SHARED ROUTER DYNAMICS IN SPARSE MOES: FROM CANONICAL GEOMETRY TO EXPERT CACHING

Anonymous authors
Paper under double-blind review

ABSTRACT

Sparse mixture-of-experts (MoE) models use an independently parameterized router at each sparse layer, yet their routing states exhibit substantial predictability across depth. We ask whether this predictability reflects reusable structure that is obscured by layer-specific coordinate systems. We isolate each router's control subspace and show that its factorization has an exact orthogonal gauge symmetry, which defines a principled quotient space for cross-layer comparison. We then fix this gauge globally with generalized orthogonal Procrustes analysis before fitting any transition model. Across four MoE architectures, a single shared transition recovers 79–90% of the held-out R^2 of independently fitted layer-specific dynamics, and the canonical forecasting advantage replicates on a second corpus across multiple architectures. Although generic residual representations are also predictable across depth, matched-rank comparisons show that canonical router coordinates preserve substantially more information about expert selection. Direct interventions nevertheless reveal a predictive–causal gap: accurate routing-state predictions need not preserve language-model loss when substituted for native routing. As a downstream application, we use canonical cross-layer forecasts to anticipate later expert demand while leaving native routing unchanged. In physical expert-offloading experiments, forecast-guided caching reduces expert transfers by up to 17.0% and improves token throughput by up to 5.7%, while preserving native routes and outputs. These results provide evidence for reusable routing geometry and dynamics across depth and show that the same structure can support MoE memory management without replacing the router itself.

1 INTRODUCTION

Sparse MoE models increase parameter capacity while activating only a subset of experts for each token. Their routers therefore play a central role in determining which computation is performed, how expert load is distributed, and which experts must be available at inference time. Existing work has studied load balancing, specialization, routing failures, and expert prefetching Shazeer et al. (2017); Fedus et al. (2022); Zhou et al. (2022); Muennighoff et al. (2025); Yoon et al. (2026). Most analyses, however, treat the router at layer ℓ as meaningful only inside that layer's own expert-index space.

Recent evidence suggests a more structured picture. Cross-layer prefetching methods show that future expert choices can be predicted from earlier routing signals Fang et al. (2026); Zhu et al. (2025); Zhao et al. (2026); mechanistic work identifies a routing-visible component of the residual stream and observes that it rotates across depth Ye et al. (2026); and router–expert parameters exhibit nontrivial within-layer geometry Ahrac et al. (2026). These results establish predictability and routing-relevant structure, but an observed cross-layer rotation does not by itself distinguish intrinsic dynamical change from arbitrary coordinate change inside the router-control subspace. Nor does generic representation alignment specify which transformations are functionally irrelevant to the router. We therefore ask whether the router factorization itself determines the nuisance symmetry, and whether quotienting out exactly that symmetry exposes reusable dynamics:

> What geometric structure underlies the observed predictability of routing across depth, and can routing-relevant states from different layers be aligned into a common representation?

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.

Answering this question requires more than applying a generic representation-alignment tool. Directly comparing router weights is not meaningful across depth because hidden representations evolve and expert identities are layer-specific. We instead isolate, at each layer, the subspace of hidden-state directions that can change relative router logits. Crucially, the resulting router-control factorization has an *exact* orthogonal gauge symmetry: rotations of its internal coordinates leave the router output unchanged. This identifies a specific nuisance degree of freedom and therefore a principled quotient space on which cross-layer comparison should be performed. Generalized orthogonal Procrustes analysis (GPA) then serves as a global gauge fixing that places the layer-wise control spaces into one canonical coordinate system.

This construction makes the shared-dynamics hypothesis falsifiable. If much of the apparent layer dependence is coordinate nuisance, then after gauge fixing a single low-capacity transition should recover most of the predictive power of separately fitted layer-specific transitions across many depths and architectures. If the rotations instead reflect genuinely unrelated dynamics, gauge fixing may increase representational similarity but should not close this transition-model gap. Our experiments support the former prediction. We then test its functional limits with causal substitutions; as a downstream application, we also use the canonical coordinates to forecast later expert demand for cache residency while native routing remains authoritative.

Our main contributions are:

- We show that the router-control factorization has an exact orthogonal gauge symmetry and formulate cross-layer comparison on the corresponding quotient space.
- Across Granite, OLMoE, Phi-tiny-MoE, and an IBM shared-expert MoE, one shared linear transition captures 79–90% of the R^2 of layer-specific dynamics after gauge fixing.
- Matched-rank, matched-readout probes distinguish generic residual smoothness from information that is specifically useful for expert selection.
- Causal interventions show reproducible local transport. A simple margin bound connects canonical-state error to exact top-k stability, and learned dynamics beats persistence at longer horizons on Phi and OLMoE.
- As a downstream application, we use same-token canonical forecasts for expert residency, reducing physical transfers and improving decoding throughput without changing native routing.

2 Related Work

MoE routing and specialization are well studied Shazeer et al. (2017); Fedus et al. (2022); Zhou et al. (2022); Muennighoff et al. (2025); Xue et al. (2024). Three nearby lines of work motivate our question but do not imply our conclusion. Cross-layer prediction and prefetching show that later expert demand is predictable Fang et al. (2026); Zhu et al. (2025); Zhao et al. (2026), but not that different depths share one routing-state dynamics. Routing-control analyses identify router-visible directions and observe that the control signal changes or rotates across depth Ye et al. (2026); without an explicit quotient construction, that observation does not separate basis-dependent rotation from intrinsic evolution, nor does it predict that one transition should approximate layer-specific dynamics after gauge fixing. Generic Procrustes/CKA alignment and layer-as-dynamics methods provide comparison tools Kornblith et al. (2019); Williams et al. (2021); Achara et al. (2026); Aswani et al. (2026), but do not identify the router-specific exact symmetry that determines which transformations are functionally irrelevant.

Our contribution is therefore not the observation that routing states rotate, nor the use of Procrustes by itself. The router-control factorization supplies an exact orthogonal gauge; quotienting out that gauge defines a canonical cross-layer routing space and yields the testable prediction that apparently layer-specific transitions become approximately shared. We evaluate that prediction across four architectures, with matched-rank, alignment, and causal controls. Further context appears in Appendix C.

3 METHOD

Figure 2 summarizes the pipeline: extract router-control states, fix their layer-wise gauges with GPA, model canonical evolution with one shared transition, and decode through the target layer's readout.

3.1 ROUTER-CONTROL FACTORIZATION

At layer ℓ, let E denote the number of routed experts and d the residual-stream width. The router weight matrix is $W_\ell \in \mathbb{R}^{E\times d}$, the token representation entering the router is $h_{\ell,t} \in \mathbb{R}^d$, and the router logits are $g_{\ell,t} = W_\ell h_{\ell,t} \in \mathbb{R}^E$. Let $\mathbf{1}_E = (1,\dots,1)^\top \in \mathbb{R}^E$ be the all-ones vector and I_E the $E \times E$ identity. Since softmax probabilities and top-k selection are invariant to a common shift of all E expert logits, we remove this common-logit direction by centering the router across experts:

$$\widetilde{W}_\ell = \left(I_E - \frac{1}{E}\mathbf{1}_E\mathbf{1}_E^\top\right) W_\ell = W_\ell - \mathbf{1}_E\bar{w}_\ell^\top, \tag{1}$$

where $\bar{w}_\ell = E^{-1}W_\ell^\top\mathbf{1}_E \in \mathbb{R}^d$ is the mean router row. Thus $\widetilde{W}_\ell$ is simply W_ℓ with the mean expert-weight vector subtracted from every row, and $\mathrm{rank}(\widetilde{W}_\ell) \le E-1$. Let

$$\widetilde{W}_\ell = A_\ell S_\ell U_\ell^\top = B_\ell U_\ell^\top, \qquad B_\ell \coloneqq A_\ell S_\ell, \tag{2}$$

be the compact SVD at rank r. The columns of $U_\ell \in \mathbb{R}^{d\times r}$ span exactly the residual-stream directions that can change *relative* router logits. We therefore define the router-control coordinates

$$x_{\ell,t} = U_\ell^\top h_{\ell,t} \in \mathbb{R}^r, \qquad \widetilde{g}_{\ell,t} = B_\ell x_{\ell,t}. \tag{3}$$

At full numerical rank, Eq. (3) reconstructs centered logits numerically and preserves native top-k decisions.

