DyClust-MoE: One Grouping State for Expert Sharing, Routing, and Placement

Anonymous authors
Paper under double-blind review

Abstract

Sparse mixture-of-experts (MoE) models usually optimize expert representation, routing, and placement separately. DyClust-MoE coordinates them through one layer-local grouping state: deterministic proposals undergo exact reconstruction checks, a mass-distilled router and bounded repair retain exactly k experts, and one transaction publishes tensors, routing, and placement. At $E = 128$, it stores 84.68% fewer inference-representation scalars than controlled same-E Switch. On eight A100 GPUs, prefill and full-run training reach 438 k and 171 k tokens/s, with paired ratios $1.26 \pm .030$ and $1.22 \pm .027$; F+B wire is $.73 \pm .027\times$. Paired quality differences are $-.08 \pm .10$ GLUE points and $+.19 \pm .08$ WikiText-103 PPL. A budget-matched, cached Decoupled-3State controller achieves slightly better local objectives, yet shared grouping improves two-/four-node prefill by $1.15 \pm .022/1.19 \pm .024\times$, training by $1.13 \pm .019\times$, and reduces wire to $.80 \pm .021\times$. Executed accelerator FLOPs are about $1.05\times$ nominal-matmul Switch: the speed advantage comes from communication and execution organization. Controlled shifts support faster recovery, and two public MoE transfers satisfy the prespecified quality and systems margins. Code: `https://anonymous.4open.science/r/SUBMIT-0001/README.md`.

1 Introduction

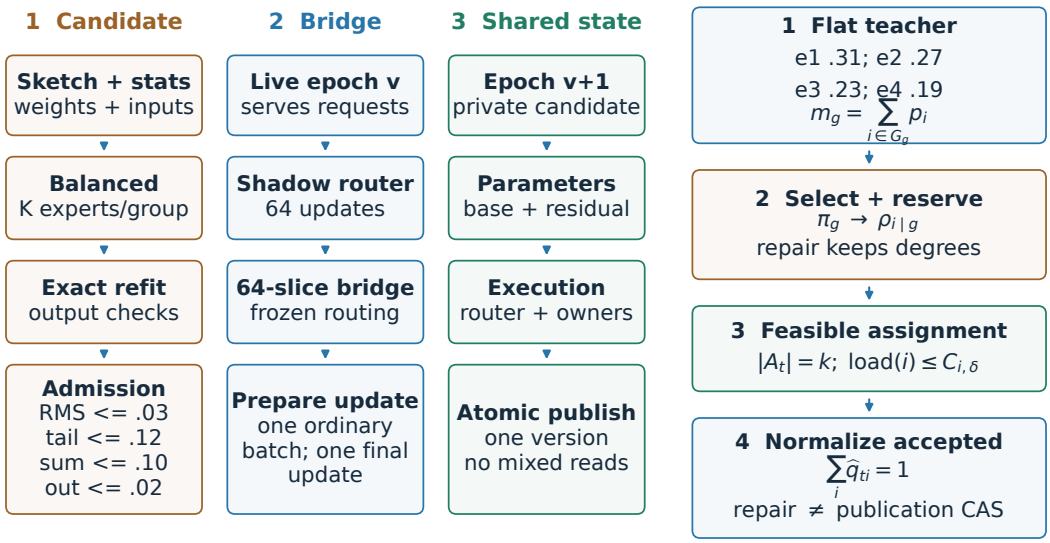
Sparse expert layers reduce per-token computation while retaining a large parameter pool and exchanging states across expert-parallel devices (Shazeer et al., 2017; Lepikhin et al., 2021; Fedus et al., 2022; Rajbhandari et al., 2022). These costs are coupled: sharing weights changes representation, restricting candidates changes load, and placement determines remote traffic. For example, two experts can share a base yet still require separate remote payloads if routing and ownership ignore that sharing. Compression alone then leaves the communication opportunity unrealized.

Our central abstraction is a *grouping state*. In every MoE layer, one balanced partition defines a shared-base residual representation, the support of a group-first router, and the units assigned to devices. Candidate admission tests reconstruction, retained flat-router mass, and exact-k feasibility. Given an admitted partition, a separate placement solve optimizes a topology-aware proxy on calibration traces; disjoint evaluation traces are used only for final systems comparisons.

Candidate generation uses deterministic sketches, activation centroids, and co-routing rows; exact refits and output checks govern retention. A group selector learns flat-router mass, a cyclic reserve initializes feasible token–expert assignments, and bounded repair improves routing scores. A training-only bridge and one activation point preserve epoch consistency under the stated publication assumptions. The evaluation measures the resulting quality–systems trade-off.

Our contributions are (i) a deterministic grouping procedure that separates a linearized proposal from exact reconstruction-based acceptance; (ii) a mass-distilled router and reserve construction with an exact-k feasibility proof and explicit routing-error terms; (iii) a fixed-charge placement model for coalesced source–owner messages and critical-rank time; and (iv) a failure-closed training transaction, exact scalar/byte ledgers, and matched quality, serving, scaling, and transfer measurements.

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.



(a) A grouping version indexes representation, routing, and placement.

(b) Mass distillation and reserve assignment.

Figure 1: Overview of DyClust-MoE. Exact refits and output checks precede activation. In panel (b), $\mathcal{A}_t = \{i : A_{ti} = 1\}$ is the accepted expert set; assignment repair changes only private routing state, whereas the later version CAS publishes parameters, routing, and placement together.

2 Related Work

Sparse routing includes GShard, Switch, GLaM, Mixtral, and Expert Choice (Lepikhin et al., 2021; Fedus et al., 2022; Du et al., 2022; Jiang et al., 2024; Zhou et al., 2022). BASE Layers uses balanced assignment (Lewis et al., 2021); StableMoE stabilizes routing, MoE++ adds zero-computation experts, and semantic priors guide visual MoE routing (Dai et al., 2022; Jin et al., 2025; Min et al., 2025). DeepSeekMoE isolates always-on shared experts (Dai et al., 2024). Our group bases share matrix parameters within routed experts; Qwen's separate shared expert remains unchanged during transfer.

Pruning and quantization compress model weights (Han et al., 2015; Hubara et al., 2018; Jacob et al., 2018); LoRA and mixtures of low-rank experts target parameter-efficient fine-tuning (Hu et al., 2022; Sun et al., 2025). Sub-MoE merges expert subspaces, while D^2-MoE decomposes expert weights into bases and deltas (Li et al., 2026; Gu et al., 2025). QMoE couples sub-1-bit compression with decoding kernels (Frantar & Alistarh, 2024); its encoded-byte savings differ from our BF16 scalar inventory. MoE-SVD, LorExperts, and dynamic expert clustering study structured compression (Li et al., 2025; Chakrabarti et al., 2026; Zhu et al., 2025). Our comparisons isolate representation and state coordination; OC-Share is a controlled sharing variant.

Tutel, FasterMoE, FlexMoE, SMILE, DeepSpeed-MoE, and MoE-Lightning optimize dispatch, scheduling, or placement (Hwang et al., 2023; He et al., 2022b; Nie et al., 2023; He et al., 2022a; Rajbhandari et al., 2022; Cao et al., 2025). MegaBlocks uses dropless block-sparse execution; ScatterMoE reduces padding and copying (Gale et al., 2023; Tan et al., 2024). These execution mechanisms address a different interface from grouping, owner selection, and destination coalescing. Our common zero-drop controls and frozen replays measure those interfaces under matched execution contracts. The analysis builds on low-rank approximation, network flow, and linearizability (Halko et al., 2011; Ahuja et al., 1993; Herlihy & Wing, 1990).

3 Method

Let E experts be partitioned into G groups of size $K = E/G$. The derivation permits $1 \le k \le E$ active experts per token; the primary small-model operating point uses $k = 2$. For an MoE layer,

training minimizes

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda_E \mathcal{L}^E_{\text{lb}} + \lambda_G \mathcal{L}^G_{\text{lb}} + \lambda_C \mathcal{L}_{\text{conf}} + \lambda_M \mathcal{L}_{\text{mass}}, \tag{1}$$

where all sums below are over the global non-padding token set $\mathcal{T}_\delta$ of capacity domain δ, and $N_\delta = |\mathcal{T}_\delta|$. Let q^{sc}_{ti} be the normalized pre-repair probability on the evaluated support (equal to $q^{\mathcal{S}}_{ti}$ in Equation (8)), $A_{ti} \in \{0,1\}$ an accepted token–expert edge, and $C_{tg} = \mathbf{1}\{\exists i \in \mathcal{G}_g : A_{ti} = 1\}$.

$$p_i = \frac{\sum_t q^{\text{sc}}_{ti}}{N_\delta}, \quad f_i = \frac{\sum_t A_{ti}}{kN_\delta}, \quad p_g = \frac{\sum_t \pi_{tg}}{N_\delta}, \quad f_g = \frac{\sum_t C_{tg}}{\sum_{t,h} C_{th}},$$

and $\mathcal{L}^E_{\text{lb}} = E \sum_i p_i f_i$, $\mathcal{L}^G_{\text{lb}} = G \sum_g p_g f_g$. If $b^{(0)}_{tg}$ marks the stop-gradient highest-scoring selector group (group-ID ties), $\mathcal{L}_{\text{conf}} = N_\delta^{-1} \sum_t (1 - \sum_g b^{(0)}_{tg} \pi_{tg})$ is a confidence regularizer, not an external supervision target. On a counter-hashed $1/64$ teacher subsample, $\mathcal{L}_{\text{mass}}$ is the token mean of $\mathrm{KL}(\mathrm{sg}(m) \| \pi)$. We use $(\lambda_E, \lambda_G, \lambda_C, \lambda_M) = (0.01, 0.01, 0.05, 0.10)$.

3.1 Deterministic grouping

For the small model let $\mathcal{P}_W = \{\text{up}, \text{down}\}$ and $P_W = |\mathcal{P}_W| = 2$. For $p \in \mathcal{P}_W$, freeze effective weights $W^{\text{pre}}_{i,p}$ and define $\epsilon_w = 10^{-12}$ and $\omega_{i,p} = (\|W^{\text{pre}}_{i,p}\|^2_F + \epsilon_w)^{-1}$. A partition has $z_{ig} \in \{0,1\}$, $\sum_g z_{ig} = 1$, $\sum_i z_{ig} = K$, and $g_z(i)$ equal to expert i's group. Let h_t be token t's input to this MoE layer, $n_i = \sum_t A_{ti}$, $\zeta_i = \mathbf{1}\{n_i > 0\}$, $N_\zeta = \max(1, \sum_i \zeta_i)$, $\mu_i = n_i^{-1} \sum_t A_{ti} h_t$ when exposed (zero otherwise), and $\mathrm{unit}(x) = x/(\|x\|_2 + 10^{-12})$. For $i < j$, set $\chi_{ij} = \sum_t A_{ti} A_{tj} / \max(1, \sum_t \binom{\sum_a A_{ta}}{2})$, reflect it symmetrically, and set $\chi_{ii} = 0$; hence $\chi = 0$ for $k = 1$. Define $u_i = \mathrm{unit}(\mu_i)$ and $c_g = \mathrm{unit}(\sum_i z_{ig} \zeta_i u_i)$ (zero for an unexposed group). Here $\overline{W}_{g,p}$ is a group base and $R_{i,p}$ a rank-r residual. The exact objective is

$$\begin{aligned} \mathcal{J}_{\text{grp}} = \frac{1}{P_W E} \sum_{i,g,p} z_{ig} \omega_{i,p} \left\| W^{\text{pre}}_{i,p} - \overline{W}_{g,p} - R_{i,p} \right\|^2_F \\ + \frac{\lambda_{\text{act}}}{N_\zeta} \sum_{i,g} z_{ig} \zeta_i \left\| \mathrm{unit}(\mu_i) - c_g \right\|^2_2 \\ + \lambda_{\text{pair}} \sum_{i<j} \chi_{ij} \mathbf{1}[g_z(i) \neq g_z(j)]. \end{aligned} \tag{2}$$

The rank constraint is $\mathrm{rank}(R_{i,p}) \leq r$, each expert has one group, and every group has K experts. We denote the first, reconstruction-only term of Equation (2) by $\mathcal{J}_{\text{cmp}}$; thus it is a normalized sum over all $P_W E$ expert–projection pairs and is not the full grouping objective.

For zero-based row-major $w_{i,p} = \mathrm{vec}(W^{\text{pre}}_{i,p})$, define a 256-bin signed CountSketch projection (Charikar et al., 2002) $\phi_{i,p}[b] = \sum_{a:\, H(a)=b} \sigma(a) w_{i,p}[a]$. The hash and sign use the deterministic byte encoding in Section A. Euclidean farthest-first selects prototype expert IDs i_g from $x_i = [\bigoplus_{p \in \mathcal{P}_W} (P_W E)^{-1/2} \sqrt{\omega_{i,p}} \phi_{i,p}; \sqrt{\lambda_{\text{act}} \zeta_i / N_\zeta}\, u_i; \sqrt{\lambda_{\text{pair}}/2}\, \chi_{i,:}]$, choosing the smallest expert ID first and then lexicographic ties. Freeze $s^-_{g,p} = \phi_{i_g,p}$, $c^-_g = u_{i_g}$, and incumbent labels g^-_i, and solve the balanced assignment with

$$\begin{aligned} C_{ig} = \frac{1}{P_W E} \sum_p \omega_{i,p} \left\| \phi_{i,p} - s^-_{g,p} \right\|^2_2 + \lambda_{\text{act}} \frac{\zeta_i}{N_\zeta} \left\| u_i - c^-_g \right\|^2_2 \\ + \frac{\lambda_{\text{pair}}}{2} \sum_{j \neq i} \chi_{ij} \mathbf{1}[g \neq g^-_j]. \end{aligned} \tag{3}$$

The factor $1/2$ prevents a symmetric unordered pair from being charged twice when the assignment sums C_{ig} over i. This pair term is a frozen separable surrogate, not the exact pair cut. The assignment is the exact balanced min-cost solution; among equal optima it chooses the lexicographically smallest vector $(g(1), \ldots, g(E))$. After assignment, two probe sweeps screen reconstruction RMS and expert-tail error; candidates that pass it proceed to the complete centroid, pair-cut, full-SVD/base-refit, and swap scan. Final retention still requires an exact-objective improvement satisfying every distortion

budget, and the scan restarts after each accepted pair swap. Alternating refits stop when relative reconstruction decrease is at most 10^{-5}, and a candidate is rejected if it has not converged after eight sweeps; the swap scan stops after a no-improvement pass over its 32-pair surrogate shortlist or $8E$ exact swap evaluations. We use $\lambda_{\text{act}} = 0.25$ and $\lambda_{\text{pair}} = 0.10$. Sketches and the pre-screen only reject or propose; they never replace the complete refit and exact objective used for acceptance. This is a deterministic local search, not a global discrete optimum. Numerical determinism is scoped to the frozen FP64 backend, reduction order, and sign rules specified in Section B, not bitwise identity across backends. A numerical example, initialization, and all gates appear in Section A.

For fixed bases, the residual refit is

$$R_{i,p} \leftarrow \mathcal{P}_r(W^{\text{pre}}_{i,p} - \overline{W}_{g(i),p}), \tag{4}$$

where $\mathcal{P}_r$ is deterministic truncated SVD. For fixed residuals and labels,

$$\overline{W}_{g,p} \leftarrow \frac{\sum_i z_{ig}\omega_{i,p}(W^{\text{pre}}_{i,p} - R_{i,p})}{\sum_i z_{ig}\omega_{i,p}}. \tag{5}$$

Eckart–Young–Mirsky and weighted least squares make both refits non-increasing for a fixed partition; no global claim is made for the discrete search.

3.2 Shared representation and training-only bridge

Each expert retains its bias and uses

$$\widetilde{W}_{i,p} = \overline{W}_{g(i),p} + A_{i,p}B^{\top}_{i,p}, \qquad \operatorname{rank}(A_{i,p}B^{\top}_{i,p}) \le r. \tag{6}$$

Define $\tau_{i,p} = \left\| W^{\text{pre}}_{i,p} - \widetilde{W}_{i,p} \right\|_F / \sqrt{\left\| W^{\text{pre}}_{i,p} \right\|^2_F + \epsilon_w}$ and $\bar{\tau} = \sqrt{\mathcal{J}_{\text{cmp}}}$. The admission ledger records $b_{\ell,e} = \bar{\tau}_{\ell,e}$ and $B_\ell = \sum_e b_{\ell,e}$. B_ℓ is an accounting ledger; it is not a bound on final model distance or semantic equivalence.

After an optimizer step, a global fence freezes the live model, primary optimizer, scheduler, data cursor, and counter-based RNG state. An overlap-remapped private candidate selector and private moments receive 64 selector-only shadow updates on a counter-hashed $7/8$ of the next ordinary batch; the untouched $1/8$ supplies KL/mass/recall gates, and all private selector state is discarded on abort. The same 262,144-token batch is then reused by old and candidate branches as $H = 64$ ordered 4,096-token slices under frozen accepted routing. Neither primary branch updates inside the loop; after slice 64, each prepares one update only for tensors whose semantic IDs are unchanged across epochs, while old expert matrices and candidate bases/factors/selectors remain frozen. Activation requires all preflight, slice-integrity, and finite-loss gates, maximum slice-level candidate-minus-old token-mean NLL < 0.020, and maximum slice-level absolute PPL gap between the gradual candidate and the non-updating immediate-swap sidecar < 0.25. Otherwise finalization retains the once-stepped old branch. Immutable versioned update buffers and phase-tagged durable snapshots make recovery idempotent; only an incomplete slice and its suffix are replayed (the whole batch only from the $s = 0$ snapshot). New base/factor moments start at zero and use a 256-update local learning-rate ramp after selection. Crash safety assumes at least one durable quorum replica retains the latest complete snapshot; exact selector remapping, semantic IDs, and phases appear in Section A.

3.3 Mass router and exact-k assignment

On a counter-hashed teacher subsample, the flat router uses $\ell_{ti} = v_i^{\top}\,\mathrm{LN}(h_t) + b_i$ and scores all E experts:

$$p^{\text{flat}}_i = \frac{e^{\ell_i}}{\sum_j e^{\ell_j}}, \quad m_g = \sum_{i \in \mathcal{G}_g} p^{\text{flat}}_i, \quad \rho_{i|g} = \frac{e^{\ell_i}}{\sum_{j \in \mathcal{G}_g} e^{\ell_j}}. \tag{7}$$

The selector is $\pi_t = \mathrm{softmax}(U\,\mathrm{LN}(h_t) + a)$ and is trained with $\mathrm{KL}(\mathrm{sg}(m)\|\pi)$. Inference starts from the highest-π group, appends groups in descending $(\pi, -g)$ order until selector mass reaches 0.98 or four additional groups have been added, and unions in every group containing a cyclic reserve expert. The cap never removes a reserve group. Expert logits are evaluated only in this union $\mathcal{S}$, and $Z_{\mathcal{S}} = \sum_{j \in \mathcal{S}} \pi_{g(j)}\rho_{j|g(j)}$. Before capacity repair, the supported distribution is

$$q^{\mathcal{S}}_i = \frac{\mathbf{1}\{i \in \mathcal{S}\}\pi_{g(i)}\rho_{i|g(i)}}{Z_{\mathcal{S}}}. \tag{8}$$

For the accepted indicators A_{ti}, define $\widehat{q}_{ti} = A_{ti} q^{\mathcal{S}}_{ti} / \sum_j A_{tj} q^{\mathcal{S}}_{tj}$. The routed MoE output is

$$y_t = \sum_{i : A_{ti}=1} \widehat{q}_{ti} F_i(h_t). \tag{9}$$

The discrete assignment A and the repair path are stop-gradient; gradients pass through $\widehat{q}$ and the selected expert parameters. With omitted teacher mass $\varepsilon = 1 - \sum_{i \in \mathcal{S}} p^{\mathrm{flat}}_i$, restricted teacher $\widetilde{p}_i = \mathbf{1}\{i \in \mathcal{S}\} p^{\mathrm{flat}}_i / (1-\varepsilon)$, selector error $\delta_{\mathrm{router}} = \mathrm{TV}(\widetilde{p}, q^{\mathcal{S}})$, and assignment error $\eta_{\mathrm{hard}} = \mathrm{TV}(q^{\mathcal{S}}, \widehat{q})$, two triangle inequalities give

$$\mathrm{TV}(p^{\mathrm{flat}}, \widehat{q}) \le \varepsilon + \delta_{\mathrm{router}} + \eta_{\mathrm{hard}}. \tag{10}$$

This is a conditional decomposition, not a guarantee that learned errors are small. The all-E teacher is charged only on its training subsample; a capped expansion that misses 0.98 remains visible through ε.

