existing experiment outputs.

Table 6 collects the results of all interventions: the median ratio of each model's MSE to its own projected baseline, the *ratio of mean MSEs*, and the number of models with increased MSE. The ratio of mean MSEs is computed by averaging intervention and baseline MSEs across models separately, then dividing the two averages. Figure 5 shows per-model ratios and their medians for selected interventions. The crossed seeds share training datasets and evaluation tasks, so these summaries describe model variation and are not confidence intervals from 25 independently sampled datasets.

G.2 Exact Input-Source Decomposition and State-Pair Matching

Write $Z_j^0 = e x_j^\top$ and $v_a = C_{OV}^a e$. Conditioning on the actual first-iteration attention gives

$$f_{c,j} = e + \sum_a \alpha_{jj}^{a,0} v_a, \qquad f_{p,j} = \sum_a \alpha_{j,j-1}^{a,0} v_a, \tag{69}$$

$$R_j = \sum_a \sum_{\ell<j-1} \alpha_{j\ell}^{a,0} v_a x_\ell^\top. \tag{70}$$

At $j = 1$, set $f_{p,1} = 0$ and $R_1 = 0$; the decomposition reduces to $Z_1^1 = f_{c,1} x_1^\top$. Thus Equation 14 is an exact accounting of input sources, including the residual connection. The coefficients depend on the input through attention; they are not fitted constants. At scored positions, first-iteration self and previous attention masses each exceed 99% when averaged over models and tasks. For each task, the remainder diagnostic is $\|R\|_F / \|Z^1\|_F$ over scored positions and all channel coordinates, then averaged over tasks and models. It equals 0.27%/0.36% on ID/OOD.

The pair-only control sets $Z^1 \leftarrow Z^1 - R$, preserving the natural coefficients. A separate hard-routing control replaces first-iteration attention by exact self and previous selection, with both selecting the first token at the boundary. Both rerun iterations two through eight; their small aggregate effects are reported in Table 6.

The second-iteration numerator is $N_{ij}^a = \mathrm{tr}[(Z_i^1)^\top C_{QK}^a Z_j^1]$. At each position j, current and previous refer to x_j and x_{j-1}, respectively. Substituting the three sources (current, previous, remainder) gives nine terms. The current–previous (CP) term is Equation 15; the current–current (CC) term

is $(f_{c,i}^\top C_{QK}^a f_{c,j})(x_i^\top x_j)$. Summing all nine terms recovers the directly computed numerator, with maximum relative error below 2.7×10^{-15}. Removing a term subtracts it from N_{ij}^a; keeping only CP replaces N_{ij}^a by that term. In either case the baseline QQ/KK normalization denominator is unchanged at this iteration. Causal masking, softmax, and all later updates are recomputed, while the other head is unchanged unless explicitly indicated. Thus "CP only" specifies one head's second-iteration numerator, not the entire predictor.

Table 6 reports these interventions for the self-role head, the previous-role head, and both heads together. In the self-role head, CP deletion consistently harms prediction, whereas CC deletion has little effect on most models. Keeping only CP preserves near-baseline performance in most models but produces substantial degradation in the data/training seed pairs (2025,2025) and (2027,2025), which remain in all summaries. The difference between the median ratio and the ratio of means highlights variation in the effects across models. Keeping only CP in both heads does not generally preserve prediction.

G.3 Two-Head Updates

We examine how the two attention heads jointly update the hidden state. Here i is the query position, $j \le i$ is a retrieved position, and $a \in \{1, 2\}$ indexes the heads. The weight $\alpha_{ij}^{a,k}$ is computed from the states Z^k and used in iteration $k+1$. At a fixed query position i and iteration $k+1$, let $\mu_a = \sum_{j \le i} \alpha_{ij}^{a,k} Z_j^k$ be the hidden-state mixture retrieved by head a. We suppress the dependence of μ_a on i, k. The total update $\Delta Z_i^k = Z_i^{k+1} - Z_i^k = C_{OV}^1 \mu_1 + C_{OV}^2 \mu_2$ can be written exactly as

$$\Delta Z_i^k = \underbrace{\tfrac{1}{2}(C_{OV}^1 + C_{OV}^2)(\mu_1 + \mu_2)}_{\text{common}} + \underbrace{\tfrac{1}{2}(C_{OV}^1 - C_{OV}^2)(\mu_1 - \mu_2)}_{\text{contrast}}. \tag{71}$$

The common term combines the sum of the OV maps with the sum of the retrieved mixtures; the contrast term combines their differences. These are two components of the joint update, not the individual heads.