3.2 EXACT ORTHOGONAL GAUGE SYMMETRY

The basis U_ℓ is defined only up to an orthogonal change of coordinates. This freedom is not an artifact of the analysis, but an exact symmetry of the router factorization. Let

$$\mathrm{O}(r) = \{R \in \mathbb{R}^{r\times r} : R^\top R = I_r\}$$

denote the set of orthogonal transformations of the r-dimensional control space.

Proposition 1 (Router gauge invariance)**.** *For any $R \in \mathrm{O}(r)$, define*

$$U'_\ell = U_\ell R, \qquad B'_\ell = B_\ell R, \qquad x'_{\ell,t} = R^\top x_{\ell,t}. \tag{4}$$

Then

$$B'_\ell x'_{\ell,t} = B_\ell x_{\ell,t} \tag{5}$$

for every token t. Thus, an orthogonal change of basis changes the control coordinates but leaves the router output unchanged.

Therefore, the particular coordinates X_ℓ are not unique: any rotated representation $X_\ell R$ describes the same router-control state. We collect all such representations into the equivalence class

$$[X_\ell] = \{X_\ell R : R \in \mathrm{O}(r)\}. \tag{6}$$

Cross-layer comparison should therefore ignore this arbitrary choice of basis. A natural way to do so is the orthogonal Procrustes distance Williams et al. (2021),

$$d_{\mathrm{P}}([X],[Y]) = \min_{R\in\mathrm{O}(r)} \|XR - Y\|_F, \tag{7}$$

which first finds the best orthogonal alignment and then measures the remaining discrepancy. This optimization has the closed form

$$d_{\mathrm{P}}^2 = \|X\|_F^2 + \|Y\|_F^2 - 2\|X^\top Y\|_*. \tag{8}$$

Thus, Procrustes alignment resolves the orthogonal basis ambiguity inherent in the router-control factorization and enables meaningful cross-layer comparison.

3.3 A GLOBALLY CONSISTENT CANONICAL GAUGE

Pairwise Procrustes fits need not define a single coordinate system across depth. We therefore use generalized Procrustes analysis (GPA). On training tokens, after layer-wise centering and scalar normalization, we solve

$$\min_{\{Q_\ell\},M} \sum_{\ell=1}^{L} \|X_\ell Q_\ell - M\|_F^2, \qquad Q_\ell \in \mathrm{O}(r). \tag{9}$$

The canonical state of a column vector $x_{\ell,t}$ is then

$$z_{\ell,t} = Q_\ell^\top \,\mathrm{norm}(x_{\ell,t}). \tag{10}$$

For fixed aligned clouds $Z_\ell = X_\ell Q_\ell$, the optimal template is their mean $M^* = L^{-1}\sum_\ell Z_\ell$. Moreover,

$$\sum_\ell \|Z_\ell - M^*\|_F^2 = \frac{1}{L}\sum_{\ell<m} \|Z_\ell - Z_m\|_F^2, \tag{11}$$

so GPA can equivalently be viewed as minimizing total pairwise disagreement among layer gauges. All Q_ℓ are fitted on training sequences only. Crucially, Eq. (9) fixes only the layer-wise orthogonal gauges: it contains no transition model, no dynamics loss, and no parameter A. The shared-dynamics hypothesis is tested only after the gauges are frozen, by fitting the transition separately and evaluating its predictive performance on held-out sequences.

3.4 SHARED DYNAMICS AS GAUGE REDUCTION

Suppose raw control coordinates admit layer-specific affine transitions. Ignoring the fixed normalization for notational clarity,

$$x_{\ell+1} \approx F_\ell x_\ell + c_\ell, \qquad z_\ell = Q_\ell^\top x_\ell \implies x_\ell = Q_\ell z_\ell, \tag{12}$$

where the last relation follows from the orthogonality of Q_ℓ.

After the canonicalization in Eq. (10), the corresponding transition becomes

$$\begin{aligned} z_{\ell+1} &= Q_{\ell+1}^\top x_{\ell+1} && (13)\\ &\approx Q_{\ell+1}^\top (F_\ell x_\ell + c_\ell) && (14)\\ &= Q_{\ell+1}^\top F_\ell Q_\ell z_\ell + Q_{\ell+1}^\top c_\ell. && (15) \end{aligned}$$

Our central dynamical hypothesis is therefore precise: there exists a low-capacity operator A such that

$$Q_{\ell+1}^\top F_\ell Q_\ell \approx A \quad \text{for many layers } \ell, \tag{16}$$

or equivalently $F_\ell \approx Q_{\ell+1} A Q_\ell^\top$. Equation (16) can be viewed as a finite-depth analogue of orthogonal reducibility: the layer-dependent transitions become approximately constant in one gauge. Here the Q_ℓ are fixed by GPA on training states before fitting dynamics, rather than chosen to simplify the F_ℓ. In the exact case, h transitions telescope to $F_{\ell+h-1}\cdots F_\ell = Q_{\ell+h} A^h Q_\ell^\top$. If each canonical one-step defect is at most ε, the h-step operator error is bounded by $(\|A\|_2+\varepsilon)^h - \|A\|_2^h$. Appendix B proves this and gives a gauge-invariant singular-spectrum check. We use reducibility only in this finite-depth, data-driven sense.

Empirically, we fit

$$\hat{z}_{\ell+1,t} = A z_{\ell,t} + b \tag{17}$$

by ridge regression on pooled layer transitions, with regularization selected on validation data. Layer-specific affine models provide an upper bound. We summarize the degree of reuse with

$$\eta_{\mathrm{share}} = \frac{R^2_{\mathrm{shared}}}{R^2_{\text{layer-specific}}}, \tag{18}$$

which equals one when a single transition is as predictive as fitting each layer independently.

3.5 ROUTING STABILITY UNDER TRANSPORT

To evaluate a predicted canonical state, we first map it back to the expert logits of the target layer. Ignoring the fixed normalization transform, which can be absorbed into the decoder and a bias, we have $x_\ell = Q_\ell z_\ell$ and therefore

$$\widetilde{g}_{\ell,t} = B_\ell x_{\ell,t} = B_\ell Q_\ell z_{\ell,t} = D_\ell z_{\ell,t}, \qquad D_\ell := B_\ell Q_\ell. \tag{19}$$

Thus, D_ℓ is the layer-specific decoder from canonical routing coordinates to centered expert logits.

Now let $\hat{z} = z + e$ be a transported or predicted canonical state, where e is its state-space error. Its decoded logits satisfy

$$D_\ell \hat{z} = D_\ell z + D_\ell e, \tag{20}$$

so $D_\ell e$ is exactly the induced error in the centered logits. Let

$$\widetilde{g}_{(1)} \geq \widetilde{g}_{(2)} \geq \cdots \geq \widetilde{g}_{(E)}$$

denote the sorted true logits, and define the top-k routing margin

$$\gamma_k = \widetilde{g}_{(k)} - \widetilde{g}_{(k+1)}. \tag{21}$$

Proposition 2 (Top-k stability)**.** *If*

$$\|D_\ell e\|_\infty < \frac{\gamma_k}{2}, \tag{22}$$

then the selected expert set is unchanged:

$$\mathrm{TopK}(D_\ell \hat{z}) = \mathrm{TopK}(D_\ell z). \tag{23}$$

A sufficient condition directly in canonical state space is

$$\|e\|_2 < \frac{\gamma_k}{2\|D_\ell\|_{2\to\infty}}, \qquad \|D_\ell\|_{2\to\infty} = \max_{j\in[E]} \|d_{\ell,j}\|_2. \tag{24}$$

Proposition 2 separates geometric prediction error from routing error: stability depends on decoded perturbation relative to the local top-k margin. Proofs and the finite-horizon corollary appear in Appendix A.