For an expert-parallel capacity domain δ, let N_δ be offered non-padding tokens and $C_{i\delta} = \lceil 1.25 k N_\delta / E \rceil$. Concatenate source ranks stably and let $\nu_t \in \{0, \dots, N_\delta - 1\}$ be token t's resulting ordinal. For the one-indexed expert IDs, set $e^{(0)}_{tj} = 1 + ((k\nu_t + j) \bmod E) \in \{1, \dots, E\}$ for $j = 0, \dots, k-1$. If $k \le E$ and $C_{i\delta} \ge \lceil k N_\delta / E \rceil$, every token has k distinct reserve experts and initial reserve loads differ by at most one. Let $\mathcal{E}_t$ contain all scored-support edges plus these reserve edges. Starting from the reserve assignment, repair decreases

$$-\sum_{t,i} A_{ti} \log(q^{\mathcal{S}}_{ti} + 10^{-12}) \quad \text{s.t.} \quad \sum_i A_{ti} = k, \quad \sum_t A_{ti} \le C_{i\delta}, \quad A_{ti} = 0 \ (i \notin \mathcal{E}_t).$$

Ranks submit hash-checked edge lists once. A deterministic parallel kernel evaluates score-improving residual moves in bounded batches and commits only a feasibility-checked batch to private assignment state. Each committed batch preserves token degree, edge uniqueness, and expert capacity; the batch journal is append-only, and one final delta list is broadcast and hash-verified. Repair stops when no admissible improving batch remains or after $R_{\max} = 16$ candidate-batch evaluation rounds, counting rejected batches. Exact-k and capacity are structural; global score optimality is not claimed.

3.4 Owner-level placement

Let fixed a_{ctsg} indicate that token t from source s invokes group g in calibration invocation c. Binary u_{ctsr} is the union indicator that this token must reach owner r, so a coalesced payload is counted once even if several invoked groups share that owner; x_{csr} marks a nonempty source–owner message. Let y_{gr} place group g, b be its resident bytes, B^{res}_r the owner budget, b_{tok} the constant payload bytes per routed token, and $(\alpha_{sr}, \beta_{sr})$ fitted launch and per-byte coefficients. We solve

$$\min_{y,u,x,T} \quad \sum_c T_c, \tag{11}$$

$$\text{s.t.} \quad \sum_r y_{gr} = 1, \quad \sum_g b y_{gr} \le B^{\mathrm{res}}_r, \tag{12}$$

$$u_{ctsr} \ge y_{gr} \quad (a_{ctsg} = 1), \quad u_{ctsr} \le \sum_{g : a_{ctsg} = 1} y_{gr}, \tag{13}$$

$$x_{csr} \ge u_{ctsr} \ (\forall t), \quad x_{csr} \le \sum_t u_{ctsr}, \tag{14}$$

$$T_c \ge \sum_s \alpha_{sr} x_{csr} + \sum_s \beta_{sr} b_{\mathrm{tok}} \sum_t u_{ctsr}, \tag{15}$$

$$y_{gr}, u_{ctsr}, x_{csr} \in \{0,1\}, \quad T_c \ge 0 \qquad (\forall c, r). \tag{16}$$

The fixed-charge formulation is solved by branch-and-bound with stable ties. Because T_c upper-envelopes every owner's fitted cost, the objective is a critical-owner proxy rather than a sum of owner times. Held-out wire, owner load, link congestion, and critical-rank time determine the systems conclusion.

Conditional invariants. For a fixed partition, Equations (4) and (5) minimize their respective continuous subproblems. The reserve construction proves exact-k feasibility under its capacity assumptions. For frozen inputs and routing, let $C_{i,p} = W_{i,p}^{\mathrm{pre}} - \widetilde{W}_{i,p}^{+}$. The bridge satisfies

$$W_{i,p}^{(s)} = \widetilde{W}_{i,p}^{+} + (1 - s/H)(W_{i,p}^{\mathrm{pre}} - \widetilde{W}_{i,p}^{+}), \tag{17}$$

so adjacent private-path states differ by $\|C_{i,p}\|_F / H$. This does not prove serving continuity or post-activation optimization stability.

Publication is conditional on a durable state record, one in-flight candidate, epoch-tagged requests, and a linearizable activation compare-and-swap. Before activation, requests see only the old epoch; after activation, they address only the complete candidate; after abort, they remain on the old epoch. The claim is crash safety under these assumptions, not liveness or correctness of an unspecified production stack.

4 Evaluation

Setup. The primary model has 12 layers ($d = 768, d_{\mathrm{ff}} = 2688$), eight top-2 MoE FFNs, $K = 8$, $r = 8$, and $E \in \{32, 64, 128\}$. Primary quality, training, serving, HBM, traffic, and downstream endpoints use six paired seeds. Mechanism and bridge/routing/repair/placement validation use four seeds, sensitivity uses five, shifts use six, and each public transfer uses ten compact/native pairs. Comparator matching variables and the complete evaluation protocol are in Section B; native-semantics neighbors are contextual controls.

Unless captioned as point diagnostics or hard checks, entries are outer-unit means $\pm$ SD. Quality and training endpoints are reduced within seed before paired training ratios are formed. Primary quality uses six-seed marginal means; the $E = 128$ same-E Switch contrast also reports paired SD. Serving, HBM, and traffic first reduce five blocks within checkpoint; placement uses four seed-level reducers. Traffic is measured per offered token before capacity or repair: fixed 131,072-token workloads report GB, scaling reports bytes/token, and accepted-token/edge views are diagnostic. Dense rows are descriptive anchors.

Training, data, and selection. Each of the 24 non-Dense primary-quality configurations starts from the same 102.550M-scalar dense parent, pretrained on 100B non-padding token instances, and receives 3B continuation token positions in each of six seeds, for a 432B primary-quality block. Exact normalized-record hashes and token-normalized 13-gram overlap remove held-out benchmark matches before either parent pretraining or continuation. GLUE-8 is the fixed macro of CoLA, SST-2, MRPC, STS-B, QQP, MNLI, QNLI, and RTE using official metrics (Wang et al., 2018); WikiText-103 uses token-level NLL and perplexity (Merity et al., 2017). Each primary configuration evaluates six architecture-local candidates on two disjoint pilot seeds using held-out NLL only; the six evaluation seeds, system traces, and downstream tasks are not visible during selection. OC-Share fixes an output-cluster partition in our base-plus-residual representation; D^2-MoE uses its pinned reference tensor path and a separately named, equivalence-tested optimized kernel (Li et al., 2026; Gu et al., 2025).

Statistical inference. Primary quality contrasts are descriptive. One-sided Bonferroni bounds apply only to shifts and public transfers. Technical repeats never increase scientific sample size; all declared outer units are retained and no non-finite value is imputed.

Quality–storage trade-off. At $E = 128$, DyClust-MoE is 84.68% below controlled same-E Switch in stored inference-representation scalars and 0.85% below the 18-expert upper bracket. Its paired descriptive differences from same-E Switch are -0.08 ± 0.10 GLUE points and $+0.19 \pm 0.08$ PPL. Relative to OC-Share and D^2-MoE, the marginal-mean GLUE/PPL differences are $+0.55/-0.49$ and $+0.51/-0.41$; relative to Decoupled-3State they are $+0.08/-0.02$. The analytical one-group path contains 103.610M active scalars, but the measured 1.23 bases/token raises the realized value to about 111.21M. This corresponds to about 5.189M expert-matrix scalars per MoE layer, or $1.20\times$ the expert-matrix work of nominal-matmul Switch. Stored-scalar reduction is therefore not interpreted as an HBM, FLOP, checkpoint-byte, or arithmetic reduction.

The complete descriptive contrasts appear in Appendix B, Table 6.

Table 1: Primary quality and representation results over six seeds. Stored scalars and nominal one-group active scalars follow the analytical ledger; realized $E = 128$ active work is reported in the text. Quality entries are means $\pm$ between-seed SD.

Configuration	E	Stored (M)	Nominal active (M)	GLUE-8 ↑	WikiText PPL ↓
Dense-active	–	102.550	102.550	83.43 ± 0.38	28.72 ± 0.24
Controlled Switch	32	1127.538	135.804	84.82 ± 0.36	24.81 ± 0.19
Switch parameter upper	5	234.811	135.638	84.14 ± 0.39	26.68 ± 0.22
Controlled Switch	64	2185.584	136.001	85.21 ± 0.32	24.42 ± 0.17
OC-Share	32	216.874	103.536	84.28 ± 0.39	25.43 ± 0.23
DyClust-MoE	32	216.874	103.536	84.63 ± 0.34	25.09 ± 0.18
DyClust-MoE	64	364.257	103.560	85.07 ± 0.31	24.66 ± 0.17
Controlled Switch	128	4301.677	136.395	85.54 ± 0.31	24.10 ± 0.11
Switch parameter upper	18	664.643	135.718	84.73 ± 0.35	25.22 ± 0.21
OC-Share	128	659.022	103.610	84.91 ± 0.38	24.78 ± 0.22
D^2-MoE+continuation	128	659.022	103.610	84.95 ± 0.36	24.70 ± 0.22
Decoupled-3State	128	659.022	103.610	85.38 ± 0.28	24.31 ± 0.18
DyClust-MoE	128	659.022	103.610	85.46 ± 0.29	24.29 ± 0.16

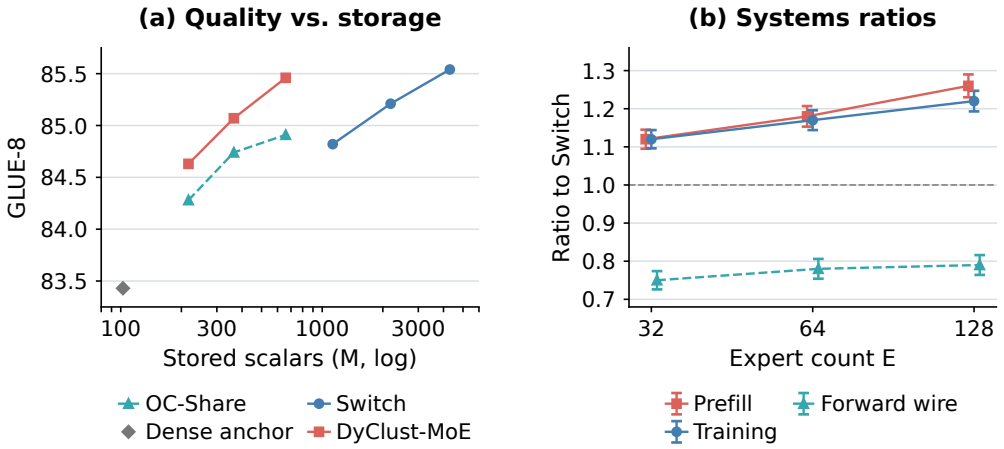


Figure 2: (a) Increasing E moves DyClust-MoE along a substantially smaller stored-representation frontier than controlled same-E Switch; Dense is a descriptive active-capacity anchor. (b) Across $E = 32/64/128$, the paired prefill and full-run-training ratios increase while the forward-wire ratios are $0.75 \pm 0.024/0.78 \pm 0.026/0.79 \pm 0.026$ relative to controlled Switch. Error bars are between-unit SDs.

The 171 k tok/s full-run rate corresponds to 4.87 hours for the declared 3B-token $E = 128$ continuation. With regroup events disabled, the ordinary-update component, excluding separately ledgered fixed-work intervals, reaches 198.5 k tok/s, corresponding to 4.20 hours for the same token count. In the separate four-seed event telemetry, the mean per-seed sum over 99 opportunities is about 35 minutes, while ordinary checkpoints, background repair, terminal fences, and other non-event fixed work contribute about 5.4 minutes. All event durations are monotonic wall-clock intervals inside the full-run ledger. Throughput is computed from unrounded fenced durations before the declared outer-unit reductions; displayed rates, duration summaries, and component totals are rounded only afterward and therefore need not be exact reciprocals or sums.

Serving and scaling. Batch-1 ITL is nearly neutral at 1.01 ± 0.028. Decode throughput is 1.18 ± 0.032, 80%-load goodput is 1.23 ± 0.038, and p99 ITL is 0.82 ± 0.034. Strong scaling at two/four nodes is $1.26 \pm 0.030/1.34 \pm 0.033$ throughput and $0.79 \pm 0.026/0.62 \pm 0.023$ wire. Against

Table 2: Two-node systems results on eight A100 GPUs over six paired outer units. Absolute throughput is the measured aggregate DyClust-MoE rate at $E = 128$; ratios and HBM differences are formed within pair. Switch absolute throughput is not reconstructed from the ratios.

Endpoint	DyClust-MoE aggregate	Paired effect	Unit
Prefill, $E = 128$	438	$1.26 \pm 0.030\times$	k tok/s
Full-run training, $E = 128$	171	$1.22 \pm 0.027\times$	k tok/s
Peak HBM, $E = 128$	–	-0.92 ± 0.11	GB/GPU
Forward wire, $E = 128$	–	0.79 ± 0.026	ratio
F+B MoE wire, $E = 128$	–	0.73 ± 0.027	ratio
All-training wire, $E = 128$	–	0.77 ± 0.031	ratio

Table 3: Frozen identical-assignment interventions at $E = 32$. Each replay changes one execution interface while preserving the accepted assignments or tensor outputs stated in the invariant column. The matched 3B training-control contract is reported in Table 24.

Contrast	Primary effect	Secondary effect	Held fixed	Interface
Fused / materialized	1.09 ± 0.03 end-to-end	–	max logit error $\leq 10^{-5}$	kernel
Hierarchical / score all	1.04 ± 0.02 end-to-end	0.60 ± 0.05 router time	accepted experts	scoring
Optimized / shuffled owners	0.78 ± 0.03 wire	0.88 ± 0.04 critical p95	assignment trace	placement
Coalescing on / off	0.65 ± 0.04 payload	–	owner map	transport

nominal-matmul Switch, OC-Share, D^2-MoE, cached Decoupled-3State, and Tutel, two-node prefill is $1.16/1.18/1.17/1.15/1.18\times$ and F+B wire is $0.74/0.78/0.77/0.80/0.77\times$. Full-run training against cached Decoupled-3State is $1.13 \pm 0.019\times$ (Table 20 in the appendix). Four-node prefill against OC-Share, D^2-MoE, and Decoupled-3State is $1.20/1.19/1.19\times$.

Mechanisms and robustness. At $E = 128$, omitted-mass, selector, and hard-assignment TV are 0.014 ± 0.002, 0.035 ± 0.004, and 0.029 ± 0.004 after token-first reduction; total actual TV is 0.064 ± 0.006 (p95 0.109). A 16,384-token layer repair takes 3.9 ± 0.3 ms, with an 8.1 ± 0.6 ms p99. On the 131,072-token workload, eight-layer repair takes 27.2 ± 1.6 ms; model forward excluding repair takes 267.6 ± 8.0 ms and other runtime work 4.6 ± 0.5 ms, giving 299.4 ± 9.0 ms end-to-end. The four validation seeds are the outer units. Exact-k, zero drops, no mixed epochs, and journal replay remain invariants under the stated failure model.

Routing estimator definitions and admission checks, together with placement and bridge diagnostics, are separated from structural guarantees in Table 10.

Shift behavior. A decision is one prescheduled post-shift grouping opportunity, visiting one layer in round-robin order (Appendix B.6). Across composition, PMC, temporal, and Chinese/Spanish shifts, no-regroup-minus-DyClust-MoE effects are $3.2/3.8/4.2/5.3$ decisions and $0.22/0.28/0.27/0.43$ PPL. One-sided $n = 6, m = 4$ lower bounds are $1.266/1.764/2.266/3.072$ decisions and $0.069/0.094/0.091/0.198$ PPL; every bound is positive. Gap upper bounds are $0.163/0.256/0.223/0.328$, below the prespecified ceilings.

4.1 Public-model transfer

OLMoE-1B-7B (Muennighoff et al., 2025) and Qwen1.5-MoE-A2.7B (Qwen Team, 2024) each use ten independently continued compact/native pairs, with 4B primary token positions per branch and matched data and optimizer schedules. Evaluation follows this continuation under native routing-weight semantics. OLMoE preserves its all-64 softmax and unnormalized top-8 weights; Qwen

Table 4: Native-routing transfer over ten compact/native pairs per model. Quality cells are compact minus native; systems cells are compact/native.

Model / native routing	Worst task difference	PPL Δ	Prefill	Decode	F+B wire
OLMoE-1B-7B / top-8	MMLU -0.18 ± 0.27	$+0.12 \pm 0.06$	1.17 ± 0.031	1.14 ± 0.029	0.77 ± 0.029
Qwen1.5-MoE-A2.7B / top-4	CMMLU -0.19 ± 0.25	$+0.14 \pm 0.06$	1.15 ± 0.028	1.13 ± 0.027	0.78 ± 0.025

preserves unnormalized native top-4 weights and its shared expert. All expert logits remain evaluated. Absolute quality scores, inventories, margins, and HBM results appear in Appendices F and G; the transfer conclusions concern those knowledge, language-understanding, and perplexity endpoints.

4.2 Quality context and secondary results

At $E = 128$, GLUE-8 is $85.54/85.46$ and WikiText PPL is $24.10/24.29$ for controlled Switch/DyClust-MoE; paired differences are -0.08 ± 0.10 points and $+0.19 \pm 0.08$ PPL (Table 6).

Appendix B.3 reports descriptive paired capability differences and the task-level GLUE panel; Tables 31 and 33 give absolute public-model scores.

Decoupled-3State uses epoch-cached expert-ID tuples and the same fused dispatcher. It attains a lower reconstruction objective ($0.957 \pm 0.012\times$), routing bound (-0.007 ± 0.003), and placement proxy ($0.979 \pm 0.009\times$), but touches 1.54 versus 1.23 bases and 1.47 versus 1.19 owners/token. Cached lookup costs $0.9 \pm 0.2\%$ of MoE-layer time; the uncached diagnostic costs $5.8 \pm 0.7\%$. With assignments and owners held identical, the shared/cached execution ratio is only 1.006 ± 0.005. The larger end-to-end gain therefore reflects the layouts and accepted routes induced by joint state, not a repeated map-join penalty.

Full-run accounting includes the complete continuation budget, accepted and rejected regrouping, refits, shadow steps, bridges, repair, placement, journaling, synchronization, and checkpoints. Wire and memory counters distinguish edges, payload, egress, byte-hops, inference HBM, and transaction HBM. Across 720 assignment-stress units and 480 injected transactions, all listed assignment, epoch, cursor/RNG/token-hash, journal, and optimizer-update violations were zero. These checks are conditional on the sampled grid, stated faults, and durable-quorum/CAS assumptions; Tables 21 and 28 give the complete boundaries and cases.

5 Conclusion

DyClust-MoE coordinates expert representation, routing, and placement through one versioned grouping state. At $E = 128$, stored scalars fall by 84.68% relative to controlled Switch, while peak inference HBM falls by 0.92 ± 0.11 GB/GPU. Paired GLUE/PPL differences are $-0.08 \pm 0.10/+0.19 \pm 0.08$; prefill/full-run training reach 438 k/171 k tokens/s. Against cached Decoupled-3State, their ratios are $1.15\times/1.13\times$ and F+B wire is $0.80\times$. The matched controls and frozen replays connect these gains to the bases, owners, and messages induced by shared grouping.

The findings concern the evaluated workloads and topologies. Equal-size groups, fixed rank, local search, and frozen-routing bridges limit generalization. Exact-k feasibility and crash safety retain their stated capacity and durability assumptions; deployment requires validated version publication and rollback. Adaptive ranks and groups are natural extensions.

AI USE STATEMENT

No generative AI or large language model tools were used in conducting this research or preparing this manuscript. The research design, implementation, data analysis, and manuscript writing and revision were carried out by the authors. The authors take full responsibility for the content of this paper.

REFERENCES

Ravindra K. Ahuja, Thomas L. Magnanti, and James B. Orlin. *Network Flows: Theory, Algorithms, and Applications*. Prentice Hall, 1993. ISBN 9780136175490.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. PIQA: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 7432–7439, 2020. doi: 10.1609/aaai.v34i05.6239.

Shiyi Cao, Shu Liu, Tyler Griggs, Peter Schafhalter, Xiaoxuan Liu, Ying Sheng, Joseph E Gonzalez, Matei Zaharia, and Ion Stoica. MoE-Lightning: High-throughput MoE inference on memory-constrained GPUs. In *Proceedings of the 30th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 1*, pp. 715–730, 2025. doi: 10.1145/3669940.3707267.

Inesh Chakrabarti, Sourjya Roy, Bowen Bao, Thiago Crepaldi, Spandan Tiwari, and Ashish Sirasao. Shape mutating expert compression: LorExperts and BTExperts. *arXiv preprint arXiv:2608.07814*, 2026. URL https://arxiv.org/abs/2608.07814.