This decomposition exposes a difference from a single standard softmax head. A single head applies $\beta_{ij} C$ to each retrieved state Z_j^k, where $\beta_{ij} \ge 0$ and the OV map C is shared across positions. These operators are nonnegative multiples of the same map: their magnitudes can vary with position, but their overall sign cannot reverse. By contrast, the contrast component is

$$\Delta Z_{i,\text{contrast}}^k = \sum_{j \le i} \frac{\alpha_{ij}^{1,k} - \alpha_{ij}^{2,k}}{2} (C_{OV}^1 - C_{OV}^2) Z_j^k. \tag{72}$$

When the two attention distributions differ, their difference sums to zero and contains both positive and negative entries. For $C_{OV}^1 \ne C_{OV}^2$, this applies opposite signs of the same OV difference to different positions. A single softmax head cannot directly reproduce this pattern of position-dependent operators on the same hidden states in one update. This comparison concerns the update operators, not the expressive power of a single-head model with different representations or multiple iterations.

Common-only replaces both heads' attention by $\bar{\alpha}_{ij}^k = (\alpha_{ij}^{1,k} + \alpha_{ij}^{2,k})/2$; contrast-only uses the signed weights $\alpha_{ij}^{a,k} - \bar{\alpha}_{ij}^k$. From iteration two onward, we recompute attention from the current state before each replacement and keep the residual connection. Common-only equivalently averages the two OV cores, as verified numerically.

Both interventions increase prediction error, with a much larger increase when the contrast component is removed (Table 6).

G.4 Updated-Memory Interventions

From iteration three ($k = 2$), key freezing replaces the key state by Z_j^1 in both the QK numerator and KK normalization; value freezing uses Z_j^1 only for values. Queries remain current, and joint freezing applies both replacements. All conditions keep the residual state and recompute subsequent updates. The degradation in Table 6 supports dependence on updated memory, without identifying a unique iterative estimation algorithm.

Intervention	Median ratio ID / OOD	Ratio of means ID / OOD	Worse / 25 ID / OOD
Exact self/previous routing	1.000 / 0.996	1.002 / 0.997	21 / 1
Remove first-iteration remainder	1.002 / 1.001	1.002 / 1.001	25 / 24
Remove CP: self-role	1.308 / 1.251	1.270 / 1.233	25 / 25
CP only: self-role	1.025 / 1.034	1.451 / 1.227	25 / 25
Remove CC: self-role	0.999 / 0.989	1.191 / 1.129	3 / 1
Remove CP: previous-role	1.046 / 1.128	1.076 / 1.136	25 / 25
CP only: previous-role	2.696 / 1.713	2.851 / 1.750	25 / 25
Remove CC: previous-role	1.118 / 0.990	1.172 / 1.016	25 / 4
Remove CP: both heads	4.414 / 1.927	4.033 / 1.855	25 / 25
CP only: both heads	2.865 / 1.768	2.798 / 1.748	25 / 25
Common only	27.357 / 13.242	27.250 / 13.286	25 / 25
Contrast only	2.558 / 3.131	3.995 / 3.468	25 / 25
Freeze keys	1.807 / 1.939	2.234 / 2.067	25 / 25
Freeze values	7.556 / 2.923	19.422 / 4.521	25 / 25
Freeze keys and values	4.945 / 3.435	7.283 / 4.251	25 / 25

Table 6: **Interventions and controls across all 25 models.** Each ID/OOD pair reports median per-model MSE ratio, ratio of mean MSEs, or number of models with increased MSE, relative to the projected baseline. The first two rows alter iteration one; CP/CC interventions alter iteration two; common/contrast interventions start at iteration two, and memory freezes start at iteration three. All subsequent computation is rerun.