Forecast-guided expert residency. As a systems application, canonical forecasts of later-layer expert demand for the same token guide eviction from a bounded GPU cache. Native routing and expert computation remain unchanged; Appendix E gives the forecast and cache policy.

4 EXPERIMENTAL SETUP

We evaluate four sparse MoE architectures spanning depth, expert count, routing sparsity, and explicit shared experts (Table 4).

Data and splits. WikiText-2 raw text is the common probe corpus. Geometry/dynamics use three sequence-level splits with 16,384 held-out tokens each; matched-readout and fixed-block causal tests use 50,176 held-out tokens per split. Alignment, dynamics, and readouts are fitted on training data and tuned on validation data. We additionally replicate the representation-only forecast experiment on CNN/DailyMail 3.0 for Granite, OLMoE, and Phi-tiny, with disjoint calibration, validation, and test articles (Appendix D.1).

Metrics. We evaluate geometry using held-out same-token cosine similarity and linear CKA. Dynamics are measured with R^2, normalized MSE, and cosine similarity. Routing fidelity is measured with top-1 accuracy, top-k recall and Jaccard, Jensen–Shannon divergence, and centered-logit MSE. We evaluate causal interventions using ΔNLL.

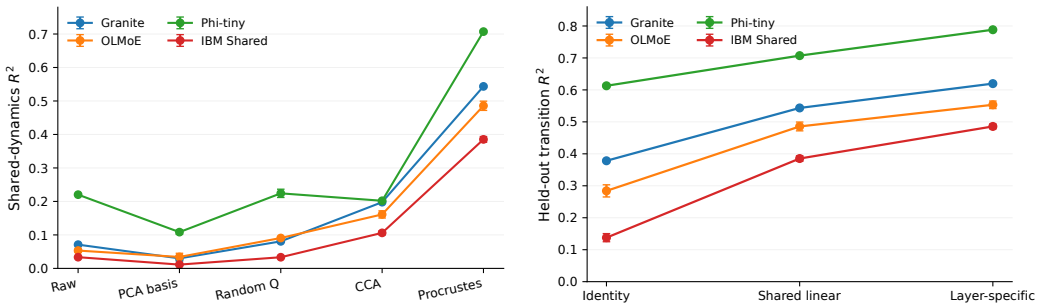
Physical expert-offloading protocol. Forecasts affect only cache eviction; native routing is unchanged. We report physical host-to-GPU loads/token and end-to-end throughput on 32 requests with 128 generated tokens and three request orderings. OLMoE uses capacities 32/64/128 and Granite 24/48/96; Appendix F gives the full protocol.

5 RESULTS

We establish the canonical geometry and routing specificity, probe causal transport, and finally use same-token forecasts for expert residency without changing native routing.

5.1 GAUGE ALIGNMENT EXPOSES REUSABLE DYNAMICS

A shared transition is weak in native coordinates and strong after gauge fixing. Across architectures, the pooled linear transition remains weak in raw coordinates and under generic alignment baselines, but becomes substantially more predictive after orthogonal Procrustes alignment (Fig. 1). Random orthogonal gauges and PCA-basis alignment do not produce the same improvement, and shuffling the correspondence between tokens removes the alignment effect almost entirely (full ablation in Table 6). Because GPA is fitted without access to the transition objective, improved geometric agreement does not by construction imply a reusable operator; the shared-transition gain is an additional held-out prediction tested after gauge fixing.



(a) Alignment baselines **(b) Shared vs. layer-specific dynamics**

Figure 1: **Gauge fixing exposes reusable cross-layer dynamics.** (a) Shared prediction remains weak under raw, PCA-basis, random-gauge, and CCA coordinates, then rises sharply after orthogonal Procrustes alignment. (b) One learned transition closes most of the gap from persistence to separately fitted layer-specific dynamics. Error bars show standard deviation across three split seeds.

One shared transition explains most of the predictable cross-layer change. We next compare the shared transition with a stronger baseline that fits a separate transition for every pair of adjacent layers. As shown in Table 1, the shared model reaches 79–90% of the R^2 obtained by these layer-specific models across all four architectures, while using far fewer parameters. This shows that the predictable evolution of router-control states is not strongly specific to individual layers. After alignment, much of this evolution can instead be captured by the same transition across depth.

Table 1: **Cross-model geometry and dynamics.** Values are means over three split seeds. "Id." is canonical-state persistence, "Shared" is one pooled linear transition, and "Layer" fits a separate linear transition per depth. $\eta_{\text{share}} = R^2_{\text{Shared}}/R^2_{\text{Layer}}$. Params are shared/layer-specific transition parameters.

Model	GPA cos.	Id. R^2	Shared R^2	Layer R^2	η_{share}	Params S/L
Granite	0.581	0.378	**0.544**	0.620	0.88	0.99k / 22.8k
OLMoE	0.548	0.284	**0.486**	0.553	0.88	4.03k / 60.5k
Phi-tiny	0.534	0.613	**0.707**	0.788	0.90	0.24k / 7.44k
IBM Shared	0.431	0.138	**0.385**	0.486	0.79	3.78k / 147.5k

The cross-layer advantage persists across corpus and architecture. On CNN/DailyMail, the canonical forecast advantage replicates on Granite, OLMoE, and Phi-tiny: across representative nonterminal horizons it consistently exceeds persistence and substantially outperforms raw, shuffled, and shared-logit controls (Appendix D.1). The models are refitted on the news corpus, so this is a cross-corpus replication rather than zero-shot transfer.

5.2 RESIDUAL SMOOTHNESS IS NOT THE SAME AS ROUTING SPECIFICITY

A possible alternative explanation is that canonical router states are predictable simply because hidden representations change smoothly across nearby layers. To test this, we compare them with matched-rank residual-PCA representations. For each representation, we train a decoder with the same parameter budget to reconstruct the true router logits on held-out tokens.

Table 2 shows a clear difference. Residual PCA is much easier to predict across layers, while router-control states recover the actual expert choices much more accurately. This shows that cross-layer predictability and routing relevance are different properties: residual representations evolve more smoothly, but router-control states retain the information that is directly used for expert selection.

Table 2: **Matched-rank, equal-budget readout comparison.** Means over three 50,176-token test splits. Higher is better for both metrics.

	Next-state R^2		Top-k recall	
Model	Router	PCA	Router	PCA
Granite	0.542	**0.855**	**0.999**	0.798
OLMoE	0.500	**0.788**	**0.998**	0.669
Phi-tiny	0.424	**0.861**	**0.989**	0.405
IBM Shared	0.371	**0.806**	**0.998**	0.589

5.3 A LOW-DIMENSIONAL SHARED CORE COEXISTS WITH ROUTING DETAIL

The rank sweep shows that the most predictable representation is not necessarily the one that best preserves routing behavior. OLMoE provides the clearest example (Table 8): very low-rank states are highly predictable, while higher ranks are needed to recover expert choices accurately and to improve causal transport. This suggests that a small shared component captures much of the cross-layer dynamics, while additional dimensions preserve finer routing information.