Moses Charikar, Kevin Chen, and Martin Farach-Colton. Finding frequent items in data streams. In *Automata, Languages and Programming (ICALP)*, volume 2380 of *Lecture Notes in Computer Science*, pp. 693–703. Springer, 2002. doi: 10.1007/3-540-45465-9_59.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Damai Dai, Li Dong, Shuming Ma, Bo Zheng, Zhifang Sui, Baobao Chang, and Furu Wei. StableMoE: Stable routing strategy for mixture of experts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7085–7095, 2022. doi: 10.18653/v1/2022.acl-long.489.

Damai Dai, Chengqi Deng, Chenggang Zhao, R.X. Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Y. Wu, Zhenda Xie, Y.K. Li, Panpan Huang, Fuli Luo, Chong Ruan, Zhifang Sui, and Wenfeng Liang. DeepSeekMoE: Towards ultimate expert specialization in mixture-of-experts language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1280–1297, 2024. doi: 10.18653/v1/2024.acl-long.70.

Nan Du, Yanping Huang, Andrew M Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, Barret Zoph, Liam Fedus, Maarten P Bosma, Zongwei Zhou, Tao Wang, Emma Wang, Kellie Webster, Marie Pellat, Kevin Robinson, Kathleen Meier-Hellstern, Toju Duke, Lucas Dixon, Kun Zhang, Quoc Le, Yonghui Wu, Zhifeng Chen, and Claire Cui. GLaM: Efficient scaling of language models with mixture-of-experts. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 5547–5569. PMLR, 2022. URL https://proceedings.mlr.press/v162/du22c.html.

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://jmlr.org/papers/v23/21-0998.html.

Elias Frantar and Dan Alistarh. QMoE: Sub-1-bit compression of trillion parameter models. *Proceedings of Machine Learning and Systems*, 6:439–451, 2024.

Trevor Gale, Deepak Narayanan, Cliff Young, and Matei Zaharia. MegaBlocks: Efficient sparse training with mixture-of-experts. *Proceedings of Machine Learning and Systems*, 5:288–304, 2023.

Hao Gu, Wei Li, Lujun Li, Qiyuan Zhu, Mark G. Lee, Shengjie Sun, Wei Xue, and Yike Guo. Delta decompression for MoE-based LLMs compression. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 20497–20514. PMLR, 2025. URL https://proceedings.mlr.press/v267/gu25c.html.

Nathan Halko, Per-Gunnar Martinsson, and Joel A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53 (2):217–288, 2011. doi: 10.1137/090771806.

Song Han, Jeff Pool, John Tran, and William Dally. Learning both weights and connections for efficient neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 28, pp. 1135–1143, 2015.

Chaoyang He, Shuai Zheng, Aston Zhang, George Karypis, Trishul Chilimbi, Mahdi Soltanolkotabi, and Salman Avestimehr. SMILE: Scaling mixture-of-experts with efficient bi-level routing. *arXiv preprint arXiv:2212.05191*, 2022a.

Jiaao He, Jidong Zhai, Tiago Antunes, Haojie Wang, Fuwen Luo, Shangfeng Shi, and Qin Li. FasterMoE: Modeling and optimizing training of large-scale dynamic pre-trained models. In *Proceedings of the 27th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming*, pp. 120–134, 2022b. doi: 10.1145/3503221.3508418.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021.

Maurice P. Herlihy and Jeannette M. Wing. Linearizability: A correctness condition for concurrent objects. *ACM Transactions on Programming Languages and Systems*, 12(3):463–492, 1990. doi: 10.1145/78969.78972.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. C-Eval: A multi-level multi-discipline chinese evaluation suite for foundation models. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

Itay Hubara, Matthieu Courbariaux, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. Quantized neural networks: Training neural networks with low precision weights and activations. *Journal of Machine Learning Research*, 18(187):1–30, 2018. URL https://jmlr.org/papers/v18/16-456.html.

Changho Hwang, Wei Cui, Yifan Xiong, Ziyue Yang, Ze Liu, Han Hu, Zilong Wang, Rafael Salas, Jithin Jose, Prabhat Ram, HoYuen Chau, Peng Cheng, Fan Yang, Mao Yang, and Yongqiang Xiong. Tutel: Adaptive mixture-of-experts at scale. *Proceedings of Machine Learning and Systems*, 5: 269–287, 2023.

Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2704–2713, 2018. doi: 10.1109/CVPR.2018.00286.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

Peng Jin, Bo Zhu, Li Yuan, and Shuicheng Yan. MoE++: Accelerating mixture-of-experts methods with Zero-Computation experts. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://iclr.cc/virtual/2025/poster/28078.

Taku Kudo and John Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 66–71, 2018. doi: 10.18653/v1/D18-2012.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. GShard: Scaling giant models with conditional computation and automatic sharding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=qrwe7XHTmYb.

Mike Lewis, Shruti Bhosale, Tim Dettmers, Naman Goyal, and Luke Zettlemoyer. BASE layers: Simplifying training of large, sparse models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 6265–6274. PMLR, 2021. URL https://proceedings.mlr.press/v139/lewis21a.html.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. CMMLU: Measuring massive multitask language understanding in chinese. *arXiv preprint arXiv:2306.09212*, 2023.

Lujun Li, Qiyuan Zhu, Jiacheng Wang, Xiaoyu Qin, Wei Li, Hao Gu, Sirui Han, and Yike Guo. Sub-MoE: Efficient mixture-of-expert LLMs compression via subspace expert merging. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(27):22994–23002, 2026. doi: 10.1609/aaai.v40i27.39464.

Wei Li, Lujun Li, Hao Gu, You-Liang Huang, Mark G. Lee, Shengjie Sun, Wei Xue, and Yike Guo. MoE-SVD: Structured mixture-of-experts LLMs compression via singular value decomposition. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 35209–35230. PMLR, 2025. URL https://proceedings.mlr.press/v267/li25az.html.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Byj72udxe.

Chengxi Min, Wei Wang, Yahui Liu, Weixin Ye, Enver Sangineto, Qi Wang, and Yao Zhao. Guiding the experts: Semantic priors for efficient and focused MoE routing. *arXiv preprint arXiv:2505.18586*, 2025.

Niklas Muennighoff, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Jacob Morrison, Sewon Min, Weijia Shi, 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://iclr.cc/virtual/2025/poster/27777.

Xiaonan Nie, Xupeng Miao, Zilong Wang, Zichao Yang, Jilong Xue, Lingxiao Ma, Gang Cao, and Bin Cui. FlexMoE: Scaling large-scale sparse pre-trained model training via dynamic device placement. *Proceedings of the ACM on Management of Data*, 1(1):1–19, 2023. doi: 10.1145/3588964.

Qwen Team. Qwen1.5-MoE: Matching 7b model performance with 1/3 activated parameters, 2024. URL https://qwenlm.github.io/blog/qwen-moe/.

Samyam Rajbhandari, Conglong Li, Zhewei Yao, Minjia Zhang, Reza Yazdani Aminabadi, Ammar Ahmad Awan, Jeff Rasley, and Yuxiong He. DeepSpeed-MoE: Advancing mixture-of-experts inference and training to power next-generation AI scale. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 18332–18346. PMLR, 2022. URL https://proceedings.mlr.press/v162/rajbhandari22a.html.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=B1ckMDqlg.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Evan Walsh, Luke Zettlemoyer, Noah Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. Dolma: An open corpus of three trillion tokens for language model pretraining research. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15725–15788, 2024. doi: 10.18653/v1/2024.acl-long.840.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. doi: 10.1016/j.neucom.2023.127063.

Mengyang Sun, Yihao Wang, Tao Feng, Dan Zhang, Yifan Zhu, and Jie Tang. A stronger mixture of low-rank experts for fine-tuning foundation models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 57712–57727. PMLR, 2025. URL https://proceedings.mlr.press/v267/sun25s.html.

Shawn Tan, Yikang Shen, Rameswar Panda, and Aaron Courville. Scattered mixture-of-experts implementation. *arXiv preprint arXiv:2403.08245*, 2024. URL https://arxiv.org/abs/2403.08245.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 353–355. Association for Computational Linguistics, 2018. doi: 10.18653/v1/W18-5446.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4791–4800, 2019. doi: 10.18653/v1/P19-1472.

Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent 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, pp. 7103–7114, 2022. doi: 10.52202/068431-0515.

Peijun Zhu, Ning Yang, Baoliang Tian, Jiayu Wei, Weihao Zhang, Haijun Zhang, and Pin Lv. Breaking the MoE LLM trilemma: Dynamic expert clustering with structured compression. *arXiv preprint arXiv:2510.02345*, 2025. URL https://arxiv.org/abs/2510.02345.

A DETERMINISTIC GROUPING AND PROOFS

A.1 CANDIDATE-GENERATION PSEUDOCODE

Grouping decisions occur at 99 evenly spaced, prespecified cursor positions through the continuation and visit one MoE layer in round-robin order, leaving a subsequent in-budget primary update for the bridge. Before continuation, immutable one-indexed expert IDs define the balanced incumbent $g^{(0)}(i) = 1 + ((i-1) \bmod G)$. Set every $R_{i,p} = 0$, apply Equation (5) once, and then alternate Equations (4) and (5); every flow candidate and trial swap repeats this cold start, so the eight-sweep limit is not path dependent. Every later event uses the last committed labels. The most recent 2^{20} primary token positions preceding each event form a fixed history window. Document hashes split it into a $7/8$ proposal view and a $1/8$ proposal-disjoint screening view; activation rereads are charged to the auxiliary ledger. Because both views come from training history, the screening view is not an independent training holdout; it is used only to screen the event candidate. Both views are disjoint from the later router-shadow batch, bridge batch, and post-commit audit blocks. The candidate procedure is:

1. On the proposal view, compute W^{pre}, ω, exposure counts, centroids, and unordered pair counts in fixed token/expert order.
2. Build the 256-bin signed CountSketch $\phi_{i,p}$ and scaled feature x_i exactly as defined in the main text.
3. Select prototype IDs by farthest-first, with smallest expert ID as the first seed and lexicographic ties; set $s^-_{g,p} = \phi_{i_g,p}$ and $c^-_g = u_{i_g}$, and freeze incumbent labels.
4. Form C_{ig} from Equation (3) and solve source–expert–group–sink min-cost flow with unit expert capacity and group capacity K. Select the lexicographically smallest assignment vector among equal-cost optima, then recompute centroids and the exact pair cut.
5. From one cold start, run at most two probe sweeps. The pre-screen retains a finite candidate only if probe reconstruction RMS is at most 0.06 and its maximum expert-relative error is at most 0.20. A failure stops the event before the complete scan; its monotonic wall-clock interval is still charged to the run ledger.
6. For a surviving proposal, apply the cold-start refit until relative reconstruction decrease is at most 10^{-5}; reject a candidate that reaches eight sweeps first. For $M \in \mathbb{R}^{m\times n}$, sort FP64 singular triplets by decreasing unrounded σ. Only bitwise-equal positive singular values form a repeated block; near but unequal values are never rotated. Within a repeated block, project $e_1, \ldots, e_n$ into its right singular subspace and use two-pass fixed-order modified Gram–Schmidt, accepting residual norm $> 10^{-12}$ until the block dimension is reached. Make the largest-magnitude entry of each v positive (smallest-index tie), set $u = Mv/\sigma$, and apply the same sign rule to simple triplets. Concatenate blocks in singular-value order and take the first r nonzero triplets; a cutoff inside an equal-value block uses this canonical within-block order and remains optimal. Finally set $A = U_r\Sigma_r^{1/2}$ and $B = V_r\Sigma_r^{1/2}$, with unused zero columns set to zero.
7. Recompute surrogate costs using the current base sketches, centroids, and labels. Shortlist at most 32 cross-group pairs with negative surrogate swap cost, ordered by predicted decrease and then immutable IDs. Each surviving pair receives a complete exact refit. A swap is eligible only when $\mathcal{J}^{\mathrm{old}}_{\mathrm{grp}} > 10^{-6}$ and its decrease is at least $\max(0.005\,\mathcal{J}^{\mathrm{old}}_{\mathrm{grp}}, 10^{-7})$. Restart after an accepted swap.
8. Stop after a no-improvement pass over the shortlisted pairs or $8E$ exact swap evaluations. Evaluate the retained event candidate against $b_{\ell,e} \le 0.03$, $\max_{i,p} \tau^{(e)}_{i,p} \le 0.12$, $B_\ell + b_{\ell,e} \le 0.10$, screening-view $D_{\mathrm{out}} \le 0.02$, and the shadow gates defined in Section A.2; any failure retains the incumbent. Subsequent post-commit audit blocks are never used for candidate selection.

The sketch byte domain is exact: the 17 ASCII bytes `dyclust-sketch-v1` are followed, without a terminator, by five unsigned 64-bit little-endian fields for run seed, zero-based layer, zero-based event, projection code, and zero-based coordinate. Projection codes are gate $= 0$, up $= 1$, and down $= 2$; the small model instantiates codes 1 and 2, whereas both public models instantiate all

three. Let $D(a)$ be the SHA-256 digest of this byte string. The bin index is $H(a) = D(a)[0]$, equivalently $\mathrm{uint64}_{\mathrm{LE}}(D(a)[0{:}8]) \bmod 256$, and the sign is $\sigma(a) = (-1)^{\mathrm{lsb}(D(a)[31])}$. This specifies a reproducible signed projection; full reconstruction refits govern acceptance. The assignment step is a proposal heuristic; the exact objective after refitting is the only quantity used for acceptance. For a numerical $E = 4, G = 2, K = 2$ example, let rows be experts and columns be groups:

$$C = \begin{bmatrix} 0.10 & 0.90 \\ 0.20 & 0.70 \\ 0.80 & 0.15 \\ 0.75 & 0.25 \end{bmatrix}.$$

The unique balanced minimum assigns experts $\{1, 2\}$ to group 1 and $\{3, 4\}$ to group 2, with proposal cost 0.70. Let $(\chi_{12}, \chi_{13}, \chi_{14}, \chi_{23}, \chi_{24}, \chi_{34}) = (0.4918, 0.0328, 0.0164, 0.0328, 0.0164, 0.4098)$, whose six entries sum to 1.0000. The incumbent raw pair cut is 0.0984, so its contribution to Equation (2) at $\lambda_{\mathrm{pair}} = 0.10$ is 0.00984. The $1/2$ in Equation (3) makes the summed linear surrogate charge this once rather than twice. Swapping experts 2 and 3 gives a recomputed raw cut of 0.9508 and a weighted objective contribution of 0.09508. The swap is nevertheless decided only after recomputing both centroids and all exact base/SVD terms.

Proposition A.1 (Continuous refit property)**.** *For a fixed partition and fixed bases, Equation* (4) *minimizes the weighted rank-r residual error. For fixed residuals and labels, Equation* (5) *minimizes the weighted base least-squares error.*

Proof. The first statement is Eckart–Young–Mirsky applied independently to each residual matrix. Differentiating the weighted quadratic in the base gives Equation (5). Neither statement implies a global optimum for the discrete partition search. □

Initialization geometry. The co-routing row in x_i describes an expert's co-routing neighborhood. For symmetric χ with zero diagonal, its contribution to a pairwise squared distance is exactly

$$\frac{\lambda_{\mathrm{pair}}}{2} \|\chi_{i,:} - \chi_{j,:}\|_2^2 = \lambda_{\mathrm{pair}} \chi_{ij}^2 + \frac{\lambda_{\mathrm{pair}}}{2} \sum_{a \notin \{i,j\}} (\chi_{ia} - \chi_{ja})^2. \tag{18}$$

Direct co-routing can therefore increase separation in the seed geometry even when the pair-cut objective favors co-location. Farthest-first selects prototypes; the balanced assignment and subsequent exact refits determine the retained partition. Its guarantee is conditional acceptance against the stated objective and gates, with the initialization remaining a local-search heuristic.

Computational cost. Let N be the history-window token count, d the activation width, $B = 256$ the sketch width, and let each of the P_W expert projections have dimensions at most $m \times n$. With k accepted edges per token, exposure centroids and pair counts require $O(Nkd + Nk^2)$ arithmetic. Building dense weight sketches costs $O(P_W Emn)$. Farthest-first distances over the concatenated features cost $O(EG(P_W B{+}d{+}E))$; the flow graph has $E{+}G{+}2$ vertices and $O(EG{+}E{+}G)$ edges. Writing $T_{\mathrm{flow}}(E, G)$ for its exact solve including lexicographic ties avoids assuming a particular flow implementation (Ahuja et al., 1993).

Let $C_{\mathrm{SVD}}(m, n)$ denote the specified full-SVD cost, $S \le 8$ the refit sweep count, and $Q \le 1 + 8E$ the initial candidate plus exact swap trials. A conservative bound for the refit and objective work is

$$O\big(Q\left[SP_W E\{C_{\mathrm{SVD}}(m, n) + mn\} + Ed + E^2\right]\big),$$

in addition to statistics, proposal construction, and T_{flow}. For a fixed number of Jacobi sweeps, dense full SVD has the usual $O(mn \min(m, n))$ arithmetic scale; retaining rank r reduces factor storage to $O(P_W Er(m + n))$, while the acceptance refit still uses the full decomposition. Explicit co-routing and flow state require $O(E^2 + EG)$ storage, beyond weights, history, and backend workspaces. The sweep and trial caps bound admitted work; batching determines realized wall-clock cost. Table 13 reports that cost separately from these operation counts. FP64 and the pinned reduction and sign rules define the numerical reproducibility contract.

A.2 EVENT DISTORTION GATES

The event ledger is

$$b_{\ell,e} = \sqrt{\frac{1}{P_W E}\sum_{i,p} d_{i,p}^{(e)}}, \qquad B_\ell = \sum_e b_{\ell,e},$$
$$d_{i,p}^{(e)} = \frac{\left\|\Delta_{i,p}^{(e)}\right\|_F^2}{\left\|W_{i,p}^{\mathrm{pre},(e)}\right\|_F^2 + \epsilon_w}. \tag{19}$$

Here $\Delta^{(e)} = W^{\mathrm{pre},(e)} - \widetilde{W}^{(e)}$. The first quantity is the event RMS displacement and the second charges repeated refactorizations in a layer. Before updating the ledger, admission requires $b_{\ell,e} \le 0.03$, $\max_{i,p} \tau_{i,p}^{(e)} \le 0.12$, $B_\ell + b_{\ell,e} \le 0.10$, and proposal-disjoint screening-view relative FFN-output RMS (including nonlinearity and bias under frozen routing)

$$D_{\mathrm{out}} = \sqrt{\frac{\sum_t \left\|y_t^+ - y_t^-\right\|_2^2}{\sum_t \left\|y_t^-\right\|_2^2 + 10^{-12}}} \le 0.02.$$

The shadow preflight additionally requires KL ≤ 0.02, retained flat mass ≥ 0.98, and flat top-k recall ≥ 0.99. Only after all gates pass do we set $B_\ell \leftarrow B_\ell + b_{\ell,e}$.

A.3 MASS AND RESERVE PROOFS

Proposition A.2 (Supported-mass decomposition)**.** *If the flat and within-group routers use identical expert logits, restriction to support $\mathcal{S}$ omits teacher mass ε, and the accepted set has nonzero supported mass, then Equation* (10) *holds exactly with the three displayed total-variation terms.*

Proof. Identical logits give $m_{g(i)}\rho_{i|g(i)} = p_i^{\mathrm{flat}}$ before the learned selector is substituted. Renormalizing a support that omits probability mass ε changes total variation by exactly ε. Applying the triangle inequality from the restricted teacher to $q^{\mathcal{S}}$ and then to $\widehat{q}$ yields Equation (10). The proof does not bound the learned selector or hard-assignment terms; both remain empirical. □

Proposition A.3 (Drop-free reserve feasibility)**.** *If $k \le E$ and every expert capacity satisfies $C_{i\delta} \ge \lceil kN_\delta/E \rceil$, the cyclic reserve is a feasible integral assignment with exactly k distinct experts per token. Every feasibility-checked batch committed by the bounded repair remains feasible.*

Proof. The stable ordinals ν_t are consecutive across source ranks. The kN_δ one-indexed reserve IDs $1 + ((k\nu_t + j) \bmod E)$ differ in load by at most one; adding one only relabels the residue classes, while $k \le E$ makes the k IDs of one token distinct. Hence the initial flow satisfies token degree and expert capacity. Before a repair batch is committed, its net edge deltas are checked for exactly k distinct edges per token and loads no larger than $C_{i\delta}$. Therefore a committed batch maps one feasible assignment to another. Induction over the bounded batch sequence completes the claim. Score optimality does not follow. □

The bounded repair starts from the cyclic reserve with costs $c_{ti} = -\log(q_{ti}^{\mathcal{S}} + 10^{-12})$. It considers a single-token edge replacement and a two-token exchange; longer augmenting paths are excluded. For each ordered expert pair (i, j), retain the eight lowest-cost proposals $(t, i \to j)$ with i accepted and j allowed but unaccepted, breaking ties by token ordinal. Direct replacements maximize the positive gain over each token's accepted origins and allowed unaccepted destinations with spare capacity; the eight-proposal bucket limit applies only to exchange-partner search. Opposite buckets supply two-token exchanges; their total cost decrease must exceed 10^{-12}. Merge eligible direct moves and exchanges in decreasing gain, then lexicographic token/expert order. A stable prefix reservation accepts a move only if its tokens are unused in the batch and the net expert loads remain feasible. Accepted moves are committed together; the next round rebuilds candidates from the updated assignment. Every evaluation round counts toward $R_{\max} = 16$, including a round with no commit. An empty admissible batch terminates repair. Degree, uniqueness, and capacity are

preserved at every commit; the cap retains the last feasible assignment. Final assignment and delta-list hashes agree across ranks before dispatch. This bounded local procedure does not claim a globally minimum-cost assignment.