In centered coordinates, the unregularized least-squares shared operator is $A^* = C_{10}C_{00}^{\dagger}$, where $C_{10} = \mathbb{E}[z_{\ell+1}z_\ell^\top]$ and $C_{00} = \mathbb{E}[z_\ell z_\ell^\top]$. The rank sweep therefore measures how many control dimensions are needed to preserve predictable cross-layer structure and how many are needed for accurate routing decisions. The IBM shared-expert model is less compressible: both predictability and routing fidelity continue to improve as rank increases (Fig. 4).

5.4 CAUSAL TRANSPORT TESTS FUNCTIONAL RELEVANCE

Good prediction alone does not show that the aligned states are actually important for model behavior. We therefore replace the native routing states at selected layers with transported states and measure the resulting change in ΔNLL.

Canonicalization is important for causal transport. For Granite layers 9–13, the shared transition causes much less degradation in canonical coordinates than in raw coordinates, with shuffled gauges, or in a random matched-rank subspace (Table 9). Results are averaged over three 50,176-token splits.

Learned dynamics outperforms persistence across replications. Table 10 shows consistent gains for OLMoE layers 8–9 and both long Phi horizons. Degradation increases with transport length, while the advantage over persistence becomes more pronounced (Appendix D.1). The substantial Phi degradation suggests systematic router dynamics, but not a practical router-skipping method.

The IBM shared-expert model provides an important counterexample to over-generalization. It still exhibits canonical geometry and shared predictability, and selected short blocks can be transported without measurable degradation under fixed-block replication, but learned A is not uniformly better than identity across horizons. The benefit of explicitly evolving the canonical state is therefore architecture- and horizon-dependent.

Table 3: **Physical expert offloading with different forecast horizons.** Lower loads/token (↓) and higher throughput (↑) are better. Values are mean ± 95% Student-t interval half-width over three request orderings ($n = 3$). Local-4 is reported in Appendix Table 11.

Method	Small cache		Medium cache		Large cache	
	Loads ↓	Tok/s ↑	Loads ↓	Tok/s ↑	Loads ↓	Tok/s ↑
OLMoE	32 slots		64 slots		128 slots	
LRU	128.00 ± 0.00	4.246 ± 0.009	128.00 ± 0.00	4.237 ± 0.010	78.55 ± 0.04	5.424 ± 0.011
Pairwise	124.25 ± 0.16	4.231 ± 0.010	118.20 ± 0.16	4.333 ± 0.012	78.50 ± 0.04	5.255 ± 0.011
Procrustes (local-2)	128.00 ± 0.00	4.070 ± 0.009	128.00 ± 0.00	4.046 ± 0.009	78.51 ± 0.04	5.224 ± 0.011
Procrustes (1×)	$\mathbf{122.40 \pm 0.19}$	$\mathbf{4.268 \pm 0.014}$	$\mathbf{117.17 \pm 0.17}$	$\mathbf{4.372 \pm 0.017}$	$\mathbf{78.45 \pm 0.04}$	5.434 ± 0.010
Procrustes (2×)	123.44 ± 0.21	4.263 ± 0.014	117.56 ± 0.22	4.369 ± 0.016	$\mathbf{78.45 \pm 0.04}$	$\mathbf{5.439 \pm 0.019}$
Granite	24 slots		48 slots		96 slots	
LRU	192.00 ± 0.00	4.353 ± 0.009	192.00 ± 0.00	4.371 ± 0.009	192.00 ± 0.00	4.369 ± 0.009
Pairwise	187.46 ± 0.22	4.143 ± 0.015	182.61 ± 0.30	4.318 ± 0.016	166.97 ± 0.30	4.468 ± 0.017
Procrustes (local-2)	192.00 ± 0.00	4.296 ± 0.016	192.00 ± 0.00	4.261 ± 0.015	192.00 ± 0.00	4.264 ± 0.015
Procrustes (1×)	$\mathbf{184.62 \pm 0.17}$	$\mathbf{4.360 \pm 0.009}$	$\mathbf{177.17 \pm 0.21}$	$\mathbf{4.429 \pm 0.009}$	$\mathbf{159.37 \pm 0.23}$	$\mathbf{4.619 \pm 0.010}$
Procrustes (2×)	185.38 ± 0.16	4.357 ± 0.009	179.09 ± 0.19	4.409 ± 0.009	161.91 ± 0.22	4.605 ± 0.010

Notes. Intervals summarize variation across three request orderings. Bold marks the best numerical value in each model/cache setting (including ties). Full protocol details are in Appendix F.

5.5 Negative results define the limits

The negative results show where the shared structure stops being useful. First, directly averaging or sharing router weights strongly degrades model quality, meaning that aligned control states do not make expert identities or router readouts interchangeable across layers. Second, repeatedly skipping routers causes errors to accumulate quickly. We therefore limit our claim to *local functional transport and reusable cross-layer dynamics*.

5.6 Cross-layer forecasts reduce physical expert transfers

We next ask whether forecasts improve expert caching without changing routing. Experts are offloaded to host memory and a bounded set is resident on GPU; misses copy the exact native-selected expert. Forecasts affect only eviction priority. We report physical loads/token and end-to-end tokens/s, including forecast and copy costs.

Local-2 predicts one or two layers ahead every two sparse layers; 1× predicts all later layers once from the first sparse layer; 2× refreshes once at mid-depth. Local-4 is in Appendix Table 11; all forecasts expire within the current token.

The horizon comparison is striking. Short-horizon Procrustes forecasts do not improve residency at the smaller capacities: Local-2 has exactly the same mean loads as LRU for the small and medium caches on both OLMoE (128.00/128.00) and Granite (192.00/192.00), while also paying forecast overhead. Local-4 shows the same qualitative behavior (Appendix Table 11). In contrast, the long-horizon 1× schedule reduces OLMoE loads by 4.4% and 8.5% at the small and medium caches, and Granite loads by 3.8%, 7.7%, and 17.0% across its three capacities. At OLMoE's medium cache, this reduction is accompanied by a throughput increase from 4.237 to 4.372 tokens/s (3.2%). Pairwise prediction also reduces loads, but 1× uses fewer loads at all three capacities on both models.

Refreshing the long-horizon forecast once at mid-depth does not improve transfer counts: 2× is slightly worse than 1× on the small and medium caches for both architectures and ties it on OLMoE's large cache. Its only numerical advantage over 1× is OLMoE large-cache throughput, 5.439 versus 5.434 tokens/s. A 4× refresh schedule shows the same pattern on both models (Appendix Table 11), further indicating that more frequent recomputation does not improve residency over one early long-horizon forecast. Thus the useful regime is not "predict again as soon as possible," but rather "give the cache a sufficiently long look-ahead."

This does not imply that long-horizon forecasts are intrinsically more accurate: cache utility depends on lead time. Short-horizon forecasts may arrive too late to prevent earlier evictions, whereas long-horizon forecasts can protect experts across intervening layers.

6 Discussion

The central result is not merely that aligned routing states are similar across layers. The router-control factorization identifies an exact nuisance symmetry, and fixing that symmetry exposes a canonical space in which a low-capacity transition becomes reusable across depth. This distinguishes a coordinate explanation of the observed rotation from a hypothesis of unrelated layer-wise dynamics.

The matched-readout result is central to this interpretation. If our observation were only generic residual smoothness, residual PCA should be an equally good routing representation. It is not. PCA predicts future residual state more accurately, while router-control coordinates preserve expert choices much more faithfully. This separates two properties that are easy to conflate: *temporal predictability* and *routing relevance*. The CNN/DailyMail replication across Granite, OLMoE, and Phi-tiny further shows that this advantage is not specific to WikiText-2 trajectories.