A.4 PLACEMENT AND PUBLICATION

The objective in Equation (11) minimizes the critical-owner variables constrained in Equation (15). The latter charges one launch intercept per nonempty source–owner message and one payload per destination-unique token. Branch-and-bound returns a feasible integer map no worse than the incumbent for this proxy; held-out traces determine the elapsed-communication comparison.

For each source–owner link class, a fit set runs 200 round-trip probes at 4 KiB, 64 KiB, 1 MiB, and 8 MiB using the runtime serialization and collective path. Their medians are fitted by deterministic nonnegative least squares to $\alpha_{sr} + \beta_{sr} m$; the active-set solve stops at relative objective change below 10^{-12} and breaks equal objectives by smaller (α, β). A disjoint calibration-validation probe set at the same sizes supplies the $R^2 \geq 0.95$ and p95-relative-error $\leq 5\%$ enablement gate. A third, request-disjoint solver-calibration trace supplies a_{ctsg} to the fixed-charge objective. Fit probes, validation probes, solver-calibration requests, and final evaluation requests are four mutually disjoint roles frozen before outcomes; only the last is used for reported optimized-versus-shuffled systems comparisons. A failed validation gate retains the incumbent owner map. Placement pins Gurobi 11.0.3 with one thread, seed 0, a 2.0-s time limit, `MIPGap`$= 10^{-4}$, and feasibility/integrality tolerances 10^{-9}. A second deterministic solve implements lexicographic owner-vector ties. Across primary $E \leq 128$ events, the reported relative gap has median/p95/maximum $0/1.2\times10^{-5}/3.7\times10^{-5}$; no solve reaches the time limit, and an incumbent is published only if its proxy is no worse than the previous owner map.

New group IDs are the lexicographic order of their sorted expert-ID tuples. Let $\omega_{gh} = |\mathcal{G}_g^+ \cap \mathcal{G}_h^-|/K$, so $\sum_h \omega_{gh} = 1$. The candidate selector is initialized coordinatewise by $(U_g^+, a_g^+) = \sum_h \omega_{gh}(U_h^-, a_h^-)$; the bias coordinate is absent in the bias-free public-model adapters. Private first- and second-moment buffers for both selector weights and biases start at zero, accumulate only the 64 shadow updates, and are published only on success. Semantic IDs common to both epochs are embeddings, attention and normalization tensors, task heads, expert-ID-keyed router rows, and retained expert biases. Old full expert matrices, new bases/factors, group-selector rows, and owner maps are epoch-private. Both primary branches clone the common optimizer state; success publishes the candidate branch and its private selector state, while abort publishes the old branch and destroys every candidate-private buffer.

The data ledger separates three quantities that are not interchangeable. The primary cursor advances over the declared N_{primary} token positions per branch—3B for a primary small-model continuation—and optimizer/scheduler state advances once for each scheduled minibatch. If B_e is the ordinary batch at event e, auxiliary rereads are $\sum_{e\in\mathcal{S}}(7/8)B_e + \sum_{e\in\mathcal{B}} 2B_e$: the shadow subset, one additional gradient-bearing bridge branch, and one forward-only immediate-swap sidecar. Gradient-bearing token–branch evaluations are $N_{\text{primary}} + \sum_{e\in\mathcal{S}}(7/8)B_e + \sum_{e\in\mathcal{B}} B_e$. Sidecars carry no gradient; sealed assessments are fresh document-disjoint reads rather than rereads, so they are recorded in a separate audit-exposure ledger and excluded from both formulas. Every auxiliary reread reuses an already scheduled ordinary-batch ID and never advances the cursor. Thus comparators match primary token positions and published updates, while full-run time, wire, and energy include the method's explicitly larger auxiliary compute.

For bridge slice s, let d_s^{NLL} be the token mean of candidate-minus-old NLL on the same microbatch, and let d_s^{PPL} be the absolute PPL gap between the gradual candidate and the non-updating immediate-swap sidecar on that microbatch. The two path gates are $\max_s d_s^{\text{NLL}} < 0.020$ and $\max_s d_s^{\text{PPL}} < 0.25$; neither reducer pools slices or treats tokens as replicates. The old branch is the same-slice, same-randomness endpoint, not a full-batch marginal average.

Publication binds tensors, router, owner map, optimizer state, data cursor, RNG, and byte ledger in one durable record. Each immutable generation advances through FENCED, SHADOW-READY, SLICE-s, UPDATE-READY, and one terminal state, COMMITTED or ABORTED. A length-prefixed snapshot stores this phase, candidate and parent epochs, layer, completed-slice index, cursor and RNG counters, branch-loss scalars, complete accumulator tensors with dtype/shape/byte length and

SHA-256 hashes, output-buffer hashes, and an all-rank completion bitmap. Updates are written to new generation-addressed buffers, never in place; ranks fsync tensor blocks before the record and fsync the containing directory before acknowledging a quorum-complete snapshot. Recovery replays only an incomplete slice plus the suffix; an UPDATE-READY generation is finalized without reapplying its optimizer update, and only FENCED at $s = 0$ replays the full batch. The compare-and-swap changes one epoch pointer after all hashes verify. Epoch-tagged requests reject stale records.

Proposition A.4 (Finite bridge increments)**.** *For fixed old and candidate endpoint tensors, adjacent private bridge states satisfy* $\|W_{i,p}^{(s+1)} - W_{i,p}^{(s)}\|_F = \|W_{i,p}^{\mathrm{pre}} - \widetilde{W}_{i,p}^{+}\|_F/H$ *for every expert and projection.*

Proof. Subtract the interpolation formula in Section 3.4 at consecutive slice indices. The common candidate endpoint cancels and the coefficient of the old-minus-candidate offset changes by exactly $1/H$. This is a parameter path identity only; it is not a bound on nonlinear loss or post-activation optimization. □

Proposition A.5 (Crash-safe epoch visibility)**.** *Assume immutable generation-addressed buffers, complete-record hash validation, epoch-tagged requests, at most one in-flight candidate, and a linearizable activation compare-and-swap. Then a request observes either the complete parent epoch or the complete candidate epoch, never a mixture of the two.*

Proof. Before the compare-and-swap, the only published pointer addresses the complete parent. Candidate writes cannot change parent buffers and an incomplete record cannot pass validation. The compare-and-swap is the single linearization point and occurs only after the candidate record and all referenced buffers validate. After it, epoch-tagged requests resolve only the candidate; stale parent-tagged requests are rejected. Abort leaves the parent pointer unchanged. The claim does not establish liveness and is conditional on the stated durable-quorum assumption. □

B Full experimental protocol

B.1 Models, data, and optimization

The small model has a tied 32,000-token SentencePiece vocabulary (Kudo & Richardson, 2018), 12 pre-LayerNorm blocks ($\epsilon = 10^{-5}$), width 768, 12 attention heads, RoPE positions (Su et al., 2024), bias-retaining two-matrix GELU FFNs of width 2688, and MoE layers $\{2, 3, 5, 6, 8, 9, 11, 12\}$. Sparse upcycling clones each affected dense FFN into E experts, adds counter-seeded rank-r perturbations of relative Frobenius norm 10^{-3} and expert-router rows $v_i \sim \mathcal{N}(0, 10^{-6}I/d)$ with bias $b_i = 0$, and gives every same-E comparator the same materialized update-0 expert tensors and router rows. For the initial balanced groups, selector row $U_g = K^{-1}\sum_{i\in\mathcal{G}_g} v_i$ and selector bias $a_g = \log K + K^{-1}\sum_{i\in\mathcal{G}_g} b_i$; subsequent accepted regroupings use the overlap remap in Section A. Sparse rows use top-2 routing, $K = 8$, and rank $r = 8$. The 102.550M-scalar dense parent is trained for 100B non-padding token instances from a fixed Dolma-v1.7-derived five-stratum manifest (60% filtered web, 15% permissively licensed code, 10% scholarly text, 10% books, and 5% encyclopedic/reference text) (Soldaini et al., 2024). The disjoint continuation manifest preserves these proportions and excludes exact normalized-record matches or documents covering at least 70% of the distinct normalized 13-grams of any evaluation record; records shorter than 13 tokens use exact normalized-string matching.

Each of the 24 non-Dense configurations in the 432B primary-quality ledger receives 3B non-padding continuation token positions at sequence length 2,048 in each of six seeds; the descriptive $E = 256$ extension receives 2B as stated in Section H. Forward/backward tensors are BF16; master weights and AdamW moments are FP32 (Loshchilov & Hutter, 2019). The selected continuation recipe is AdamW $(\beta_1, \beta_2, \epsilon) = (0.9, 0.95, 10^{-8})$, peak learning rate 2×10^{-4} after a 30M-token (1%) linear warm-up, cosine decay to zero, matrix-only weight decay 0.10, gradient clip 1.0, and zero dropout/stochastic depth. The global batch is 262,144 tokens/update with gradient accumulation 2, giving approximately 11,444 optimizer updates over 3B tokens. Each of the 24 primary configurations evaluates six candidates on two pilot seeds for 2B tokens per run, totaling 576B pilot tokens; selection uses held-out token NLL only. Six disjoint confirmatory seeds supply the primary estimates; the pilots never enter reported results. Controlled same-E Switch, parameter brackets, nominal-matmul

Switch, OC-Share, GlobalBase-LR, StaticShared, and D^2-MoE share parent, token order, active k, precision, primary token positions, checkpoint cadence, and selection budget within each declared scale.

Table 5: Experiment-token ledger. Each branch budget counts primary token positions, and reused references are charged once to their originating block. Auxiliary reads, failed reruns, and verification beyond the named validation block are covered by the stated reserve.

Experiment block	Construction	Budget
Primary quality	24 non-Dense configurations $\times$6 seeds $\times$3B	432B
Full factorial	32 cells $\times$4 seeds $\times$2.5B	320B
OLMoE transfer	10 pairs $\times$2 branches $\times$4B	80B
Qwen transfer	10 pairs $\times$2 branches $\times$4B	80B
$E = 256$ extension	2 configurations $\times$6 seeds $\times$2B	24B
Distribution shifts	4 shifts $\times$6 seeds $\times$3 branches $\times$1B	72B
Pilot selection	24 configurations $\times$6 candidates $\times$2 seeds $\times$2B	576B
Mechanisms and ablations	6 variants $\times$4 seeds $\times$3B	72B
Bridge/repair/placement validation	8 scenarios $\times$4 seeds $\times$1B	32B
Sensitivity surface	11 configurations $\times$5 seeds $\times$3B	165B
Downstream training/evaluation	GLUE fine-tuning and evaluation exposure	$\approx$ 4.6B
Measured-result subtotal	–	$\approx$ 1857.6B
Rerun, auxiliary-read, and verification reserve	–	52B
Allocated program including reserve	–	$\approx$ 1909.6B
Dense parent pretraining	One shared parent	100B
Total	–	$\approx$ **2009.6**B $\approx$ **2.01**T

Comparator contracts. Controlled same-E Switch preserves the expert count; the parameter-upper bracket is the smallest integer full-expert count whose stored inventory is at least ours. Nominal-matmul Switch uses the nearest tensor-core-aligned width, 1,408, selected once by a parent-only FFN morph; its top-2 expert multiply-add inventory is 2.03% above ours rather than exactly equal. Every small-model Switch or structured exact-k control uses the same capacity domains, factor 1.25, reserve/repair implementation, and zero-drop rule; only its representation, scored support, router, or placement varies as named. OC-Share uses capacity-constrained output clustering, $K = 8$, rank 8, and our unchanged base-plus-residual layout under the same canonical byte cap. It retains all expert identities and a flat router. This is an output-cluster sharing control, not the expert-merging algorithm of Sub-MoE (Li et al., 2026); GlobalBase-LR replaces group bases with one global base and takes the largest integer residual rank below that cap. StaticShared freezes the first learned grouping and disables later regrouping. The D^2-MoE port pins `lliai/D2MoE` commit `b3092f030babf277218593b23cb5f98c66464154`, uses its reference tensor path, the same cap, and the same continuation budget. Tutel pins `ac0e51b943bb7084d4a6a7243e44a4afe04b9911`; its dispatcher receives the same accepted-edge trace and frozen autotuning cache. Native StableMoE and Expert Choice retain their published assignment semantics and are reported only as contextual neighbors.

GLUE uses official train/dev splits, max length 128, a seeded terminal-token linear head, end-to-end fine-tuning, AdamW $(0.9, 0.999, 10^{-8})$, weight decay 0.01, clip 1.0, and 10% linear warm-up/decay (Wang et al., 2018). The fixed (epochs, global batch, peak learning rate) tuples are CoLA $(10, 32, 10^{-5})$, SST-2 $(3, 32, 2\times10^{-5})$, MRPC $(10, 16, 10^{-5})$, STS-B $(10, 16, 2\times10^{-5})$, QQP $(3, 32, 2\times10^{-5})$, MNLI $(3, 32, 2\times10^{-5})$, QNLI $(3, 32, 2\times10^{-5})$, and RTE $(10, 16, 10^{-5})$. The terminal checkpoint at the prespecified epoch is always reported; dev performance never selects

an epoch. Within each task and evaluation seed, compact and control runs share head initialization, shuffle, batch order, and counter-based randomness, and task scores are paired before the fixed GLUE-8 macro is formed. WikiText-103 uses length 2,048 and stride 1,024, scoring each token once (Merity et al., 2017). MMLU uses fixed five-shot answer-token likelihood; HellaSwag/ARC-Challenge/PIQA use 10/25/0-shot length-normalized candidate likelihood; HumanEval uses greedy $n = 1$ pass@1, a 512-token cap, and a five-second test timeout (Hendrycks et al., 2021; Zellers et al., 2019; Clark et al., 2018; Bisk et al., 2020; Chen et al., 2021).

The reference system uses one, two, and four nodes with four A100-80GB GPUs per node, four-GPU NVLink islands, and a 200-Gb/s-class GPUDirect-RDMA fabric. Clocks and process order are fixed; CUDA 12.4, NCCL 2.21.5, and PyTorch 2.5.1 are used throughout. Acceptance refits use CUDA 12.4 cuSOLVER FP64 `gesvdj` on A100, tolerance 10^{-12}, at most 100 internal Jacobi sweeps, descending singular values, and a column-sign rule that makes the largest-magnitude right-vector entry positive with smallest-index ties. TP=PP=1; on the primary two-node mapping, $E = 32$ uses DP=2/EP=4 and $E \in \{64, 128, 256\}$ uses DP=1/EP=8. The $E = 128$ strong-scaling path keeps DP=1 and uses EP=4/8/16 on one/two/four nodes; weak scaling holds 16,384 offered tokens per source rank. The same fused dispatcher, variable-split `all_to_all_single`, compilation flags, topology file, and post-pilot autotuning cache are used for every baseline. Prefill fixes 64 sequences of length 2,048 (131,072 offered tokens); decode fixes a 2,048-token prompt, 256 generated tokens, batch 32, and greedy decoding. Continuous batching replays a hashed 10,000-request log-normal renewal trace (arrival CV 1.5) for ten minutes after two minutes of warm-up at common 30% and 80% Switch-derived rates. A request meets the SLO when time-to-first-token is at most 1.0 s and its p99 inter-token latency is at most 50 ms. Goodput counts completed output tokens only from requests meeting both thresholds; load is fixed from the controlled-Switch service rate. Each process block has 200 warm-up and 1,000 measured steps. Primary training throughput divides the declared 3B continuation positions by fenced wall-clock over the complete run; the $E = 256$ extension analogously times its declared 2B positions. Both include every accepted/rejected proposal, SVD/refit, shadow update, bridge, repair, placement solve, journal/fsync, and checkpoint, and terminal fences drain background work. Every event uses its measured monotonic start-to-finish interval, including proposals rejected by the pre-screen.

The OLMoE experiment pins `allenai/OLMoE-1B-7B-0924` at commit `6d84c48581ece794365f2b8e9cfb043c68ade9c5`. Each of ten DyClust/control pairs contains two independently continued branches, and each branch consumes 4B primary token positions at native 2,048-token packing. The adapter uses a bias-free selector, opens two groups, expands to selector mass 0.995 or six additional groups, and never removes reserve groups. The resulting union plus reserve experts is the candidate edge set for the same $k = 8$, factor-1.25 bounded parallel repair; its costs use native all-64 softmax weights without accepted-support renormalization. Both sides use identical BF16 optimizer/data/update-mask schedules under DP=8 and EP=8 on eight eight-GPU H100 nodes.

B.2 Statistics and units

Primary quality, training-throughput, serving, HBM, traffic, and downstream results use six paired seeds. Mechanism, routing-error, bridge, repair, and placement results use four seeds; the factorial design uses four seeds; the sensitivity surface uses five paired seeds; each distribution shift uses six paired seeds; and each public-model transfer uses ten compact/native pairs. Requests, token gaps, layers, events, assessment blocks, and repeated launches are technical subsamples and never increase these outer-unit counts. Results are descriptive means $\pm$ SD unless a table explicitly reports point estimates, exact checks, or a simultaneous bound. Shift bounds use one-sided Bonferroni t_5 limits with $m = 4$ per response. OLMoE uses one-sided Bonferroni t_9 limits with $m = 3$ separately for quality and systems; Qwen uses t_9 limits with $m = 5$ and $m = 3$. Failed and excluded outer units retain reason codes; no non-finite value is imputed.

B.3 Quality contrasts and downstream guards

Table 6: Descriptive quality contrasts over six seeds. Each cell is DyClust-MoE minus the named control and is computed before display rounding. The $E = 128$ same-E Switch row reports the paired mean and SD; all other rows are differences of the unrounded marginal means in Table 18.

Control	E	GLUE difference	PPL difference
Controlled Switch	32	-0.19	$+0.28$
Controlled Switch	64	-0.14	$+0.24$
Controlled Switch	128	-0.08 ± 0.10	$+0.19 \pm 0.08$
Parameter-upper Switch	32	$+0.49$	-1.59
Parameter-upper Switch	128	$+0.73$	-0.93
Nominal-matmul Switch	32	$+0.40$	-0.62
Nominal-matmul Switch	64	$+0.45$	-0.73
Nominal-matmul Switch	128	$+0.52$	-0.63
OC-Share	32	$+0.35$	-0.34
OC-Share	128	$+0.55$	-0.49
GlobalBase-LR	32	$+0.32$	-0.37
GlobalBase-LR	128	$+0.43$	-0.48
D^2-MoE+continuation	32	$+0.27$	-0.31
D^2-MoE+continuation	128	$+0.51$	-0.41
Decoupled-3State	128	$+0.08$	-0.02

GlobalBase-LR GLUE/PPL means at $E = 32/64/128$ are $84.31 \pm 0.38/25.46 \pm 0.24$, $84.73 \pm 0.36/25.09 \pm 0.23$, and $85.03 \pm 0.36/24.77 \pm 0.21$. Its frozen global residual ranks are 64/73/78 at $E = 32/64/128$. Specifically, $N_{\text{global}}(q) = N_0 + L\{2dd_f + G(d+1) + E[2q(d+d_f) + (d+d_f) + (d+1)]\}$; these are the largest integers satisfying $N_{\text{global}}(q) \le N_{\text{ours}}^{\text{store}}$, after which the canonical serializer verifies the BF16 tensor order, actual headers, 256-byte alignment, and router manifest. No dead tensor pads either side.

Table 7: Descriptive task-level GLUE differences at $E = 128$ over six seeds. Scores use official task metrics multiplied by 100; estimates are DyClust-MoE minus Switch.