Our results also make the phrase "shared dynamics" more precise. We do not find a universal transition that makes all layers dynamically identical. Rather, Eq. (16) says that apparently different raw transitions are approximately related by layer-specific gauge changes, and one shared transition captures most of the predictive power of independently fitted layer-specific transitions with far fewer parameters. Its causal advantage over identity is clearest at longer horizons on Phi and OLMoE; the explicit shared-expert architecture exhibits weaker universality. The evidence therefore supports reusable cross-layer structure, not exact dynamical equivalence. Proposition 2 further separates representation error from routing error: functional stability depends on the transported-state error measured through the logit decoder relative to the local top-k margin.

As a supporting application, the cache experiment shows that the canonical coordinates also carry actionable long-range routing information: one early forecast reduces physical transfers without changing native routing. We view this as practical evidence for the usefulness of the representation rather than the central novelty claim.

7 Limitations

The main controlled experiments use WikiText-2, with cross-corpus replication on CNN/DailyMail for Granite, OLMoE, and Phi-tiny. Broader domains and zero-shot transfer of a fixed canonicalization remain open. Phi-tiny-MoE is instruction-tuned, so its causal results are used for intervention comparisons rather than perplexity claims. Router skipping still accumulates error. Cache experiments cover OLMoE and Granite and target offloaded-expert serving; broader serving configurations remain a natural direction for future evaluation.

8 Conclusion

Across four sparse MoE architectures, the router-control factorization induces an exact orthogonal gauge whose removal exposes a canonical cross-layer space with reusable dynamics. A single transition captures most of the predictive power of layer-specific models, while matched-rank and causal controls show that the effect is routing-specific and functionally meaningful but not equivalent to safe router replacement. As a downstream application, the same coordinates also provide useful long-range expert-residency forecasts without changing native routing.

AI Use Statement

Generative AI tools were used to assist with manuscript drafting and language editing, figure formatting, and generation of icons for the pipeline illustration. All research ideas, experimental design, implementation, analysis, interpretation, and conclusions are the authors' own.

References

Akshit Achara, Tatiana Gaintseva, Mateo Mahaut, Pritish Chakraborty, Viktor Stenby Johansson, Melih Barsbey, Emanuele Rodolà, and Donato Crisostomi. Multi-way representation alignment, 2026. URL https://arxiv.org/abs/2602.06205.

Sagi Ahrac, Noya Hochwald, and Mor Geva. Routers learn the geometry of their experts: Geometric coupling in sparse mixture-of-experts, 2026. URL https://arxiv.org/abs/2605.12476.

Nishant Suresh Aswani, Saif Jabari, and Muhammad Shafique. Representing neural network layers as linear operations via koopman operator theory. In *Proceedings of UniReps: the Third Edition of the Workshop on Unifying Representations in Neural Models*, volume 322 of *Proceedings of Machine Learning Research*, pp. 70–80. PMLR, 2026. URL https://proceedings.mlr.press/v322/aswani26a.html.

Zhiyuan Fang, Xingfan Yu, Yuegui Huang, Zicong Hong, Yufeng Lyu, Wuhui Chen, Yue Yu, and Fan Yu. Fate: Fast edge inference of mixture-of-experts models via cross-layer gate. In *Proceedings of the ACM Web Conference 2026*, 2026. URL https://arxiv.org/abs/2502.12224.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022. URL https://www.jmlr.org/papers/v23/21-0998.html.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529. PMLR, 2019. URL https://proceedings.mlr.press/v97/kornblith19a.html.

Niklas Muennighoff, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Jacob Morrison, Sewon Min, Weijia Shi, Evan Pete Walsh, Oyvind Tafjord, Nathan Lambert, Yuling Gu, Shane Arora, Akshita Bhagia, Dustin Schwenk, David Wadden, Alexander Wettig, Binyuan Hui, Tim Dettmers, Douwe Kiela, Ali Farhadi, Noah A. Smith, Pang Wei Koh, Amanpreet Singh, and Hannaneh Hajishirzi. OLMoE: Open mixture-of-experts language models. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/9b224ace8963c9385ad5e2b5c9039b97-Abstract-Conference.html.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017. URL https://arxiv.org/abs/1701.06538.

Alex H. Williams, Erin Kunz, Simon Kornblith, and Scott W. Linderman. Generalized shape metrics on neural representations. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/252a3dbaeb32e7690242ad3b556e626b-Abstract.html.

Fuzhao Xue, Zian Zheng, Yao Fu, Jinjie Ni, Zangwei Zheng, Wangchunshu Zhou, and Yang You. OpenMoE: An early effort on open mixture-of-experts language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 55625–55655. PMLR, 2024. URL https://proceedings.mlr.press/v235/xue24c.html.

Charles Ye, Bo Yuan, and Lee Sharkey. Polysemantic experts, monosemantic paths: Routing as control in MoEs, 2026. URL https://arxiv.org/abs/2604.17837.

Youngsik Yoon, Siwei Wang, Wei Chen, and Jungseul Ok. When are experts misrouted? counterfactual routing analysis in mixture-of-experts language models, 2026. URL https://arxiv.org/abs/2605.07260.

Yingnan Zhao, Razvan Bunescu, Ahmed Louri, Avinash Karanth, and Ke Wang. A spatio-temporal expert prefetching framework for efficient MoE-based LLM inference, 2026. URL https://arxiv.org/abs/2606.15453.

Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Y. Zhao, Andrew M. Dai, Zhifeng Chen, Quoc V. Le, and James Laudon. Mixture-of-experts with expert choice routing. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/2f00ecd787b432c1d36f3de9800728eb-Abstract-Conference.html.

Shien Zhu, Samuel Bohl, Robin Oester, and Gustavo Alonso. Pre-attention expert prediction and prefetching for mixture-of-experts large language models, 2025. URL https://arxiv.org/abs/2511.10676.

A Mathematical Details

Proof of Proposition 1. For $R \in \mathrm{O}(r)$,

$$B'_\ell x'_\ell = (B_\ell R)(R^\top x_\ell) = B_\ell (RR^\top) x_\ell = B_\ell x_\ell. \tag{25}$$

Thus all centered logits, softmax probabilities, and top-k decisions are identical. The same argument applies token-wise to a data matrix X_ℓ, whose coordinate representation transforms as $X_\ell \mapsto X_\ell R$.

Procrustes closed form. For centered X, Y,

$$\begin{aligned} \min_{R \in \mathrm{O}(r)} \|XR - Y\|_F^2 &= \|X\|_F^2 + \|Y\|_F^2 - 2 \max_{R \in \mathrm{O}(r)} \operatorname{tr}(R^\top X^\top Y) \\ &= \|X\|_F^2 + \|Y\|_F^2 - 2\|X^\top Y\|_*, \end{aligned} \tag{26}$$

where the maximum is attained by the standard orthogonal Procrustes solution obtained from the SVD of $X^\top Y$.

GPA pairwise-dispersion identity. Let $M^* = L^{-1} \sum_\ell Z_\ell$. Expanding squared distances gives

$$\sum_\ell \|Z_\ell - M^*\|_F^2 = \sum_\ell \|Z_\ell\|_F^2 - L\|M^*\|_F^2 = \frac{1}{L} \sum_{\ell < m} \|Z_\ell - Z_m\|_F^2. \tag{27}$$

Hence minimizing the GPA objective minimizes the average pairwise disagreement among all aligned layer representations.

Proof of Proposition 2. Let i be any selected expert and j any unselected expert. By definition of the top-k margin, $g_i - g_j \geq \gamma_k$. With logit perturbation $\delta g = D_\ell e$,

$$\hat{g}_i - \hat{g}_j \geq \gamma_k - |\delta g_i| - |\delta g_j| > 0 \tag{28}$$

whenever $\|\delta g\|_\infty < \gamma_k/2$. Therefore no selected expert can cross an unselected expert, so the top-k set is unchanged. Finally, $\|D_\ell e\|_\infty \leq \|D_\ell\|_{2\to\infty} \|e\|_2$ yields Eq. (24).