Task	Point difference
CoLA	-0.17
SST-2	-0.02
MRPC	-0.08
STS-B	-0.06
QQP	-0.03
MNLI	-0.09
QNLI	-0.05
RTE	-0.14

The fixed macro of the displayed task-level point differences is -0.08; its paired SD is computed from the six seed-level macro differences. CoLA uses Matthews correlation, STS-B mean Pearson/Spearman, MRPC and QQP mean accuracy/F1, MNLI the matched/mismatched mean, and the remaining tasks accuracy. The task-level panel shows the pattern hidden by the macro alone. Secondary $E = 128$ DyClust-minus-Switch effects are MMLU -0.2 ± 0.35, HellaSwag -0.3 ± 0.40, ARC-Challenge -0.3 ± 0.45, PIQA -0.2 ± 0.35, and HumanEval pass@1 -0.3 ± 0.70 points.

Table 8: Secondary descriptive native-router neighbors at $E = 128$. Quality cells are DyClust-MoE minus the neighbor; systems cells are DyClust-MoE/neighbor. Native assignment semantics are retained, so these are neither exact-k controls nor confirmatory families.

Neighbor	GLUE diff	PPL diff	Prefill ratio	F+B-wire ratio
Native Expert Choice	$+0.22 \pm 0.14$	-0.11 ± 0.07	1.24 ± 0.024	0.74 ± 0.021
StableMoE	$+0.07 \pm 0.15$	-0.01 ± 0.08	1.23 ± 0.025	0.76 ± 0.022

B.4 Traffic and memory boundaries

Forward wire is sender-egress dispatch plus combine bytes over all ranks and eight MoE layers for 131,072 offered non-padding tokens entering the layer before capacity or repair. Backward and all-training wire add gradients, teacher, regroup, both bridge branches, journal, failed/retried transactions, and checkpoint control. Logical edge bytes, destination-unique payload, collective padding, sender-egress wire, physical byte-hops, and exposed critical-rank time occupy disjoint counters and reconcile to NIC/NVLink counters. HBM includes replicated weights, router state, KV/activations, dispatch buffers, kernels, workspaces, allocator reservation, and fragmentation for inference; optimizer and transaction state are reported separately for training. Decimal GB is used in tables.

Table 9: Representation and byte ledger. Active entries are nominal one-group analytical paths. The INT4 column is an analytical lower bound with BF16 bases and unchanged tensors, one FP16 scale per 128 residual weights, and no zero points.

E	Inventory	Exact count
32	Ours stored scalars	216,874,272
32	Switch stored scalars	1,127,537,920
32	Ours nominal active	103,535,712
32	Switch nominal active	135,804,160
64	Ours stored scalars	364,256,832
64	Switch stored scalars	2,185,584,128
64	Ours nominal active	103,560,320
64	Switch nominal active	136,001,024
128	Ours stored scalars	659,021,952
128	Switch stored scalars	4,301,676,544
128	Ours nominal active	103,609,536
128	Switch nominal active	136,394,752
256	Ours stored scalars	1,248,552,192
256	Switch stored scalars	8,533,861,376
256	Ours nominal active	103,707,968
256	Switch nominal active	137,182,208
32	Ours 2-byte dense	433,748,544
32	Switch 2-byte dense	2,255,075,840
32	INT4 analytical lower bound	412,736,064
64	Ours 2-byte dense	728,513,664
64	Switch 2-byte dense	4,371,168,256
64	INT4 analytical lower bound	686,488,704
128	Ours 2-byte dense	1,318,043,904
128	Switch 2-byte dense	8,603,353,088
128	INT4 analytical lower bound	1,233,993,984
256	Ours 2-byte dense	2,497,104,384
256	Switch 2-byte dense	17,067,722,752
256	INT4 analytical lower bound	2,329,004,544

Let $N_0 = 69{,}491{,}712$ be the unchanged inventory, $L = 8$ the number of MoE layers, $d = 768$, $d_f = 2688$, $G = E/K$, $K = 8$, $r = 8$, and $k = 2$. In stored counts, $G(d+1)$ accounts for the group

selector and $E(d+1)$ for the expert-router rows shared by the flat teacher and within-group scoring. Counting every matrix, bias, and router row once gives

$$\begin{aligned}
N_{\text{ours}}^{\text{store}} &= N_0 + L\{G[2dd_f + (d+1)] \\
&\quad + E[2r(d+d_f) + (d+d_f) + (d+1)]\}, \\
N_{\text{Switch}}^{\text{store}} &= N_0 + LE[2dd_f + (d+d_f) + (d+1)], \\
N_{\text{ours}}^{\text{active}} &= N_0 + L\{2dd_f + k[2r(d+d_f) + (d+d_f)] \\
&\quad + K(d+1) + G(d+1)\}, \\
N_{\text{Switch}}^{\text{active}} &= N_0 + Lk[2dd_f + (d+d_f)] + LE(d+1).
\end{aligned}$$

Thus stored counts equal $N_0 + 4{,}605{,}705E$ and $N_0 + 33{,}063{,}944E$, while nominal active counts equal $103{,}511{,}104 + 769E$ and $135{,}607{,}296 + 6{,}152E$. The nominal DyClust-MoE count is the analytical one-group top-2 path: one group-base pair, two residual/bias paths, all K within-group expert-router rows, and all G selector rows. It is neither a worst-case cross-group count nor measured FLOPs. At $E = 128$, the measured 1.23 bases/token corresponds to about 111.21M realized active scalars and about 5.189M expert-matrix scalars per MoE layer. For the INT4 lower bound, $N_{\text{res}} = 8E \cdot 55{,}296$ and $B_{\text{INT4}} = B_{\text{FP16}} - 1.5N_{\text{res}} + 2(N_{\text{res}}/128)$, which gives the integer values in Table 9.

B.5 Routing diagnostics

For every token we first compute omitted mass, selector TV, hard-assignment TV, and their sum in Equation (10); only then do we reduce within seed and average over four seeds. This preserves the theorem's estimand rather than summing separately averaged terms. The learned magnitudes are reduced token-first within seed and then summarized over four seeds. They form a dedicated descriptive mechanism family and never treat tokens as replicates; exact-k and capacity remain conditional invariants. Within each seed, correlations are computed over a fixed set of document-disjoint macroblocks and Fisher-z transformed; the reported point correlations are then formed over the four seed-level values, not over nested macroblocks.

Table 10: Routing-TV decomposition and measured placement and bridge diagnostics over four seeds. Router means and SDs use token-first reductions; p95/max are token-distribution diagnostics after the declared outer-unit reduction. Bridge rows use seed maxima over accepted events and slices.

Domain	Details
Router $E = 32$	**Quantity:** $\varepsilon/\delta_{\text{router}}/\eta_{\text{hard}}$ **Result:** $0.010 \pm 0.002/0.029 \pm 0.004/0.025 \pm 0.004$; bound/actual $0.064 \pm 0.006/0.052 \pm 0.005$ **Interpretation:** p95 $0.110/0.091$; max $0.205/0.174$
Router $E = 64$	**Quantity:** $\varepsilon/\delta_{\text{router}}/\eta_{\text{hard}}$ **Result:** $0.012 \pm 0.002/0.032 \pm 0.004/0.027 \pm 0.004$; bound/actual $0.071 \pm 0.006/0.058 \pm 0.005$ **Interpretation:** p95 $0.122/0.101$; max $0.218/0.183$
Router $E = 128$	**Quantity:** $\varepsilon/\delta_{\text{router}}/\eta_{\text{hard}}$ **Result:** $0.014 \pm 0.002/0.035 \pm 0.004/0.029 \pm 0.004$; bound/actual $0.078 \pm 0.007/0.064 \pm 0.006$ **Interpretation:** p95 $0.132/0.109$; max $0.231/0.194$
Router correlations	**Quantity:** total TV–NLL / hard TV–load CV **Result:** 0.43 / 0.47 **Interpretation:** seed-level Fisher-z point estimates
Router correlations	**Quantity:** omitted mass–opened groups / selector TV–selector KL **Result:** -0.61 / 0.72 **Interpretation:** seed-level Fisher-z point estimates
Grouping	**Quantity:** cumulative-ledger admission / accepted-event violations **Result:** $B_\ell \leq 0.10$ / 0 **Interpretation:** compliance only; bound is not a realized maximum

Continued on next page

Table 10 (continued)

Domain	Details
Grouping	**Quantity:** screening-output admission / accepted-event violations **Result:** $D_{\text{out}} \leq 0.02$ / 0 **Interpretation:** compliance only; bound is not a realized maximum
Placement	**Quantity:** calibration-validation R^2 / p95 relative error **Result:** 0.985 ± 0.006 / $3.8 \pm 1.0\%$ **Interpretation:** disjoint proxy gate before solving
Placement	**Quantity:** optimized/shuffled held-out replay **Result:** wire 0.78 ± 0.03; critical p95 0.88 ± 0.04 **Interpretation:** disjoint-trace effect
Bridge	**Quantity:** max candidate-minus-old path NLL **Result:** 0.010 ± 0.003 nats/token **Interpretation:** seed maximum over accepted events and slices; gate < 0.020
Bridge	**Quantity:** max |gradual–immediate-swap| PPL gap **Result:** 0.12 ± 0.04 **Interpretation:** seed maximum over accepted events and slices; gate < 0.25

Table 11: Measured $E = 128$ bounded-repair and prefill timing. The 131,072-token repair interval is included in the end-to-end window. Values are four-seed means $\pm$ SD; each seed first reduces calls, then sums disjoint timing components. Per-layer repair averages eight layers.

Offered tokens	Timed boundary	Time
16,384	One-layer repair	3.9 ± 0.3 ms
16,384	One-layer repair p99	8.1 ± 0.6 ms
131,072	Repair per layer (eight-layer average)	3.4 ± 0.2 ms
131,072	Eight MoE-layer repairs	27.2 ± 1.6 ms
131,072	Model forward excluding repair	267.6 ± 8.0 ms
131,072	Other end-to-end runtime overhead	4.6 ± 0.5 ms
131,072	End-to-end prefill	299.4 ± 9.0 ms

The 438 k tok/s headline is the aggregate throughput reducer over the complete prefill benchmark blocks, whereas the 299.4 ms row is the four-seed instrumented replay mean used for the timing partition. Component means add; their SDs do not, because the intervals are correlated. Aggregate throughput is reduced separately over the six primary checkpoint pairs.

Table 12: Regroup opportunity funnel over four seeds. Each scale has 99 opportunities per seed; acceptance statistics use seeds as outer units.

E	Accepted events by seed	Acceptance
32	$(33, 38, 42, 43)$	$39.4 \pm 4.6\%$
64	$(29, 32, 35, 39)$	$34.1 \pm 4.3\%$
128	$(25, 28, 31, 34)$	$29.8 \pm 3.9\%$

Table 13: $E = 128$ regroup event costs and gate distributions. Event times are monotonic wall-clock measurements inside the full-run ledger. Gate quantiles are descriptive event distributions with equal seed weight and do not increase n. A rejected column contains only candidates whose mutually exclusive first failure is the named gate.

Quantity	Result / activated	First-fail
Event costs		
Accepted event, total	43 s	–
Accepted refit/swap component	31 s	–
Other accepted-event work	12 s	–
Rejected event after pre-screen	12 s	–
Cold-start calls, accepted/rejected	$\approx 29/\approx 1$	–
Mean cold-start calls per opportunity	≈ 9.3	–
Gate distributions: activated / first-fail p50/p95		
Exact objective decrease	$0.013/0.027$	$0.000/0.004$
Event RMS $b_{\ell,e}$	$0.016/0.025$	$0.035/0.052$
Output drift D_{out}	$0.009/0.016$	$0.027/0.041$
Selector KL	$0.011/0.018$	$0.030/0.048$
Bridge max NLL	$0.010/0.016$	$0.028/0.043$

The timing decomposition in Table 13 is scoped to $E = 128$. Each full-expert sweep contains 256 SVDs; a complete refit uses at most eight sweeps, or 2,048 SVDs. A single-matrix SVD takes about 4 ms, and the implementation executes these operations in parallel batches. The cheap pre-screen runs after two probe sweeps in one cold start, so rejected events do not enter the complete swap scan. Across the 99 opportunities, the mean per-seed measured event time totals about 35 minutes; ordinary checkpoints, background repair, terminal fences, and other non-event fixed work add about 5.4 minutes. The ordinary-update component, excluding those separately ledgered fixed-work intervals, is 198.5 k tok/s, corresponding to 4.20 hours over 3B positions, whereas the complete run is 171 k tok/s, corresponding to 4.87 hours. Throughput is computed from unrounded fenced durations before the declared outer-unit reductions; displayed rates, duration summaries, and component totals are rounded only afterward and therefore need not be exact reciprocals or sums. Candidate-minus-no-regroup PPL is -0.07 ± 0.05 at the +256-update audit and -0.11 ± 0.06 at +1,000 updates. Delayed rollback retains the previous complete epoch and never exposes a mixed epoch.

B.6 Bridge and shift measurements

The bridge compares immediate swap, interpolation, and no-regroup controls on the same fenced inputs; Table 10 reports the two path endpoints, while the 256-update ramp is part of ordinary post-activation training rather than a second selection view. Composition, source-disjoint PMC, 2025-Q4 temporal, and Chinese/Spanish shifts use six paired seeds with matched Switch and no-regroup branches. Each seed first yields a layer-visit count, a global-decision count, and a terminal PPL gap; absolute endpoints are maxima over the six seed-level values, whereas no-regroup contrasts are paired-seed means±SD. Neither reducer treats adaptation decisions as a quality score.

From the same prespecified checkpoint, the composition shift changes the five source proportions from $(0.60, 0.15, 0.10, 0.10, 0.05)$ to $(0.30, 0.10, 0.35, 0.10, 0.15)$ without loss reweighting. The source-disjoint shift instead uses license-filtered PMC Open Access records. The temporal fork uses only records first published in 2025-Q4, and the multilingual fork uses equal token mass from language-ID-filtered Chinese and Spanish records. Every replacement manifest is disjoint by document ID and normalized 13-grams from parent, continuation, validation, and evaluation roles. For each shift and seed, the DyClust-MoE, controlled same-E Switch, and no-regroup branches each consume 1B token positions and share post-shift token order, optimizer, counter-based randomness, and sealed assessment blocks.

One global decision is one prescheduled post-shift grouping opportunity, with layers visited round-robin. For layer ℓ at its jth visit, a qualifying commit is an activated candidate that passes every

reconstruction, routing, bridge, and publication gate and whose one-sided lower bound

$$\overline{d} - t_{31,1-\alpha_{\ell j}} s_d/\sqrt{32} > 0.02, \qquad \alpha_{\ell j} = 0.05/[8j(j+1)],$$

uses 32 frozen document-disjoint assessment macroblocks and $d_b = \mathrm{NLL}_{\mathrm{incumbent,b}} - \mathrm{NLL}_{\mathrm{candidate,b}}$. Missing, duplicated, non-finite, or unbalanced blocks fail the visit; blocks never increase n. Within a seed, the visit endpoint is the maximum over layers of visits through each layer's last qualifying commit; a layer with none retains a failure rather than being dropped. After every global decision, a fresh sealed shard, disjoint from proposal, shadow, commit, and all earlier assessment shards, computes the signed gap $\Delta_{\mathrm{PPL}} = \mathrm{PPL}_{\text{DyClust-MoE}} - \mathrm{PPL}_{\mathrm{Switch}}$. Recovery is the first decision after the last qualifying commit with $\Delta_{\mathrm{PPL}} \le c_s$, maintained on the next two sealed assessments, which confirm but do not increment the recovery endpoint, where $c_s = 0.20/0.30/0.30/0.40$ for composition/PMC/temporal/multilingual. The terminal endpoint uses the same signed contrast on a never-touched terminal shard; no-regroup improvements are $\mathrm{PPL}_{\text{no-regroup}} - \mathrm{PPL}_{\text{DyClust-MoE}}$ and the analogous excess decision count. Control recovery uses the same global grid and the same DyClust post-commit start decision; only the branch supplying the signed PPL gap changes.

C Scope, controls, and frozen protocol

C.1 Closest-work boundary

Table 14 separates the synthesis evaluated here from the individual primitives on which it builds. The contribution is not low-rank structure, hierarchical routing, capacity repair, coalescing, or placement in isolation. It is the use of one deterministic, versioned grouping state across representation, routing, and placement, together with interventions that hold the other interfaces fixed. Accordingly, a favorable compression ratio cannot by itself support a routing or systems claim, and exact-k feasibility cannot be interpreted as either routing quality or low latency.

Table 14: Scope comparison with the closest mechanism families. Entries summarize the primary mechanism evaluated by each line of work; they do not rule out compatible extensions.

Line of work	Details
Sub-MoE and MoE-LoRA (Li et al., 2026; Sun et al., 2025)	**Representation lever:** Expert-subspace merging or low-rank adapters **Routing / assignment lever:** Merged experts or routed adapters **Systems lever:** Weight compression or parameter-efficient adaptation **Evidence scope:** Compression for Sub-MoE; fine-tuning for MoE-LoRA
D^2-MoE (Gu et al., 2025)	**Representation lever:** Merged or pruned base plus low-rank deltas **Routing / assignment lever:** Existing sparse router retained **Systems lever:** Static delta decompression **Evidence scope:** Post-training delta decomposition; no online regroup publication
Stable, adaptive, and semantic routers (Dai et al., 2022; Jin et al., 2025; Min et al., 2025)	**Representation lever:** Learned experts; MoE++ also uses zero-computation experts **Routing / assignment lever:** Stable, adaptive, or semantic routing **Systems lever:** Allocation of expert work **Evidence scope:** Routing studies; semantic priors evaluated on visual MoE
Expert Choice (Zhou et al., 2022)	**Representation lever:** Full experts **Routing / assignment lever:** Expert-selects-token capacity regime **Systems lever:** Balanced expert capacity **Evidence scope:** Assignment quality and scaling
Tutel and MoE-Lightning (Hwang et al., 2023; Cao et al., 2025)	**Representation lever:** Orthogonal to weight sharing **Routing / assignment lever:** Existing sparse routing semantics **Systems lever:** Dispatch, caching, scheduling, or placement **Evidence scope:** Runtime systems evaluation

Continued on next page

Table 14 (continued)

Line of work	Details
SMILE (He et al., 2022a)	**Representation lever:** Full experts **Routing / assignment lever:** Bi-level routing **Systems lever:** Parallel routing execution **Evidence scope:** Joint routing/systems evaluation
MegaBlocks and ScatterMoE (Gale et al., 2023; Tan et al., 2024)	**Representation lever:** Expert tensors supplied to the runtime **Routing / assignment lever:** Execution of sparse assignments **Systems lever:** Dropless block-sparse computation, or padding/copy reduction **Evidence scope:** GPU execution efficiency; distinct from grouping-state coordination
DeepSeekMoE (Dai et al., 2024)	**Representation lever:** Fine-grained routed experts and separate shared experts **Routing / assignment lever:** Shared experts execute for every token **Systems lever:** Sparse activation with shared capacity **Evidence scope:** Architectural specialization; shared experts differ from shared matrix bases
QMoE (Frantar & Alistarh, 2024)	**Representation lever:** Sub-1-bit encoding of expert weights **Routing / assignment lever:** Existing sparse computation **Systems lever:** Custom compressed format and GPU decoding **Evidence scope:** Encoded-byte/quality trade-off, separate from BF16 scalar counts
DyClust-MoE	**Representation lever:** Group bases plus expert residuals **Routing / assignment lever:** Mass hierarchy plus reserve-feasible exact-k **Systems lever:** Accepted-payload coalescing and owner placement **Evidence scope:** One versioned group state with frozen-interface interventions

C.2 Frozen method constants

Table 15 collects constants that otherwise appear across the main method and experimental protocol. The 99 proposal positions are fixed as evenly spaced cursor fractions and stop before the terminal boundary so that every event has a subsequent in-budget primary update for its bridge. Thresholds are applied to the reducers stated in the table, not to individual tokens, layers, slices, or technical launches.

Table 15: Complete primary $E \leq 128$ small-model method configuration. Algorithmic constants and decision thresholds were fixed before the six confirmatory seeds were evaluated.

Component	Details
Sparse structure	**Frozen value:** $K = 8$, $r = 8$, top-2 **Operational definition:** Every same-E comparator starts from the same materialized update-0 experts and expert-router rows; expert-specific biases are retained.
Training auxiliaries	**Frozen value:** $(\lambda_E, \lambda_G, \lambda_C, \lambda_M) = (0.01, 0.01, 0.05, 0.10)$ **Operational definition:** Global non-padding reducers; the flat teacher is evaluated only on the counter-hashed $1/64$ sample.
Grouping objective	**Frozen value:** $\lambda_{\text{act}} = 0.25$, $\lambda_{\text{pair}} = 0.10$, $\epsilon_w = 10^{-12}$ **Operational definition:** Reconstruction is normalized over all $P_W E$ expert–projection pairs; co-routing counts unordered accepted pairs once.