Least-squares shared operator. Ignoring the intercept after centering, the pooled objective is $\mathbb{E}\|z_{\ell+1} - A z_\ell\|_2^2$. The minimum-norm solution is

$$A^* = C_{10} C_{00}^\dagger, \qquad C_{10} = \mathbb{E}[z_{\ell+1} z_\ell^\top], \quad C_{00} = \mathbb{E}[z_\ell z_\ell^\top]. \tag{29}$$

This makes explicit that the shared dynamics are determined by cross-layer covariance expressed in the canonical gauge.

B Approximate Reducibility and Finite-Horizon Transport

The following statements concern finite sequences of fitted linear transitions. They do not assert reducibility of an infinite cocycle or of the nonlinear MoE computation.

Proposition 3 (Finite-depth transport bound)**.** *Let $F_\ell \in \mathbb{R}^{r \times r}$ be fitted one-step maps and $Q_\ell \in \mathrm{O}(r)$ the independently GPA-fitted gauges. Suppose*

$$Q_{\ell+1}^\top F_\ell Q_\ell = A + E_\ell, \qquad \|E_\ell\|_2 \leq \varepsilon \tag{30}$$

for $\ell, \ldots, \ell + h - 1$. Then

$$\left\| F_{\ell+h-1} \cdots F_\ell - Q_{\ell+h} A^h Q_\ell^\top \right\|_2 \leq (\|A\|_2 + \varepsilon)^h - \|A\|_2^h. \tag{31}$$

More precisely, the right-hand side can be replaced by $\sum_{j=0}^{h-1} \left(\prod_{i=j+1}^{h-1} (\|A\|_2 + \|E_{\ell+i}\|_2) \right) \|E_{\ell+j}\|_2 \|A\|_2^j$.

Proof. Orthogonality cancels each adjacent pair $Q_m^\top Q_m$, so the product is $Q_{\ell+h}(A + E_{\ell+h-1}) \cdots (A + E_\ell) Q_\ell^\top$. Subtract $Q_{\ell+h} A^h Q_\ell^\top$, telescope the difference of the two products one factor at a time, and apply submultiplicativity and the triangle inequality. Bounding every defect by ε gives the difference of powers in Eq. (31). □

For the affine state forecast, write the observed canonical transition as $z_{m+1} = Az_m + b + \xi_m$ and initialize $\hat{z}_\ell = z_\ell$. Iteration yields

$$z_{\ell+h} - \hat{z}_{\ell+h} = \sum_{j=0}^{h-1} A^{h-1-j} \xi_{\ell+j}, \qquad \|z_{\ell+h} - \hat{z}_{\ell+h}\|_2 \le \sum_{j=0}^{h-1} \|A^{h-1-j}\|_2 \|\xi_{\ell+j}\|_2. \tag{32}$$

Corollary 1 (Finite-horizon top-k stability)**.** *Under the state recursion above, the native top-k set at layer $\ell + h$ equals the set decoded from $\hat{z}_{\ell+h}$ whenever*

$$\sum_{j=0}^{h-1} \|A^{h-1-j}\|_2 \|\xi_{\ell+j}\|_2 < \frac{\gamma_k^{(\ell+h)}}{2\|D_{\ell+h}\|_{2\to\infty}}. \tag{33}$$

Proof. Equation (32) bounds the target-state error; apply Proposition 2 with the target-layer decoder and margin. □

This is an *a posteriori* certificate: the realized residuals and target-layer margin need not be known when prefetch decisions are made. It makes no claim about language-model loss.

C Detailed Related Work

MoE routing and specialization. Sparse gating underlies large-scale MoE systems such as sparsely gated MoE and Switch Transformers Shazeer et al. (2017); Fedus et al. (2022). Later work studies routing schemes and expert specialization in models such as Expert Choice, OLMoE, and OpenMoE Zhou et al. (2022); Muennighoff et al. (2025); Xue et al. (2024). We instead study how routing-relevant representations relate *across depth*.

Routing structure across depth. Polysemantic Experts, Monosemantic Paths decomposes hidden states into routing-visible control and routing-invisible content, $h_\ell = h_\ell^{\mathrm{ctrl}} + h_\ell^{\mathrm{content}}$, and observes that control rotates across layers Ye et al. (2026). Fate and later prefetching methods likewise show that future expert choices are predictable from earlier routing signals Fang et al. (2026); Zhu et al. (2025); Zhao et al. (2026). These results reveal cross-layer routing structure, but not whether it reflects shared dynamics or layer-specific coordinates. We address this by aligning router-control states and studying their shared geometry and dynamics.

Representation alignment. CKA and Procrustes-based methods provide tools for comparing neural representations beyond their native coordinate systems Kornblith et al. (2019); Williams et al. (2021), while recent work extends orthogonal alignment to multiple representations using GPA Achara et al. (2026). We instead apply this idea to different depths of the same MoE, using the orthogonal non-identifiability of the router-control factorization to define a common coordinate system for routing states.

Layer-wise dynamics. Viewing network depth as discrete time has precedents in dynamical-systems and Koopman-style analyses Aswani et al. (2026). Rather than modeling the full hidden state, we study router-control states and test whether their apparently layer-specific transitions become shared after alignment.

D Additional Experimental Details

D.1 Cross-Corpus Replication on CNN/DailyMail

We repeat the representation-only routing forecast experiment on CNN/DailyMail 3.0 for Granite, OLMoE, and Phi-tiny. Each reported test run contains 4,096 held-out tokens and is repeated for

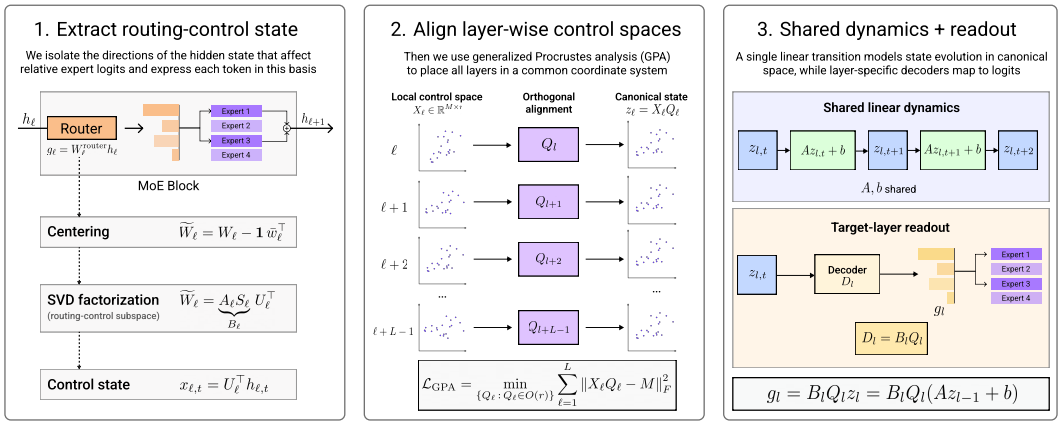


Figure 2: **Method overview.** We extract router-control states from layer-wise MoE routers, align them into a shared canonical space, and model their evolution with shared dynamics. For cache management, cross-layer forecasts in the same canonical coordinates guide expert residency; native routing remains unchanged.

Table 4: Evaluated MoE architectures. Rank is the numerical rank of the centered router.