Continued on next page

Table 15 (continued)

Component	Details
Primary proposal schedule	**Frozen value:** 99 evenly spaced cursor positions for $E \leq 128$ **Operational definition:** One layer is visited round-robin; each event uses its preceding document history split $7/8$ for proposal and $1/8$ for proposal-disjoint screening.
Sketch and refit	**Frozen value:** 256 bins; two probe sweeps; eight full-refit sweeps; 32-pair shortlists; $8E$ exact evaluations **Operational definition:** The probe screen requires reconstruction RMS ≤ 0.06 and expert tail ≤ 0.20 on the 2^{20}-token history; only survivors enter exact refinement. Acceptance uses CUDA 12.4 cuSOLVER FP64 `gesvdj` with tolerance 10^{-12}, at most 100 Jacobi sweeps, descending singular values, the stated largest-entry sign rule, and stable discrete ties.
Objective gate	**Frozen value:** relative decrease 0.005; floors 10^{-6} and 10^{-7} **Operational definition:** An event is skipped when the incumbent objective is at the floor; otherwise the retained exact-objective decrease must exceed the larger relative or absolute threshold.
Distortion gates	**Frozen value:** $b_{\ell,e} \leq 0.03$, $\max_{i,p} \tau_{i,p}^{(e)} \leq 0.12$, $B_\ell + b_{\ell,e} \leq 0.10$, $D_{\text{out}} \leq 0.02$ **Operational definition:** All four gates must hold before the event ledger advances; the output reducer uses frozen routing on the proposal-disjoint screen.
Selector preflight	**Frozen value:** 64 shadow steps; KL ≤ 0.02; mass ≥ 0.98; top-k recall ≥ 0.99 **Operational definition:** Private selector state is trained on $7/8$ of the next ordinary batch and checked on its untouched $1/8$; abort discards all private state.
Bridge	**Frozen value:** $H = 64$ slices of 4,096 tokens; local ramp 256 updates **Operational definition:** One 262,144-token ordinary batch is partitioned into the slices. Slice maxima require candidate-minus-old NLL < 0.020 nats/token and gradual-versus-immediate-swap absolute PPL gap < 0.25.
Assignment	**Frozen value:** capacity 1.25; selector mass 0.98; four additional groups; $R_{\max} = 16$ repair rounds **Operational definition:** Reserve groups are retained. Each candidate-batch evaluation counts toward the cap, including rejected batches; accepted batches preserve degree, uniqueness, and capacity.
Placement	**Frozen value:** four probe sizes; 200 probes per link class **Operational definition:** Nonnegative least squares fits launch and byte coefficients; the fixed-charge MIP minimizes a critical-owner proxy and is evaluated only on request-disjoint traces.
Publication	**Frozen value:** one in-flight candidate; immutable generations; epoch CAS **Operational definition:** Tensor state, selector, owner map, optimizer, cursor, RNG, and byte ledger become visible together or the old epoch remains visible.

C.3 Comparator roles and matching variables

The comparison matrix in Table 16 prevents one control from silently answering several different questions. Controlled same-E Switch tests the quality cost relative to a full expert pool; parameter brackets test storage; nominal-matmul Switch tests approximate active expert work; structured controls test the representation; Decoupled-3State tests the shared-state claim; and frozen replays isolate execution interfaces.

Table 16: Comparator roles. Unless a row names the changed axis, controls share the parent, data order, active k, precision, token budget, and blind architecture-local selection budget.

Question	Details
Expert-pool quality	**Configuration:** Controlled same-E Switch top-2 **Axis isolated:** Quality relative to an uncompressed pool with the same expert count under the common reserve/repair/zero-drop framework.
Stored capacity	**Configuration:** Lower/upper integer Switch brackets **Axis isolated:** The nearest smaller/larger full-expert inventories; the upper member is the harder quality comparison.
Nominal active work	**Configuration:** Switch with $d_{\text{ff}} = 1408$ **Axis isolated:** Nearest tensor-core-aligned top-2 expert matmul inventory to the one-group analytical path; realized DyClust-MoE expert-matrix work is $1.20\times$ this comparator at $E = 128$.
Structured sharing	**Configuration:** OC-Share; GlobalBase-LR **Axis isolated:** Group-local versus global-base low-rank structure at the same canonical byte cap.
Static compression	**Configuration:** D^2-MoE+continuation **Axis isolated:** A pinned delta-decomposition path given the same continuation budget and byte cap.
State coupling	**Configuration:** Decoupled-3State **Axis isolated:** Three independently optimized states; expert-ID tuples are precompiled once per epoch and read by the same fused dispatcher. Bytes, active k, tokens, search budget, and hardware match.
Production runtime	**Configuration:** Tutel dispatcher **Axis isolated:** A pinned mature runtime under the same checkpoint, accepted-edge trace, software stack, and pilot-only autotuning.
Adaptation	**Configuration:** StaticShared; fixed balanced random groups **Axis isolated:** Dynamic regrouping relative to a learned frozen grouping and seed-nested random maps.
Routing	**Configuration:** Flat global; no mass loss; native Expert Choice; StableMoE **Axis isolated:** Candidate scoring, distillation, exact-k assignment, and contextual native-routing behavior.
Execution	**Configuration:** Materialized experts; score-all; shuffled owners; no coalescing **Axis isolated:** Kernel, scoring, placement, and payload aggregation under identical accepted assignments or tensor outputs.

C.4 Inference units and family boundaries

Primary quality, full-run training, serving, HBM, traffic, and downstream results use six paired seeds as outer units. Mechanism results and bridge, repair, and placement validation use four seeds; factorial results use four seeds; sensitivity uses five paired seeds; distribution shifts use six paired seeds; the $E = 256$ extension uses six seeds; and each public model uses ten paired continuations. Technical process blocks are reduced inside the paired outer unit. System requests, token gaps, layers, events, assessment blocks, and repeated launches are subsamples or technical repeats, never additional scientific replicates. Tables report descriptive mean$\pm$SD unless they explicitly print a simultaneous bound.

Table 17: Result registry. "Diff" is DyClust-MoE minus control and "ratio" is DyClust-MoE/control. Exact implementation invariants have no statistical margin.

Result block	Details
Primary quality	**Estimand and outer unit:** GLUE/PPL marginal means; paired SD for the $E = 128$ same-E contrast, six seeds **Reporting:** descriptive **Role:** Quality–representation trade-off across scales and controls
Primary systems	**Estimand and outer unit:** Paired throughput ratios, HBM differences, and wire ratios, six outer units **Reporting:** descriptive **Role:** Measured execution behavior on the tested A100 topology
Mechanisms	**Estimand and outer unit:** Matched 3B training-control protocol and frozen replays, four seeds **Reporting:** protocol plus descriptive replay effects **Role:** Interface attribution under matched endpoints
Bridge/repair/ placement	**Estimand and outer unit:** Seed-level validation reducers, four seeds **Reporting:** descriptive plus hard checks **Role:** Event path, assignment, and publication behavior
Sensitivity	**Estimand and outer unit:** Paired response surface, five seeds **Reporting:** descriptive **Role:** Local operating-point trends
Factorial	**Estimand and outer unit:** Prespecified contrast matrix, four seeds **Reporting:** descriptive **Role:** Main, pairwise, and higher-order effects
Shift adaptation	**Estimand and outer unit:** Decisions saved, PPL gain, and signed compact–Switch gap, six seeds **Reporting:** one-sided $m = 4$ bounds per response **Role:** Adaptation under four fixed distribution changes
Public-model transfer	**Estimand and outer unit:** Paired quality differences and systems ratios, ten pairs/model **Reporting:** OLMoE $m = 3/3$; Qwen $m = 5/3$ bounds **Role:** Native-routing transfer on two public architectures
$E = 256$ extension	**Estimand and outer unit:** Quality endpoints and full-run-training order-of-magnitude check, six seeds **Reporting:** descriptive **Role:** Larger-expert-count endpoint

C.5 Full absolute quality matrix

Table 18 gives the six-seed marginal means used for the descriptive contrasts in Table 6. Marginal SDs describe between-seed dispersion and are not combined to reconstruct paired uncertainty.

Table 18: Complete six-seed small-model quality results. Parameter-bracket expert counts are the nearest integer full-expert inventories below or above the corresponding compact row.

Model	E	GLUE-8 ↑	WikiText-103 PPL ↓
Dense-active	–	83.43 ± 0.38	28.72 ± 0.24
Controlled Switch	32	84.82 ± 0.36	24.81 ± 0.19
Controlled Switch	64	85.21 ± 0.32	24.42 ± 0.17
Controlled Switch	128	85.54 ± 0.31	24.10 ± 0.11
DyClust-MoE	32	84.63 ± 0.34	25.09 ± 0.18
DyClust-MoE	64	85.07 ± 0.31	24.66 ± 0.17
DyClust-MoE	128	85.46 ± 0.29	24.29 ± 0.16
Switch parameter lower	4	83.93 ± 0.41	27.03 ± 0.24
Switch parameter upper	5	84.14 ± 0.39	26.68 ± 0.22
Switch parameter lower	8	84.32 ± 0.39	26.12 ± 0.23
Switch parameter upper	9	84.43 ± 0.38	25.91 ± 0.23
Switch parameter lower	17	84.61 ± 0.36	25.42 ± 0.21
Switch parameter upper	18	84.73 ± 0.35	25.22 ± 0.21
Switch nominal-matmul	32	84.23 ± 0.35	25.71 ± 0.23
Switch nominal-matmul	64	84.62 ± 0.35	25.39 ± 0.23
Switch nominal-matmul	128	84.94 ± 0.35	24.92 ± 0.22
OC-Share	32	84.28 ± 0.39	25.43 ± 0.23
OC-Share	64	84.74 ± 0.37	25.08 ± 0.22
OC-Share	128	84.91 ± 0.38	24.78 ± 0.22
GlobalBase-LR	32	84.31 ± 0.38	25.46 ± 0.24
GlobalBase-LR	64	84.73 ± 0.36	25.09 ± 0.23
GlobalBase-LR	128	85.03 ± 0.36	24.77 ± 0.21
D^2-MoE+continuation	32	84.36 ± 0.37	25.40 ± 0.21
D^2-MoE+continuation	128	84.95 ± 0.36	24.70 ± 0.22
Decoupled-3State	128	85.38 ± 0.28	24.31 ± 0.18

The continuation-matched DyClust-minus-D^2-MoE point differences at $E = 32/128$ are GLUE $+0.27/+0.51$ points and PPL $-0.31/-0.41$. The equal-byte operating point is outside the 40–60% reduction range used in the frozen-checkpoint overlap panel, so this continuation contrast is an extreme-budget control rather than a general claim about D^2-MoE.

D Complete systems and resource results

D.1 Two-node measurements

Table 19 expands the main systems table to all three expert counts. Prefill uses 131,072 offered non-padding tokens. Full-run training includes proposal generation, exact refits, shadow work, accepted and rejected bridges, publication, and checkpoints. Forward wire is sender-egress dispatch plus combine traffic over all ranks and all eight MoE layers. Absolute throughput is reported only for the measured $E = 128$ DyClust-MoE run.

Table 19: Two-node systems measurements on eight A100 GPUs over six paired outer units. The 438/171 k tok/s entries are $E = 128$ aggregate DyClust-MoE throughput; Switch absolute throughput is not inferred from ratios.

E	Endpoint	Result
32	DyClust-MoE prefill	–
32	DyClust-MoE train	–
32	Prefill ratio	1.12 ± 0.025

Continued on next page

Table 19 (continued)

E	Endpoint	Result
32	Train ratio	1.12 ± 0.024
32	HBM diff (GB)	-0.40 ± 0.10
32	Forward-wire ratio	0.75 ± 0.024
64	DyClust-MoE prefill	–
64	DyClust-MoE train	–
64	Prefill ratio	1.18 ± 0.027
64	Train ratio	1.17 ± 0.026
64	HBM diff (GB)	-0.42 ± 0.11
64	Forward-wire ratio	0.78 ± 0.026
128	DyClust-MoE prefill	438
128	DyClust-MoE train	171
128	Prefill ratio	1.26 ± 0.030
128	Train ratio	1.22 ± 0.027
128	HBM diff (GB)	-0.92 ± 0.11
128	Forward-wire ratio	0.79 ± 0.026

The paired prefill and training ratios and HBM differences are shown directly in Table 19; the forward-wire ratios are 0.75 ± 0.024, 0.78 ± 0.026, and 0.79 ± 0.026 at $E = 32/64/128$. At $E = 128$, backward expert wire, F+B MoE wire, all-training wire, exposed communication time, and prefill energy/token are 0.70 ± 0.027, 0.73 ± 0.027, 0.77 ± 0.031, 0.75 ± 0.032, and 0.82 ± 0.032 relative to Switch.

D.2 Matched systems controls

Table 20 reports systems comparisons against controls that are closer in nominal active work or representation. Every effect is formed within pair. These modest differences characterize the tested kernels, workloads, and topologies, not hardware-independent dominance.

Table 20: Matched $E = 128$ systems matrix. HBM is DyClust-MoE minus control in GB; all other cells are DyClust-MoE/control.

Control	Paired effects
Nominal-matmul Switch	**Prefill:** 1.16 ± 0.019
	Full-run train: 1.13 ± 0.019
	80% goodput: 1.16 ± 0.024
	F+B MoE wire: 0.74 ± 0.020
	Peak HBM Δ: -0.69 ± 0.09
	Energy/token: 0.86 ± 0.027
OC-Share	**Prefill:** 1.18 ± 0.021
	Full-run train: 1.16 ± 0.020
	80% goodput: 1.20 ± 0.026
	F+B MoE wire: 0.78 ± 0.021
	Peak HBM Δ: -0.68 ± 0.11
	Energy/token: 0.86 ± 0.029
D^2-MoE+continuation	**Prefill:** 1.17 ± 0.022
	Full-run train: 1.15 ± 0.019
	80% goodput: 1.18 ± 0.025
	F+B MoE wire: 0.77 ± 0.022
	Peak HBM Δ: -0.66 ± 0.09
	Energy/token: 0.86 ± 0.028

Continued on next page

Table 20 (continued)

Control	Paired effects
Decoupled-3State	**Prefill:** 1.15 ± 0.022 **Full-run train:** 1.13 ± 0.019 **80% goodput:** 1.17 ± 0.024 **F+B MoE wire:** 0.80 ± 0.021 **Peak HBM Δ:** -0.43 ± 0.08 **Energy/token:** 0.89 ± 0.027
Tutel `ac0e51b943bb`	**Prefill:** 1.18 ± 0.024 **Full-run train:** 1.14 ± 0.020 **80% goodput:** 1.20 ± 0.028 **F+B MoE wire:** 0.77 ± 0.023 **Peak HBM Δ:** -0.78 ± 0.12 **Energy/token:** 0.87 ± 0.030

The production-runtime control pins Tutel commit `ac0e51b943bb7084d4a6a7243e44a4afe04b9911`, CUDA 12.4, NCCL 2.21.5, PyTorch 2.5.1, BF16 kernels, `-O3`, and `sm_80`; autotuning is performed only on pilot traces and then frozen. It consumes the same checkpoint, accepted-edge trace, split sizes, capacity factor, warm-up, and measured blocks, so the row compares production dispatch rather than changing routing semantics. Its descriptive paired effects appear in Table 20; the row uses the same six-outer-unit reduction as the other primary systems controls.

At four nodes, prefill ratios against OC-Share, D^2-MoE, and Decoupled-3State are 1.20 ± 0.024, 1.19 ± 0.025, and 1.19 ± 0.024. Decoupled-3State obtains lower reconstruction objective (0.957 ± 0.012 ratio), routing bound (-0.007 ± 0.003), and placement proxy (0.979 ± 0.009 ratio), but touches 1.54 versus 1.23 bases/token and 1.47 versus 1.19 owners/token. Its three states use the same 99 event opportunities and the same total candidate-refit budget, split cyclically among representation, routing, and placement proposals. Representation minimizes reconstruction, routing minimizes teacher-mass KL, and placement minimizes the held-out calibrated owner proxy. Expert-ID tuples are materialized once at publication; the fused dispatcher does no per-token hash join. Rebuild cost is charged to full-run training. Cached lookup consumes $0.9 \pm 0.2\%$ of MoE-layer time, versus $5.8 \pm 0.7\%$ for an uncached diagnostic using identical checkpoints and accepted edges. The shared/uncached prefill ratio is 1.19 ± 0.023, whereas the primary shared/cached ratio is 1.15 ± 0.022. In a metadata-only replay using identical base/factor tensors, assignments, and physical owners, the shared/cached ratio falls to 1.006 ± 0.005. These replays separate metadata handling from the execution consequences of the independently optimized partitions; they reuse checkpoints and add no continuation tokens.

D.3 Counter boundaries and reconciliation

System conclusions use the counter hierarchy in Table 21. The levels are deliberately not interchangeable: payload coalescing can reduce destination-unique bytes without changing logical accepted edges, and a lower sender-egress count need not yield lower elapsed time if congestion or overlap changes. All byte counters use the immutable offered-token denominator even if a measured unit later fails a hard invariant.

Table 21: Traffic, time, and memory counter boundaries. A valid process block reconciles adjacent byte levels and the device memory component sum before it contributes to a paired checkpoint mean.

Counter	Details
Offered tokens	**Included quantity:** Non-padding source tokens entering an MoE layer before capacity, reserve repair, or any drop decision **Primary use:** Common denominator
Accepted edges	**Included quantity:** Exactly k selected expert edges per valid offered token **Primary use:** Assignment invariant and diagnostic

Continued on next page

Table 21 (continued)

Counter	Details
Logical edge bytes	**Included quantity:** Payload implied by all accepted token–expert edges before shared-base or destination reuse **Primary use:** Algorithmic traffic baseline
Destination-unique payload	**Included quantity:** One token payload per destination owner even when several invoked groups or experts share it **Primary use:** Coalescing effect
Collective-padded bytes	**Included quantity:** Actual collective buffers after split metadata, alignment, and padding **Primary use:** Runtime buffer accounting
Sender-egress wire	**Included quantity:** Dispatch plus combine bytes summed over all senders, ranks, and eight MoE layers **Primary use:** Primary communication endpoint
Physical byte-hops	**Included quantity:** Directional link-level bytes after topology routing **Primary use:** Congestion and path diagnostic
Exposed critical-rank time	**Included quantity:** Non-overlapped communication on the slowest owner path **Primary use:** Placement/latency endpoint
Inference peak HBM	**Included quantity:** Weights, routers, KV/activations, dispatch buffers, kernels, workspaces, reservation, and fragmentation **Primary use:** Prefill memory endpoint
Training-event peak HBM	**Included quantity:** Inference components plus optimizer, candidate, shadow, bridge-sidecar, journal, and repair state **Primary use:** Transaction boundary

Forward, backward, teacher, regroup, shadow, both bridge branches, failed or retried transactions, journal control, optimizer, and checkpoint traffic occupy disjoint training counters whose union is compared with NIC/NVLink totals. Inference HBM excludes training-only transaction state because no transaction occurs in that workload; the separate training-event peak prevents those bytes from disappearing from the study. Tables use decimal GB, while exact scalar and byte inventories remain integers.

D.4 Serving and strong/weak scaling

The offline and continuous-serving endpoints use the same checkpoints as the prefill study. Strong scaling fixes 131,072 global offered tokens and reports global throughput; weak scaling fixes 16,384 offered tokens per source GPU and reports per-GPU throughput. Wire is sender-egress bytes per offered token and p95 is exposed critical-rank communication time.

Table 22: Normalized serving and scaling at $E = 128$. Ratios are DyClust-MoE/Switch; latency and wire are lower-is-better. SLO attainment is the displayed percentage-point difference rather than a ratio.

Scenario	Result	Outer-unit reducer
Offline decode, batch 32	1.18 ± 0.032 throughput	checkpoint mean
Batch-1 p50 inter-token latency	1.01 ± 0.028	checkpoint mean
Continuous, 30% load, p95 latency	0.95 ± 0.037	frozen-trace checkpoint mean
Continuous, 80% load, goodput / p99 ITL	1.23 ± 0.038 / 0.82 ± 0.034	frozen-trace checkpoint mean
Continuous, SLO attainment	$+11.6 \pm 2.4$ percentage points	frozen-trace checkpoint mean
Strong, 1 node: throughput / p95	1.08 ± 0.026 / 0.92 ± 0.029	checkpoint mean

Continued on next page

Table 22 (continued)

Scenario	Result	Outer-unit reducer
Strong, 2 nodes: throughput / wire / p95	1.26 ± 0.030 / 0.79 ± 0.026 / 0.72 ± 0.028	checkpoint mean
Strong, 4 nodes: throughput / wire / p95	1.34 ± 0.033 / 0.62 ± 0.023 / 0.64 ± 0.026	checkpoint mean
Weak, 1 node: per-GPU throughput / p95	1.06 ± 0.026 / 0.94 ± 0.030	checkpoint mean
Weak, 2 nodes: per-GPU throughput / wire / p95	1.22 ± 0.027 / 0.73 ± 0.024 / 0.77 ± 0.026	checkpoint mean
Weak, 4 nodes: per-GPU throughput / wire / p95	1.28 ± 0.030 / 0.65 ± 0.024 / 0.68 ± 0.027	checkpoint mean

D.5 Executed work, serialized bytes, and memory pressure

Nominal active scalars are not executed FLOPs. We therefore count accelerator floating-point work per offered token over the full declared continuation, including group/expert scoring, flat-teacher calls, sketches, every refit, assignment repair, shadow updates, and both bridge branches. Hardware counters are reconciled against a static operation ledger. CPU placement, serialization, fsync, and control traffic remain in their native wall-clock, CPU-second, I/O-byte, and energy ledgers rather than being converted to FLOPs.