Model	Layers	Experts	Top-k	Rank
Granite	24	32	8	31
OLMoE-SFT	16	64	8	63
Phi-tiny	32	16	2	15
IBM Shared	40	62	6	61

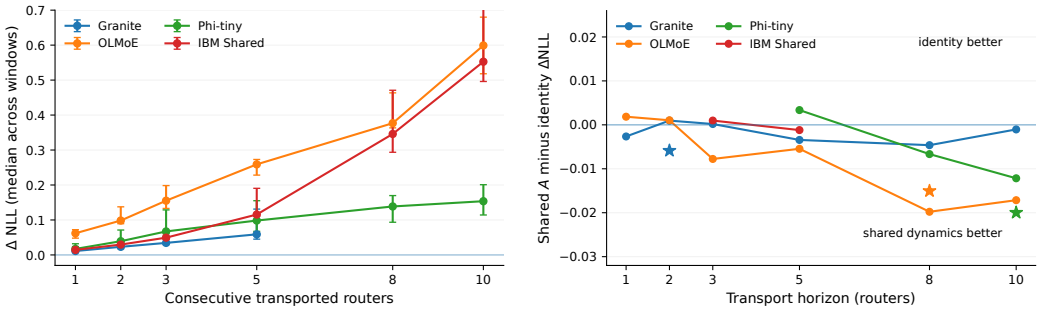
seeds 101, 202, and 303. The canonicalization and forecast models are fitted on the news corpus, so this experiment tests replication across corpus distributions rather than zero-shot transfer from WikiText-2.

Table 5 reports mean top-k recall across the three seeds. The pattern is consistent across all three architectures: canonical forecasts remain above persistence and strongly outperform raw and shuffled coordinates throughout the nonterminal horizon range. OLMoE and Phi-tiny additionally include a direct shared-logit control, which is also substantially weaker. At the terminal horizon, the direct controls coincide, while the canonical forecast remains above persistence. Taken together, these results show that the routing geometry identified on WikiText-2 persists on a qualitatively different news corpus and across multiple MoE architectures.

Table 5: **Cross-corpus routing forecast on CNN/DailyMail 3.0.** Mean top-k recall over seeds 101/202/303; 4,096 held-out test tokens per seed. Higher is better. Shared-logit was not reported for Granite.

Model	Horizon	Canonical	Persistence	Raw	Shuffled	Shared-logit
Granite	1	0.783	0.769	0.514	0.534	–
Granite	4	0.739	0.721	0.501	0.526	–
Granite	8	0.698	0.676	0.504	0.523	–
Granite	16	0.641	0.607	0.513	0.527	–
Granite	23	0.663	0.592	0.663	0.663	–
OLMoE	1	0.722	0.707	0.278	0.353	0.394
OLMoE	4	0.666	0.639	0.308	0.376	0.377
OLMoE	8	0.615	0.571	0.333	0.377	0.370
OLMoE	15	0.462	0.386	0.462	0.462	0.462
Phi-tiny	1	0.531	0.516	0.232	0.227	0.151
Phi-tiny	4	0.518	0.497	0.240	0.229	0.141
Phi-tiny	8	0.483	0.460	0.252	0.246	0.144
Phi-tiny	16	0.425	0.408	0.246	0.251	0.156
Phi-tiny	31	0.406	0.397	0.406	0.406	0.406

Causal horizon sweep. To characterize how transport degrades with distance, we sweep the number of consecutive transported routers over contiguous candidate windows on one exploratory split. Figure 3 shows two complementary effects. Absolute ΔNLL generally increases with transport horizon, indicating error accumulation under longer interventions. At the same time, the learned shared transition becomes increasingly competitive with identity persistence at longer horizons, particularly for OLMoE and Phi-tiny. The highlighted blocks are then fixed before evaluation on the additional splits reported in Table 10.



(a) Transport horizon vs. ΔNLL **(b) Learned dynamics vs. persistence**

Figure 3: **Causal transport has a clear horizon structure.** (a) Median ΔNLL over contiguous candidate windows on the exploratory split, with interquartile ranges. (b) Shared A minus identity ΔNLL; values below zero favor learned dynamics, and stars mark blocks fixed before evaluation on the additional splits.

Reconstruction checks. At full numerical rank, centered-router reconstruction reproduces native top-k decisions exactly across all evaluated layers and models.

Complete alignment controls. Table 6 reports the alignment baselines used in Fig. 1. Values are mean shared-dynamics R^2 across the three split seeds. "Random Q" applies independently sampled orthogonal gauges; "PCA basis" aligns via PCA coordinate conventions rather than token-wise Procrustes.

Table 6: **Complete alignment ablation.** Mean shared-dynamics R^2 across three split seeds.

Model	Raw	PCA basis	Random Q	CCA	Procrustes
Granite	0.071	0.030	0.081	0.198	**0.544**
OLMoE	0.053	0.034	0.091	0.161	**0.486**
Phi-tiny	0.220	0.108	0.224	0.202	**0.707**
IBM Shared	0.033	0.011	0.033	0.106	**0.385**

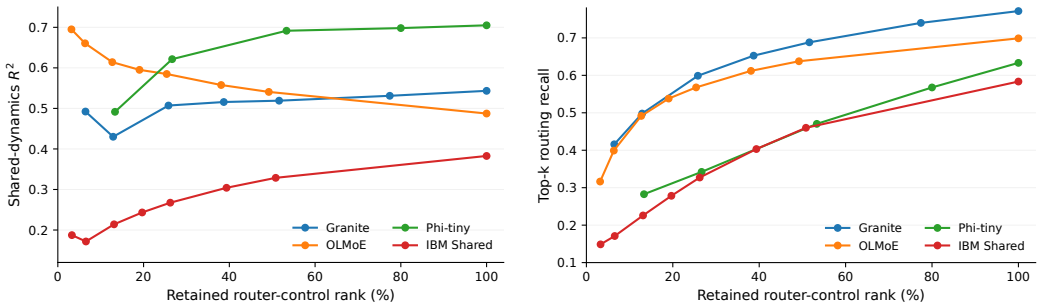
Matched readout. Router-control and residual-PCA representations use learned linear logit decoders with the same per-layer parameterization $(r+1)E$. Exact cross-model means are reported in Table 2.

Fixed-block replication across splits. Table 7 expands Table 10. Blocks identified on the exploratory split are fixed before evaluation on the additional splits and are not re-selected.

Rank sweeps. Rank truncation is applied before gauge alignment. Table 8 shows representative OLMoE points and Fig. 4 the full cross-model curves. We treat the sweep as structural evidence rather than a final benchmark.

Table 7: **Fixed-block causal replication across splits.** Lower ΔNLL is better; Shared−Id. < 0 favors learned dynamics.

Model	Block	Seed	Identity	Shared	Shared−Id.
OLMoE	8–9	101	0.0386	0.0326	−0.0060
		202	0.0457	0.0418	−0.0039
		303	0.0281	0.0204	−0.0077
Phi-tiny	24–31	101	0.2552	0.2471	−0.0080
		202	0.2819	0.2654	−0.0166
		303	0.2639	0.2434	−0.0206
Phi-tiny	22–31	101	0.3112	0.3001	−0.0110
		202	0.3378	0.3120	−0.0258
		303	0.3161	0.2930	−0.0231



(a) Shared-dynamics R^2 **(b) Top-k routing recall**

Figure 4: **A low-dimensional shared structure captures much of the predictable dynamics, while accurate routing requires more dimensions.** OLMoE shows this most clearly: low-rank states are easy to predict, but higher ranks are needed to recover expert choices accurately. Rank is normalized by each model's full centered-router rank.

E Forecast-Guided Expert Residency

For cache management, we use the canonical state to predict expert demand at later layers $\ell + d$ of the *same token*. For each horizon d, a ridge forecast on calibration pairs gives

$$\hat{z}_{\ell+d,t} = T_d z_{\ell,t} + c_d, \qquad \hat{g}_{\ell+d,t} = D_{\ell+d}\hat{z}_{\ell+d,t}. \tag{34}$$

At inference, we algebraically fold the source-router projection, canonical alignment, forecast, and target readout into a small map from the *already computed* source router logits to predicted target logits. Forecasts expire after their target layer or when the next token begins; no signal from token $t + 1$ is used.