Table 23: Executed work and transient memory at $E = 128$. FLOPs count one multiply-add as two operations and include the complete continuation. Memory is decimal GB/GPU. Six primary pairs supply work and inference estimates; four validation seeds supply event peaks.

Quantity	Mean $\pm$ SD	Boundary
Nominal-matmul Switch work	0.862 ± 0.006 GFLOP/token	Complete continuation
DyClust-MoE work	0.906 ± 0.009 GFLOP/token	Complete continuation
Paired executed-work ratio	1.051 ± 0.008	Within-pair ratio
Controlled Switch inference peak	8.73 ± 0.12 GB	Prefill
DyClust-MoE inference peak	7.81 ± 0.11 GB	Prefill
Ordinary training peak	11.18 ± 0.24 GB	Outside regroup events
Training-event peak	16.55 ± 0.38 GB	Candidate plus old epoch

The event peak adds 2.31 GB of candidate/FP64 workspaces, 2.78 GB of bridge activations, and 0.28 GB of shadow/publication buffers to the 11.18 GB ordinary training footprint, giving 16.55 GB. These are same-instant component means, not sums of independently timed maxima. The work ledger includes dense, expert, router, teacher, and refit kernels; FMA and tensor-core instructions are converted with their executed shapes. The maximum counter-to-ledger discrepancy is 0.8% across the paired runs.

At $E = 128$, the measured 1.23 bases/token yields about 5.189M expert-matrix scalars per MoE layer, $1.20\times$ the nominal-matmul Switch expert work. The executed-work ratio therefore rules out reduced arithmetic as the explanation for the throughput gain. The measured advantage is instead consistent with lower wire, topology-aware placement, destination coalescing, and scheduling.

The canonical BF16 inference artifacts, including tensor headers, alignment, router/group state, and container metadata, have paired byte ratios $0.35/0.32/0.30$ against nominal-matmul Switch at $E = 32/64/128$. These ratios are separate from the analytical scalar and dense-byte ledger in Table 9; neither is used as a proxy for HBM or latency.

For the memory-pressure experiment, allocator caps are applied before loading, variant order is randomized, and each run uses the same fragmentation warm-up followed by a last-success/first-OOM binary search. On A100-80GB constrained to 24 GB at 8k context, the maximum offered-token ratio is 1.25 ± 0.06. This is a capped-device stress result, not physical 24-GB deployment evidence. The same serializer, warm-up, trace, and search on two four-GPU A10-24GB nodes gives 1.22 ± 0.07.

The $E = 256$ cap study instead returns to the primary two-node A100-80GB DP=1/EP=8 mapping and applies allocator caps on that same mapping; its descriptive ratios over six paired outer units (five technical blocks nested within pair) are 1.35 ± 0.08 at 24 GB/8k and 1.25 ± 0.06 at 40 GB/16k. These are allocator-cap stress results, not physical A10 or 24/40-GB accelerator measurements. A 10,000-request $E = 128$ soak has peak HBM drift below 2%, zero OOMs, and zero runtime fallbacks. Measured deployment results use resident BF16 tensors; the INT4 ledger is an analytical byte bound.

D.6 Equal-byte D^2-MoE overlap

At $E = 128$, a representation-only comparison evaluates both methods over four paired continuation seeds at exact 40%, 50%, and 60% reductions from the same continued Switch checkpoint within each seed. Integer residual rank is selected under the same canonical byte cap before quality or systems evaluation. DyClust-minus-D^2-MoE results at the three caps are GLUE $+0.08 \pm 0.04$, $+0.12 \pm 0.05$, and $+0.18 \pm 0.07$; PPL -0.05 ± 0.025, -0.10 ± 0.035, and -0.18 ± 0.05; prefill ratios 1.01 ± 0.015, 1.02 ± 0.015, and 1.03 ± 0.02; and peak-HBM differences -0.05 ± 0.04, -0.08 ± 0.05, and -0.12 ± 0.06 GB. This frozen-checkpoint panel isolates the byte-matched representation and does not compare the methods' training trajectories. The panel is descriptive and makes no multiplicity-adjusted superiority claim.

E Mechanism attribution and robustness

E.1 Matched retraining contracts and frozen replays

The six mechanism variants use the matched training contract in Table 24. Each is a 3B continuation in four seeds, giving a 72B mechanism block. Within a seed, the variant and its reference share the same parent, token order, primary-token budget, optimizer schedule, training length, and terminal endpoint. Table 24 gives the terminal retraining effects. The frozen replays in Table 3 instead isolate execution while holding assignments fixed; their estimands are not interchangeable.

Table 24: Matched $E = 32$ retraining ablations over four paired seeds. Every variant receives 3B tokens with the same parent, token order, optimizer, and endpoint. Quality is variant minus the full method; prefill and wire are variant/full. The full-method reference reuses the first four primary seeds.

Variant	GLUE Δ	PPL Δ	Prefill	F+B wire
StaticShared	-0.28 ± 0.13	$+0.22 \pm 0.10$	0.98 ± 0.018	1.06 ± 0.023
Balanced random groups	-0.51 ± 0.18	$+0.47 \pm 0.13$	0.95 ± 0.024	1.10 ± 0.031
No activation statistic	-0.12 ± 0.09	$+0.10 \pm 0.07$	0.99 ± 0.014	1.02 ± 0.019
No co-routing statistic	-0.10 ± 0.08	$+0.08 \pm 0.06$	0.97 ± 0.017	1.08 ± 0.026
Flat global router	$+0.02 \pm 0.06$	-0.01 ± 0.04	0.96 ± 0.019	1.01 ± 0.017
No flat-mass distillation	-0.17 ± 0.10	$+0.16 \pm 0.08$	0.98 ± 0.017	1.04 ± 0.023

E.2 Factorial interactions

Marginal ablations do not distinguish additive effects from interactions. We therefore train the complete $2^5 = 32$-cell factorial at $E = 32$ for each of four seeds, with 2.5B token positions per cell and a total budget of 320B. The factors are dynamic grouping (D), activation statistics (A), co-routing statistics (C), hierarchy (H), and topology-aware ownership (P), coded $x_f \in \{-1, +1\}$. For every nonempty factor subset $q \subseteq \{D, A, C, H, P\}$ and seed s, the high-minus-low effect is

$$\widehat{\theta}_{q,s} = \sum_{c=1}^{32} w_{q,c} Y_{c,s}, \qquad w_{q,c} = 2^{-4} \prod_{f \in q} x_{f,c}.$$

Together with the design matrix below, this formula gives every entry of all 31 non-intercept contrast vectors; hence the five main, ten two-way, and 16 higher-order effects are all identified without aliasing. GLUE and PPL use raw differences; throughput uses percentage-ratio points $100(T/T_{\text{Switch}} - 1)$; and wire is computed on the log-ratio scale. The table reports the exponentiated contrast on the ratio scale. The 32 cells remain nested within seed and do not increase n.

Table 25: Complete 32×5 factorial design matrix. Minus and plus denote -1 and $+1$; the two blocks show cells 1–16 and 17–32.

Cell	D	A	C	H	P	Cell	D	A	C	H	P
1	−	−	−	−	−	17	+	−	−	−	−
2	−	−	−	−	+	18	+	−	−	−	+
3	−	−	−	+	−	19	+	−	−	+	−
4	−	−	−	+	+	20	+	−	−	+	+
5	−	−	+	−	−	21	+	−	+	−	−
6	−	−	+	−	+	22	+	−	+	−	+
7	−	−	+	+	−	23	+	−	+	+	−
8	−	−	+	+	+	24	+	−	+	+	+
9	−	+	−	−	−	25	+	+	−	−	−
10	−	+	−	−	+	26	+	+	−	−	+
11	−	+	−	+	−	27	+	+	−	+	−
12	−	+	−	+	+	28	+	+	−	+	+
13	−	+	+	−	−	29	+	+	+	−	−
14	−	+	+	−	+	30	+	+	+	−	+
15	−	+	+	+	−	31	+	+	+	+	−
16	−	+	+	+	+	32	+	+	+	+	+

Table 26: Complete main and two-way point estimates from the independently retrained 32-cell full-factorial design over four seeds. Throughput is in percentage points relative to Switch; wire is an exponentiated log-ratio contrast.

Effect	GLUE	PPL	Throughput pp	Wire ratio
D	$+0.24$	-0.21	$+1.4$	0.980
A	$+0.09$	-0.06	$+0.7$	0.990
C	$+0.07$	-0.05	$+0.8$	0.970
H	-0.02	$+0.01$	$+4.7$	0.930
P	$+0.00$	$+0.00$	$+2.3$	0.820
D×A	$+0.07$	-0.03	$+0.4$	0.996
D×C	$+0.05$	-0.05	$+0.5$	0.988
D×H	0.00	0.00	$+0.2$	0.993
D×P	0.00	0.00	$+0.1$	0.991
A×C	$+0.03$	-0.02	0.0	0.994
A×H	$+0.01$	0.00	$+0.1$	0.995
A×P	0.00	≈ 0.00	0.0	0.998
C×H	$+0.01$	-0.01	$+0.3$	0.989
C×P	0.00	≈ 0.00	$+0.5$	0.985
H×P	0.00	0.00	$+1.5$	0.960

The $H \times P$ throughput interaction of $+1.5$ percentage points and wire contrast of 0.960 are consistent with hierarchy and ownership affecting the same communication path: hierarchy determines which groups are touched, while ownership determines how those groups share destinations. The $C \times P$ wire contrast of 0.985 similarly links co-routing structure to owner reuse. These descriptive contrasts quantify dependence between interventions; the frozen assignment/owner replays in Table 3 provide the complementary interface-specific evidence.

The table reports the five main and ten two-way contrast point estimates from the complete design. The four seeds remain the outer units; no multiplicity-adjusted decision is asserted from this descriptive panel.

Table 27: Complete third- and higher-order point estimates from the 32-cell design over four seeds. Throughput is in percentage points and wire is the exponentiated log-ratio contrast.

Effect	GLUE	PPL	Throughput pp	Wire ratio
D×A×C	$+0.018$	-0.012	$+0.8$	0.986
D×A×H	$+0.010$	-0.007	$+0.6$	0.990
D×A×P	$+0.006$	-0.004	$+0.5$	0.992
D×C×H	$+0.012$	-0.009	$+0.7$	0.988
D×C×P	$+0.004$	-0.003	$+0.4$	0.994
D×H×P	$+0.003$	-0.002	$+0.5$	0.989
A×C×H	$+0.008$	-0.006	$+0.6$	0.995
A×C×P	$+0.002$	-0.001	$+0.3$	0.991
A×H×P	-0.001	$+0.001$	$+0.2$	0.996
C×H×P	$+0.005$	-0.004	$+0.5$	0.992
D×A×C×H	$+0.007$	-0.005	$+0.4$	0.993
D×A×C×P	$+0.001$	0.000	$+0.3$	0.997
D×A×H×P	-0.003	$+0.002$	$+0.2$	0.998
D×C×H×P	$+0.004$	-0.003	$+0.4$	0.995
A×C×H×P	0.000	$+0.001$	$+0.3$	0.997
D×A×C×H×P	$+0.002$	0.000	$+0.2$	0.998

The ranges are GLUE $[-0.003, +0.018]$, PPL $[-0.012, +0.002]$, throughput $[+0.2, +0.8]$ points, and wire $[0.986, 0.998]$. These high-order values are reported as point estimates rather than multiplicity-adjusted decisions.

E.3 Sensitivity response surface

A separate report-only response surface uses five paired seeds and varies rank, group size, bridge length, and expansion cap one factor at a time. The shared default and ten non-default configurations each receive 3B continuation tokens per seed, totaling 165B. This frozen design is not used to select the primary operating point. The default reuses the same run on both sides, so its self-contrast has zero variance. Relative to the default, rank $r = 4/8/16$ changes PPL by $+0.22 \pm 0.10/0.00 \pm 0.00/-0.08 \pm 0.10$. Group size $K = 4/8/16$ changes PPL by $-0.10 \pm 0.10/0.00 \pm 0.00/+0.22 \pm 0.12$ and prefill by $0.96 \pm 0.03/1.00 \pm 0.00/1.05 \pm 0.04$. For bridge lengths $H = 16/32/64/128$, the maximum path PPL gaps are $0.28 \pm 0.10/0.19 \pm 0.08/0.12 \pm 0.06/0.09 \pm 0.05$, while complete-run overheads are $0.22 \pm 0.08/0.28 \pm 0.10/0.35 \pm 0.12/0.55 \pm 0.18\%$. Expansion caps $0/2/4/8$ retain flat mass $0.940 \pm 0.010/0.975 \pm 0.006/0.986 \pm 0.004/0.992 \pm 0.003$. The trends expose the expected trade-offs: larger groups increase base reuse but eventually harm quality; longer bridges smooth the path but add work; and aggressive support caps reduce scoring at the cost of omitted mass.

E.4 Bridge and publication checks

The two continuous bridge endpoints are statistical measurements; the remaining rows in Table 28 are hard implementation checks. The NLL reducer compares candidate and old branches on the same slice. The PPL reducer compares the gradual candidate with a non-updating immediate-swap sidecar on that slice. Neither pools tokens or slices as replicates.

Table 28: Bridge path and publication results at $E = 128$. The first two rows use four seeds; hard checks aggregate the declared stress or fault units shown in the final column.

Quantity	Details
Maximum candidate-minus-old NLL	**Result:** 0.010 ± 0.003 nats/token **Required condition:** < 0.020 **Unit / reducer:** seed max over accepted events and slices

Continued on next page

Table 28 (continued)

Quantity	Details
Maximum gradual–immediate-swap PPL gap	**Result:** 0.12 ± 0.04 **Required condition:** < 0.25 **Unit / reducer:** seed max over accepted events and slices
Exact-k violations	**Result:** 0 **Required condition:** $= 0$ **Unit / reducer:** all 720 stress units and every injected assignment boundary
Capacity violations / silent drops	**Result:** $0/0$ **Required condition:** $= 0/ = 0$ **Unit / reducer:** all 720 stress units and every injected assignment boundary
Mixed-epoch publications	**Result:** 0 **Required condition:** $= 0$ **Unit / reducer:** every injected transaction boundary
Cursor/RNG/token-hash divergence	**Result:** 0 **Required condition:** $= 0$ **Unit / reducer:** every transaction
Unrecoverable journal steps	**Result:** 0 **Required condition:** $= 0$ **Unit / reducer:** every injected boundary
Accumulator/finalizer hash mismatches	**Result:** 0 **Required condition:** $= 0$ **Unit / reducer:** every injected boundary
Extra or missing optimizer updates	**Result:** 0 **Required condition:** $= 0$ **Unit / reducer:** both branches update once
Dual or empty branch selections	**Result:** 0 **Required condition:** $= 0$ **Unit / reducer:** exactly one branch retained
Requests served from interpolation state	**Result:** 0 **Required condition:** $= 0$ **Unit / reducer:** old epoch until activation CAS

E.5 CAPACITY, TOPOLOGY, AND FAULT BOUNDARIES

The assignment stress design fixes $E = 128$ and sequence length 2,048, then crosses capacity factors $\{1.0, 1.1, 1.25, 1.5\}$, Zipf exponents $\{0, 1.0, 1.5, 2.0\}$, group sizes $\{4, 8, 16\}$, and $k \in \{1, 2, 4\}$. The 144 configurations each use five independent counter-hashed outer blocks, giving 720 configuration–block units; within a unit, methods share the expert permutation and Gumbel stream. This grid is a hard feasibility/invariant stress rather than a throughput or routing-tail endpoint: exact-k, capacity, and silent-drop counts are the declared outcomes. These tests cover only the sampled grid and do not imply feasibility if the aggregate capacity assumption in Section 3.4 is violated. Across all 720 declared units, exact-k violations, capacity violations, and silent drops are respectively $0/0/0$.

Topology degradation disables one high-bandwidth path, rate-limits a declared inter-node link, and injects a deterministic slow rank while holding the accepted assignment trace fixed. The placement solve is rerun only from the perturbed calibration matrix and is evaluated on request-disjoint traces. This diagnoses adaptation of the fixed-charge proxy; it does not convert the proxy into a guarantee about elapsed communication.

Fault injection covers checksum mismatch, stale parent epoch, failed SVD, selector-gate failure, rank slowdown, missing payload, coordinator loss, partial activation, rank restart, replayed nonce, lost acknowledgement, and network partition. The suite executes 480 injected transactions. For each of four seeds, the ten phase-agnostic classes are injected before and after FENCED, SHADOW-READY, final SLICE-H, UPDATE-READY, and the activation CAS (400 transactions); failed-SVD and selector-gate faults are instead injected at their call sites in ten counter-hashed candidate events per seed (80

transactions). Each case either aborts while the old epoch remains visible or recovers from the latest complete snapshot. The zero mixed-epoch, exact-k, capacity, silent-drop, and journal divergences in Table 28 are conditional on this fault model and on at least one durable quorum replica retaining the latest complete snapshot.

E.6 DISTRIBUTION-SHIFT CONTROLS

Using the visit, decision, qualifying-commit, and signed-PPL reducers in Section B.6, the four forks freeze their manifests before any outcome is inspected. Composition changes only the five source proportions; PMC is a source-disjoint biomedical replacement; the temporal condition uses license-filtered records first published from 2025-10-01 through 2025-12-31; and the multilingual condition uses equal Chinese/Spanish token mass with language-ID posterior at least 0.95. Document hashes, normalized 13-gram filters, and sealed terminal shards prevent overlap with all parent, pre-shift, assessment, and benchmark roles.

Table 29: Adaptation under four prespecified shifts. "Saved" and "gain" are no-regroup minus DYCLUST-MOE paired-seed effects; gap is $\mathrm{PPL}_{\text{DyClust-MoE}} - \mathrm{PPL}_{\text{Switch}}$. Bounds are one-sided Bonferroni limits for one $n = 6, m = 4$ family per response. Worst endpoints are maxima of the six seed-level reducers. Means and SDs are rounded only for display; all bounds use the unrounded paired-seed values. TV is terminal total routing TV.

Shift	Results and bounds
Composition	**Decisions saved / LCB:** $3.2 \pm 1.5/1.266$
	PPL gain / LCB: $0.22 \pm 0.12/0.069$
	Dy–Switch gap / UCB (ceiling): $0.11 \pm 0.04/0.163$ (0.20)
	Worst visits / decisions / gap: $2/12/0.18$
	TV mean / p95: $0.082/0.137$
PMC biomedical	**Decisions saved / LCB:** $3.8 \pm 1.6/1.764$
	PPL gain / LCB: $0.28 \pm 0.15/0.094$
	Dy–Switch gap / UCB (ceiling): $0.18 \pm 0.06/0.256$ (0.30)
	Worst visits / decisions / gap: $3/19/0.27$
	TV mean / p95: $0.088/0.145$
2025-Q4 temporal	**Decisions saved / LCB:** $4.2 \pm 1.5/2.266$
	PPL gain / LCB: $0.27 \pm 0.14/0.091$
	Dy–Switch gap / UCB (ceiling): $0.16 \pm 0.05/0.223$ (0.30)
	Worst visits / decisions / gap: $3/19/0.25$
	TV mean / p95: $0.086/0.143$
Chinese/Spanish	**Decisions saved / LCB:** $5.3 \pm 1.8/3.072$
	PPL gain / LCB: $0.43 \pm 0.18/0.198$
	Dy–Switch gap / UCB (ceiling): $0.24 \pm 0.07/0.328$ (0.40)
	Worst visits / decisions / gap: $4/26/0.37$
	TV mean / p95: $0.094/0.149$

Simultaneous lower bounds are positive for decisions saved and terminal PPL gain in all four shifts; every terminal gap UCB remains inside its practical ceiling. Thus improvement is supported across all four prespecified shifts, rather than only by positive point estimates. These controlled shifts span mixture, domain, time, and language axes, but do not establish open-ended continual learning or subgroup fairness.

F COMPLETE OLMOE TRANSFER CONTRACT

The public-model experiment is not a scale-up of the small-model constants, and no unlisted small-model event constant is inherited implicitly. Table 30 states the architecture-specific contract. It keeps OLMoE's all-64 softmax and unnormalized native top-8 weights, so the grouped adapter reduces accepted support, bases touched, and placement traffic but does not reduce expert-logit computation.

Let p_t be OLMoE's all-64 softmax, T_t its native top-8 set, and A_t the final repaired set. Native and compact routed weights are sub-probability measures $r_t(i) = p_t(i)\mathbf{1}\{i \in T_t\}$ and $\widehat{r}_t(i) = p_t(i)\mathbf{1}\{i \in A_t\}$. Their exact discrepancy is

$$\|r_t - \widehat{r}_t\|_1 = p_t(T_t \setminus A_t) + p_t(A_t \setminus T_t). \tag{20}$$

This identity, rather than normalized TV in Equation (10), governs the native transfer. It preserves missing mass explicitly and makes no calibration claim about renormalized accepted weights.

Table 30: OLMoE-specific experiment contract. Protocol constants are separated from the measured results in the text.

Component	Contract
Parent and control	`allenai/OLMoE-1B-7B-0924` commit `6d84c48581ec e794365f2b8e9cfb043c68ade9c5`; ten compact/native pairs are continued independently, and each branch receives 4B primary token positions with matched optimizer, data order, and update mask.
Continuation data	A fixed Dolma-v1.7-derived manifest uses the same five source strata and normalized-record/13-gram benchmark exclusion rule as the small-model continuation. Documents are ordered by a counter hash of pair seed and normalized document hash; the compact and native branches consume exactly the same packed token IDs and masks.
Optimizer	Fresh AdamW with $(\beta_1, \beta_2, \epsilon) = (0.9, 0.95, 10^{-8})$, peak learning rate 10^{-4} after a matched linear warm-up, cosine decay to zero, matrix-only weight decay 0.10, gradient clip 1.0, and zero dropout; forward/backward tensors are BF16 and master weights and moments are FP32.
Representation	16 sparse layers, $d = 2048$, routed $d_{\text{ff}} = 1024$, $E = 64$, native top-8, $G = 8$, $K = 8$, and rank 32 for gate, up, and down matrices; deterministic base refit plus truncated SVD initializes from the pinned tensors.
Objective and refit	$\epsilon_w = 10^{-12}$, $\lambda_{\text{act}} = 0.25$, $\lambda_{\text{pair}} = 0.10$, a 256-bin sketch, eight outer refit sweeps, and an $8E = 512$ exact-swap cap. The cold refit stops at relative reconstruction decrease $\leq 10^{-5}$ and fails if the eighth sweep does not meet it. An eligible swap requires incumbent objective $\mathcal{J}_{\text{grp}}^{\text{old}} > 10^{-6}$ and decrease $\geq \max(0.005\,\mathcal{J}_{\text{grp}}^{\text{old}}, 10^{-7})$. Exact refits use CUDA 12.4 cuSOLVER FP64 `gesvdj`, tolerance 10^{-12}, at most 100 Jacobi sweeps, descending singular values, the largest-magnitude-right-vector-entry sign rule, and stable discrete ties.
Event schedule	After warm-up, 99 opportunities are evenly spaced through the 4B-token continuation in round-robin layer order. The preceding document history is counter-hash split $7/8$ for proposal and $1/8$ for proposal-disjoint screening. Event gates are mean reconstruction ≤ 0.02, per-expert tail ≤ 0.08, cumulative layer budget ≤ 0.06, and screening output drift ≤ 0.015.
Selector preflight	An overlap-remapped private bias-free selector receives 64 shadow steps on $7/8$ of the next ordinary batch; the untouched $1/8$ requires KL ≤ 0.02, retained mass ≥ 0.98, and native-top-8 recall ≥ 0.99. Failure discards selector weights and both moment buffers.
Routing and capacity	Start from two groups, expand to selector mass 0.995 or at most six additional groups, preserve all reserve groups, and use owner-global $C_{i\delta} = \lceil 1.25 \cdot 8N_\delta/64 \rceil$. Repair uses $R_{\max} = 16$ evaluation rounds, counting rejected batches. Accepted edges retain native unnormalized weights.

Continued on next page

Table 30 (continued)

Component	Contract
Bridge and publication	One scheduled 1,048,576-token primary update is partitioned into $H = 128$ real continuation slices of 8,192 global tokens. Twin branches consume the same IDs and each prepares exactly one optimizer/scheduler update; activation requires maximum candidate-minus-old NLL < 0.020 nats/token and gradual–immediate-swap absolute PPL gap < 0.25, followed by the same immutable-generation and epoch-CAS publication protocol. New base/factor moments start at zero with a 256-update local learning-rate ramp.
Placement roles	Link-fit probes, disjoint calibration-validation probes, request-disjoint solver-calibration traces, and final evaluation traces are four mutually disjoint roles. Only validation supplies the $R^2 \geq 0.95$/p95-error $\leq 5\%$ gate, only solver calibration supplies objective counts, and only final traces support reported systems effects.
Implementation parity	The uncompressed adapter matches native router logits, top-8 IDs/weights, and gate/up/down outputs; the compact fused path matches its materialized compact implementation; serialization preserves the declared tensor inventory. Float tolerances are 10^{-6} absolute and 10^{-5} relative.
Systems hardware/workload	Eight nodes each contain eight H100-SXM-80GB GPUs joined by NVSwitch and dual 400-Gb/s NDR InfiniBand GPUDirect-RDMA links; CUDA 12.4, NCCL 2.21.5, and PyTorch 2.5.1 are fixed. Prefill uses 64 sequences of length 2,048. Decode uses batch 32, a 2,048-token prompt, and 256 generated tokens timed end-to-end. Each of ten pairs reduces five process blocks after 200 warm-ups and 1,000 measured repetitions.

The parent contains 6,442,450,944 expert-matrix and 476,710,912 unchanged scalars. The compact model contains 805,306,368 group-base, 301,989,888 rank-32 factor, 476,710,912 unchanged, and 262,144 selector scalars, for 1,584,269,312 total and an exact scalar ratio of 0.2289683845. At BF16 this is 3,168,538,624 tensor bytes before container metadata. For h accepted groups, the exact active expert-matrix ratio is

$$\frac{h(3dd_{\text{ff}}) + 8 \cdot 3r(d + d_{\text{ff}})}{8(3dd_{\text{ff}})} = \frac{h}{8} + 0.046875.$$

For $h = 1/2/3/4/8$, this gives $0.171875/0.296875/0.421875/0.546875/1.046875$. The final accepted expert set touches 1.35 ± 0.10 distinct group bases on average (p95 at most two); this is not the number of selector-opened candidate groups. Flat-top-8 candidate recall and retained mass are both at least 0.995, and reserve edges are below 0.5%. Because all 64 native logits are intentionally scored, these diagnostics rule out only an unreported full-support expert/base execution fallback, not score-all logit computation.

Table 31: OLMoE transfer over ten compact/native pairs. Bounds use one-sided Bonferroni t_9, with $m = 3$ separately for quality and systems. Native/compact columns give absolute quality means; effects are compact minus native for quality and HBM, and compact/native for systems.

Endpoint	Native / compact	Effect $\pm$ SD	Bound	Margin
MMLU points	$54.12/53.94$	-0.18 ± 0.27	LCB -0.394	> -0.60
HellaSwag points	$79.68/79.55$	-0.13 ± 0.23	LCB -0.313	> -0.60
WikiText PPL	$11.43/11.55$	$+0.12 \pm 0.06$	UCB 0.168	< 0.25
Prefill ratio	–	1.17 ± 0.031	LCB 1.145	> 1.05
Decode ratio	–	1.14 ± 0.029	LCB 1.117	> 1.05
F+B wire ratio	–	0.77 ± 0.029	UCB 0.793	< 0.90
Peak HBM (GB/GPU)	–	-1.36 ± 0.17	descriptive	–

All prespecified quality and systems bounds satisfy their practical margins: task-score and PPL changes remain within tolerance, while prefill, decode, and F+B-wire improvements exceed the required thresholds. These results test compatibility with one public native top-8 architecture and do not by themselves establish the trade-off for frontier-scale or differently routed models.

G Complete Qwen1.5-MoE transfer contract

Qwen1.5-MoE supplies a second public architecture with fewer routed experts, native top-4 selection, unnormalized accepted weights, and an always-on shared expert (Qwen Team, 2024). The shared expert, its gate, attention, embeddings, and normalization tensors are unchanged and excluded from every routed-expert ratio below.

Table 32: Qwen1.5-MoE-A2.7B-specific contract. No small-model or OLMoE event constant is inherited unless printed here.

Component	Contract
Parent and architecture	`Qwen/Qwen1.5-MoE-A2.7B` commit `1a758c50ecb6350748b9ce0a99d2352fd9fc11c9`; 24 sparse layers, $d = 2048$, 60 routed experts, native top-4, routed $d_{\text{ff}} = 1408$, and `norm_topk_prob=false`. The native shared expert and shared-expert gate remain unchanged.
Pairs and exposure	Ten compact/native pairs are continued independently. Each branch receives 4B primary token positions with identical packed IDs, masks, optimizer schedule, and update masks; auxiliary reads use the separate ledger in Section B.
Continuation data	A frozen license-filtered English/Chinese manifest uses 55% web/books/reference English, 25% Chinese web/reference, 10% code, and 10% scholarly text. Normalized-record and 13-gram filters remove MMLU, C-Eval, CMMLU, HellaSwag, and WikiText evaluation overlap before either branch runs.
Optimizer	Fresh AdamW $(0.9, 0.95, 10^{-8})$, peak learning rate 2×10^{-5} after a matched linear warm-up, cosine decay to zero, matrix-only weight decay 0.10, clip 1.0, BF16 forward/backward, and FP32 master weights and moments.
Representation	$G = 10$, $K = 6$, rank 32 for each routed gate, up, and down matrix. Deterministic full-SVD/base refits initialize only routed experts; the shared expert is neither grouped nor counted as a reserve edge.

Continued on next page

Table 32 (continued)

Component	Contract
Event schedule and gates	After warm-up, 99 opportunities are evenly spaced through the 4B-token continuation in round-robin layer order. Each event uses a 256-bin sketch, eight refit sweeps, an $8E = 480$ exact-swap cap, and 64 selector shadow updates. The preceding document history is split $7/8$ proposal and $1/8$ screen. Gates are event RMS ≤ 0.025, expert tail ≤ 0.10, cumulative layer ledger ≤ 0.08, output drift ≤ 0.018, selector KL ≤ 0.02, retained native mass ≥ 0.98, and native-top-4 recall ≥ 0.99.
Routing/capacity	Open two groups, expand to selector mass 0.995 or six additional groups, retain reserve groups, and repair to exactly four routed experts at capacity factor 1.25 with $R_{\max} = 16$ evaluation rounds, including rejected batches. All 60 logits and their softmax are evaluated; final accepted weights remain unnormalized. The shared expert executes independently on every token.
Bridge/publication	One scheduled 1,048,576-token batch is split into $H = 128$ slices. Twin branches share IDs and randomness; maximum candidate-minus-old NLL must be < 0.020 and gradual–immediate PPL gap < 0.25. One epoch CAS publishes routed tensors, selector, owner map, private moments, cursor, and RNG together.
Systems path	Eight eight-H100-SXM-80GB nodes, NVSwitch, dual 400-Gb/s NDR, CUDA 12.4, NCCL 2.21.5, and PyTorch 2.5.1. Prefill and decode match the OLMoE workloads; ten pairs each reduce five blocks after 200 warm-ups and 1,000 measured repetitions. Placement uses the pinned Gurobi configuration and disjoint trace roles in Section A.

Across 24 layers, the native routed gate/up/down matrices contain 12,457,082,880 scalars. Compact routed storage contains 2,076,180,480 group-base, 477,757,440 residual-factor, and 491,520 selector scalars, totaling 2,554,429,440 and giving an exact routed-representation ratio of 0.2050583965. The unchanged embedding, attention, normalization, shared-expert, and output tensors contain 1,858,701,312 scalars. Thus the complete native/compact model inventories are 14,315,784,192/4,413,130,752 scalars (ratio 0.3082702766), or 28,631,568,384/8,826,261,504 analytical BF16 tensor bytes before allocator and container overhead. For h distinct bases touched by the final accepted four-expert set, the active routed-matrix ratio is

$$\frac{h}{4} + \frac{32(2048 + 1408)}{2048 \cdot 1408} = \frac{h}{4} + 0.03835227. \tag{21}$$

The accepted set touches 1.62 ± 0.12 distinct bases on average (p95 two), so the mean analytical active ratio is 0.44335227. These are routed-expert matrix ratios, not whole-model storage, executed FLOPs, or HBM ratios.

Table 33: Qwen transfer over ten pairs. Quality uses one-sided Bonferroni t_9 with $m = 5$; systems uses $m = 3$. Native/compact columns give absolute quality means; effects are compact minus native for quality and HBM, and compact/native for systems.

Endpoint	Native / compact	Effect ± SD	Bound	Margin
MMLU points	$63.27/63.11$	-0.16 ± 0.24	LCB -0.374	> -0.60
C-Eval points	$68.43/68.31$	-0.12 ± 0.22	LCB -0.316	> -0.60
CMMLU points	$67.86/67.67$	-0.19 ± 0.25	LCB -0.413	> -0.60
HellaSwag points	$79.21/79.11$	-0.10 ± 0.21	LCB -0.287	> -0.60
WikiText PPL	$10.72/10.86$	$+0.14 \pm 0.06$	UCB 0.194	< 0.25
Prefill ratio	–	1.15 ± 0.028	LCB 1.128	> 1.05
Decode ratio	–	1.13 ± 0.027	LCB 1.109	> 1.05
F+B wire ratio	–	0.78 ± 0.025	UCB 0.800	< 0.90
Peak HBM (GB/GPU)	–	-2.21 ± 0.28	descriptive	–

C-Eval and CMMLU use their official zero-shot multiple-choice protocols and macro averages (Huang et al., 2023; Li et al., 2023). All four task-score bounds, the PPL bound, and the three systems bounds satisfy their prespecified practical margins. The unchanged shared expert prevents attributing the effect to deleting shared capacity.

H DESCRIPTIVE $E = 256$ EXTENSION

The $E = 256$ quality and full-run-training extension uses six paired training seeds. Each displayed Switch/DYCLUST-MOE quality entry is the between-seed mean±SD; pairing is used for contrasts but this panel remains descriptive. It starts from the same dense parent and each branch receives 2B continuation token positions. Stored scalars are 1,248,552,192 for DYCLUST-MOE and 8,533,861,376 for Switch. Grouping opportunities are evenly spaced through the declared continuation at 99 prespecified positions and stop before the terminal boundary, using the preceding document history with the same $7/8$ proposal and $1/8$ proposal-disjoint split. If the final event activates, its local learning-rate ramp advances only through the remaining declared schedule and the terminal checkpoint reports that realized state without extrapolation. Full-run throughput uses the same two-node A100-80GB TP=PP=1, DP=1/EP=8 mapping as the $E = 128$ primary run and times the complete 2B-token continuation including all method overhead. The 24 GB/8k and 40 GB/16k results are allocator caps on this same A100 mapping, not physical A10 or 24/40-GB accelerator results. Measured Switch/DYCLUST-MOE quality results are MMLU $31.5 \pm 0.6/31.3 \pm 0.6$, HumanEval pass@1 $6.0 \pm 0.9/5.8 \pm 0.9$, and WikiText PPL $27.8 \pm 0.12/28.0 \pm 0.13$. On the fixed two-node A100 mapping, the $E = 256$ full-run training rate remains in the same order of magnitude as the $E = 128$ primary endpoint. The shorter continuation and descriptive status prevent this extension from replacing any primary $E \leq 128$ result.

I EXTENDED LIMITATIONS

Representation and grouping. The grouping objective uses a first-moment activation summary, normalized reconstruction, and a heuristic pair cut. It neither proves semantic equivalence of nonlinear FFNs nor finds a global discrete optimum. A candidate can satisfy every event-local distortion gate while repeated accepted events still move the model in ways not captured by the additive ledger. Equal-size groups and fixed residual rank may be poor choices for heterogeneous experts.

Routing and assignment. The normalized mass decomposition does not govern either public transfer. OLMoE and Qwen retain native unnormalized accepted weights under the contracts in Sections F and G; Equation (20) gives the applicable sub-probability identity, and Qwen's shared expert executes outside the grouped routed path. Exact-k reserve feasibility proves token and expert degrees under its capacity assumptions; it does not imply high routing score, calibrated probabilities, balanced realized work, or low tail latency. Support expansion can approach score-all behavior under a severe shift.

Bridge and publication. The bridge bounds a private, frozen-routing path for one fenced primary batch and does not prove long-run optimizer stability after activation. Crash safety is conditional on immutable generation-addressed buffers, epoch-tagged requests, a linearizable activation CAS, and at least one durable quorum replica with the latest complete snapshot. The experiments test the stated injected faults, not all failures of a production storage or consensus system.

Placement and hardware. Placement optimizes a calibrated fixed-charge critical-owner proxy. A change in message-size distribution, topology, collective implementation, or contention can invalidate that calibration. Analytical active scalars, executed FLOPs, serialized bytes, HBM, wire bytes, latency, and energy are distinct estimands; none can be substituted for another. The A100 cap study is not evidence about all 24-GB devices, and the physical A10 result does not establish behavior on newer accelerators. The measured 1.051 ± 0.008 executed-work ratio against nominal-matmul Switch accompanies lower traffic on the evaluated systems. The balance between these costs depends on arithmetic throughput, memory bandwidth, interconnect bandwidth, and overlap, so the magnitude of the speedup is specific to the hardware and workload contract.

Scale and generalization. The primary study uses an English 12-layer model on A100 hardware, and transfer covers two pinned public checkpoints on H100. The $E = 256$ extension has a shorter continuation. These experiments do not establish frontier-model behavior, multilingual quality beyond the stated Qwen tasks, broad downstream safety, or a universal quality–storage Pareto improvement. The four controlled shifts are not evidence of open-ended continual learning.

Lifecycle accounting. Smaller inference artifacts can reduce replication and traffic, but refits, shadow routing, bridges, control runs, and confirmatory evaluation consume energy and temporary storage. Environmental benefit depends on the workload horizon over which those development and adaptation costs are amortized; a steady-state per-token ratio alone is insufficient.

Statistical scope. Paired designs remove substantial seed and checkpoint variation, but four, five, six, or ten outer units still provide limited resolution for small effects and tail behavior. The one-sided t procedures used for shifts and public-model transfers rely on the distribution of outer-unit paired differences, while the Bonferroni bounds are conservative. Secondary capability, sensitivity, stress, and $E = 256$ results are descriptive and should not be treated as additional confirmatory families or independent replications.

Comparator scope. The primary quality comparison concerns 3B continuation tokens after dense-parent pretraining, with an equal architecture-local pilot budget. Relative quality is interpreted at that training horizon. Common exact-k reserve/repair and zero-drop controls hold the assignment-feasibility regime fixed; destination coalescing and placement are evaluated under the stated runtime contract. These controls support interface-level attribution within that contract. The precision, hardware, kernels, and workloads determine the comparison with other compression and execution systems.

OC-Share isolates a fixed output partition under the same tensor inventory; it is not an upstream Sub-MoE reproduction. The continuation-matched D^2-MoE port includes implementation choices beyond its original deployment. Parity, serializer, and tensor tests reduce this ambiguity but do not make either port identical to every upstream deployment. Native Expert Choice and StableMoE keep their different assignment semantics and therefore serve only as contextual neighbors, not exact-k substitutions or evidence of state-of-the-art rank.

Data and contamination boundary. The normalized-record and 13-gram filters remove declared overlap with the evaluation sets but cannot rule out paraphrases, shared facts, or undocumented upstream contamination. The primary corpus mixture is English-dominant and its license metadata and source categories do not establish representativeness. The four shifts cover mixture, biomedical-domain, recent-time, and two-language conditions, but remain curated forks rather than a general adaptation or fairness benchmark.

Reimplementation boundary. Reimplementation must match the tie rules, reducers, workloads, versions, and tensor inventories, while SVD behavior, collective scheduling, filesystem durability,

and kernels may still change outcomes. Determinism is scoped to CUDA 12.4 cuSOLVER FP64 `gesvdj` with the specified tolerance, sweep cap, ordering/sign rules, and reduction order. Reproduce structural invariants first; measured ratios remain hardware- and workload-conditional.

Deployment and governance. Deployment should freeze a checksummed version of grouping, compact tensors, router, owner map, and precision; subsequent adaptation requires evaluation and rollback. Compression may reduce deployment cost, while group or owner changes can redistribute quality and latency even when aggregate gates pass. Deployment also requires license and data-provenance review, privacy controls, abuse monitoring, subgroup exposure audits, rollback authority, and explicit energy budgets. These experiments do not establish model safety or fairness.