We place experts' immutable weights in host memory and keep a fixed number of layer–expert entries on the GPU. A forecast assigns higher retention priority to resident experts likely to be requested at later layers. When the cache fills, the policy evicts an unprotected entry if possible, then uses forecast score and recency to choose among protected entries. It does *not* proactively copy experts in the reported eviction-only comparison. Each target layer still evaluates its native router, loads any missing exact experts, and executes those experts with the native gate weights. Thus forecasting can change transfers and runtime, but not the model's computation or generated tokens.

F Physical Cache Protocol

The physical-cache experiments use `allenai/OLMoE-1B-7B-0125-SFT` and `ibm-granite/granite-3.0-1b-a400m-base`. The text source is `Salesforce/wikitext`, configuration `wikitext-2-raw-v1`, train split. Calibration, validation, forecast evaluation, and the physical benchmark use disjoint documents. We use 128-token segments, 128 calibration sequences, 16 validation sequences, 392 forecast-evaluation sequences, and 32 physical benchmark requests; benchmark documents are excluded from fitting and model selection.

Table 8: **OLMoE rank trade-off** on one exploratory split. Lower causal ΔNLL and higher R^2/recall are better. The full centered-router rank is 63.

Rank	Shared R^2	Top-k recall	Causal ΔNLL
2	**0.695**	0.316	0.593
8	0.614	0.492	0.309
16	0.585	0.568	0.230
31	0.541	0.638	0.176
63	0.487	**0.699**	**0.118**

Table 9: **Granite causal controls, layers 9–13.** Mean ΔNLL ± standard deviation across seeds 101/202/303. Lower is better.

Transport variant	ΔNLL
Canonical + shared A	0.0070 ± 0.0016
Canonical + identity	0.0142 ± 0.0032
Raw coordinates + shared A	0.1120 ± 0.0046
Shuffled Q + shared A	0.1012 ± 0.0060
Random subspace + shared A	0.1495 ± 0.0099

Rank and ridge are fixed before the physical benchmark; no Procrustes parameter is tuned on benchmark requests. Canonical forecasts are fitted for horizons $d = 1, \ldots, L-1$ with ridge coefficient 10^{-3}. Local-2 predicts one or two layers ahead every two sparse layers; Local-4 predicts one to four layers ahead every four sparse layers. The $1\times$ schedule predicts all later sparse layers from the first sparse layer, $2\times$ refreshes once at mid-depth, and $4\times$ refreshes four times across the sparse-layer sequence. Cache capacities are 32/64/128 layer–expert entries for OLMoE and 24/48/96 for Granite. All policies see the same native routing traces and three request orderings; only cache residency decisions differ.

Pairwise control. PAIRWISE is the native-logit forecasting control used in the cache tables. It predicts future routing directly in native router-logit coordinates rather than through the canonical control space, and uses a substantially larger parameterization than the canonical forecast. It therefore tests whether the cache gains require the canonical representation rather than merely a sufficiently expressive routing predictor.

Each physical run contains 32 requests with 128 generated tokens each, for 4,096 autoregressive decoding steps per request ordering. This corresponds to 524,288 expert accesses for OLMoE and 786,432 for Granite, so transfer rates aggregate hundreds of thousands of cache-access events per run.

For throughput, the table reports $\bar{x} \pm t_{0.975,2}\, s/\sqrt{3}$ over the three request orderings, using each run aggregate as one observation. These intervals summarize run-to-run variation on the fixed benchmark workload. All Granite cache runs, including LRU, Pairwise, and Procrustes schedules, were measured on the same NVIDIA V100 GPU. The systems experiment targets offloaded-expert serving, where host-to-GPU transfers contribute materially to runtime.

Table 10: **Fixed-block causal replications.** Means over three fixed 50,176-token splits. The final column is Shared minus Identity; negative values favor learned dynamics.

Model	Block	Identity	Shared	Shared−Id.
OLMoE	8–9	0.0375	**0.0316**	−0.0059
Phi-tiny	24–31	0.2670	**0.2520**	−0.0150
Phi-tiny	22–31	0.3217	**0.3017**	−0.0200

Table 11: **Full physical expert-offloading schedule comparison.** Lower loads/token (↓) and higher throughput (↑) are better. Values are mean ± 95% Student-t interval half-width over the same three request orderings as Table 3.

Method	Small cache		Medium cache		Large cache	
	Loads ↓	Tok/s ↑	Loads ↓	Tok/s ↑	Loads ↓	Tok/s ↑
OLMoE	32 slots		64 slots		128 slots	
LRU	128.00 ± 0.00	4.246 ± 0.009	128.00 ± 0.00	4.237 ± 0.010	78.55 ± 0.04	5.424 ± 0.011
Pairwise	124.25 ± 0.16	4.231 ± 0.010	118.20 ± 0.16	4.333 ± 0.012	78.50 ± 0.04	5.255 ± 0.011
Procrustes (local-2)	128.00 ± 0.00	4.070 ± 0.009	128.00 ± 0.00	4.046 ± 0.009	78.51 ± 0.04	5.224 ± 0.011
Procrustes (local-4)	128.00 ± 0.00	4.156 ± 0.012	128.00 ± 0.00	4.152 ± 0.015	78.49 ± 0.03	5.393 ± 0.011
Procrustes (1×)	$\mathbf{122.40 \pm 0.19}$	$\mathbf{4.268 \pm 0.014}$	$\mathbf{117.17 \pm 0.17}$	$\mathbf{4.372 \pm 0.017}$	$\mathbf{78.45 \pm 0.04}$	5.434 ± 0.010
Procrustes (2×)	123.44 ± 0.21	4.263 ± 0.014	117.56 ± 0.22	4.369 ± 0.016	$\mathbf{78.45 \pm 0.04}$	$\mathbf{5.439 \pm 0.019}$
Procrustes (4×)	124.09 ± 0.03	4.148 ± 0.005	118.39 ± 0.04	4.265 ± 0.022	$\mathbf{78.45 \pm 0.03}$	5.246 ± 0.038
Granite	24 slots		48 slots		96 slots	
LRU	192.00 ± 0.00	4.353 ± 0.009	192.00 ± 0.00	4.371 ± 0.009	192.00 ± 0.00	4.369 ± 0.009
Pairwise	187.46 ± 0.22	4.143 ± 0.015	182.61 ± 0.30	4.318 ± 0.016	166.97 ± 0.30	4.468 ± 0.017
Procrustes (local-2)	192.00 ± 0.00	4.296 ± 0.016	192.00 ± 0.00	4.261 ± 0.015	192.00 ± 0.00	4.264 ± 0.015
Procrustes (local-4)	192.00 ± 0.00	4.287 ± 0.015	192.00 ± 0.00	4.268 ± 0.015	192.00 ± 0.00	4.232 ± 0.015
Procrustes (1×)	$\mathbf{184.62 \pm 0.17}$	$\mathbf{4.360 \pm 0.009}$	$\mathbf{177.17 \pm 0.21}$	$\mathbf{4.429 \pm 0.009}$	$\mathbf{159.37 \pm 0.23}$	$\mathbf{4.619 \pm 0.010}$
Procrustes (2×)	185.38 ± 0.16	4.357 ± 0.009	179.09 ± 0.19	4.409 ± 0.009	161.91 ± 0.22	4.605 ± 0.010
Procrustes (4×)	186.15 ± 0.08	4.300 ± 0.004	180.60 ± 0.11	4.347 ± 0.006	164.24 ± 0.05	4.520 ± 0.004

Notes. Intervals summarize variation across three request orderings. Bold marks the best numerical value in each model/cache setting (including ties).