INTERVENING BEFORE SELECTION: COUNTERFACTUAL EFFECTS IN ADAPTIVE SPARSE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Adaptive sparse models select which components participate in a computation. A component's contribution to the selected computation can therefore differ from the model's dependence on its availability. We formalize this distinction through suppression after selection and exclusion before selection, which admits eligible alternatives. Across mixtures of experts, dynamic token selection, and fitted sparse semantic fields, the two interventions yield opposite effect signs and different maximizing components, even where rankings by representation change nearly coincide, and recomputing subsequent expert assignments can reverse their loss ordering. We derive a sufficient condition for invariance of the maximizing component across interventions and use suppression ranking margins to allocate additional measurements under a reference intervention. On inputs separate from threshold calibration, this allocation reduces disagreement with the reference maximizer relative to uniform random allocation at the same number of additionally measured groups, and motivates calibration objectives that also account for the effect gaps of the remaining errors.

1 INTRODUCTION

A central challenge in sparse modeling is to select and combine a limited set of components while preserving task performance under computational or representational constraints. Adaptive expert networks select expert transformations for each input (Dai et al., 2024; Muennighoff et al., 2025; Shi et al., 2025), dynamic vision transformers retain patch tokens for subsequent processing (Rao et al., 2021), and sparse semantic fields express Gaussian features through codebook entries selected during scene fitting (Li et al., 2025). Deciding which components to retain requires understanding how their joint computation supports model behavior. Importance estimates guide expert pruning (Lu et al., 2024; Huang et al., 2025; Lasby et al., 2026) which are essential for improving model efficiency.

To estimate the importance of a target component, a common way is to intervene on the target and measure the response to a specified modification (Covert et al., 2021). However, the intervention operation and response measurement do not directly make the importance of the target component precisely interpretable. It instead depends on how the intervention is implemented and on how the surrounding computation adjusts to carry out the subsequent processing. For example, if an expert in MoE is ablated after routing, its contribution is removed and the downstream computation is affected only by the remaining experts. However, if the same expert is ablated before routing, another expert is selected and the downstream computation differs from that in the former case.

For example, existing research shows that the choice of corruption changes which components activation patching identifies in dense transformers (Zhang & Nanda, 2024). Some other work studies how the computational relationships among components affect functional interpretations of the intervention response. A component's influence depends on the alternatives available alongside it, the coefficients combining their outputs, and the subsequent computation receiving those outputs. Coalition-based importance estimates account for the first of these dependencies (Huang et al., 2025), while pruning analyses examine substitution and renormalization (Lasby et al., 2026). Later computation can further compensate for a local perturbation (Rushing & Nanda, 2024). Accounting for these dependencies within an intervention, however, does not establish whether the resulting attribution is stable across removal specifications. Therefore, an importance measurement reflects

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.

both the targeted component and the response of the surrounding computation. Specifically, we are interested in the following questions: Which conclusions about component function remain reliable across these responses? When do changes in those conclusions affect component choices, and how can uncertain choices be checked efficiently?

In this work, we distinguish a component's contribution to the selected computation from the model's dependence on its availability through suppression and exclusion, respectively. Under a fixed score order, exclusion preserves the surviving selected components and admits eligible alternatives. In language MoEs, the two interventions produce nearly identical rankings by representation change but can disagree on effect signs and component choice. Controlled replay and complementary analyses examine how subsequent routing, response geometry, candidate scoring, and coefficient normalization affect these comparisons. We further investigate how effect aggregation influences temporal rank transfer and whether selection criteria that preserve loss also preserve predictions under joint exclusion. A sufficient condition for selection stability motivates using suppression ranking margins to prioritize additional measurements, which we evaluate through both agreement with the reference choice and selection regret.

We summarize our contributions as follows:

1. We formalize suppression and exclusion, quantify their attribution disagreements, and isolate the contribution of subsequent routing through controlled replay (Sections 3 and 4).
2. We show that effect aggregation changes temporal rank transfer and that lower loss deviation under joint expert exclusion need not imply fewer prediction changes (Sections 5.1 and 5.2).
3. We use separate inputs for threshold calibration and evaluation. At matched group counts, margin allocation reduces remaining disagreement relative to uniform random measurement, though not always with lower mean selection regret.

2 Related Work

2.1 Sparse Selection and Model Compression

Sparse expert networks separate routing from expert computation (Shazeer et al., 2017; Lepikhin et al., 2021; Fedus et al., 2022), with variants for language modeling (Qwen Team, 2024; Muennighoff et al., 2025; Dai et al., 2024) and forecasting (Shi et al., 2025). Vision architectures select or reorganize tokens (Rao et al., 2021; Liang et al., 2022), or merge their representations (Bolya et al., 2023). Gaussian scene representations (Kerbl et al., 2023) extend to language features (Qin et al., 2024), including sparse codebook representations (Li et al., 2025). Compression methods include expert pruning and dynamic skipping (Lu et al., 2024), learned pruning rates across layers (Bai et al., 2025), Shapley-value contributions (Huang et al., 2025), and activation criteria weighted by router probabilities (Lasby et al., 2026). Other approaches merge experts through hierarchical clustering (Chen et al., 2025) or identify collaborative expert groups for compression (Tang et al., 2026). LightGaussian combines pruning, spherical harmonic distillation, and vector quantization (Fan et al., 2024). We examine how available alternatives, coefficient construction, and subsequent computation affect the interpretation of component importance.

2.2 Interventions and Component Attribution

Removal-based explanations are characterized by how features are removed, which model behavior is explained, and how influence is summarized (Covert et al., 2021). Related work studies attribution for model editing (Shah et al., 2024), ablation baselines (Li & Janson, 2024), and the sensitivity of activation patching to the corruption procedure (Zhang & Nanda, 2024). Visual attribution includes occlusion analysis (Zeiler & Fergus, 2014) and relevance propagation (Chefer et al., 2021). Expert steering and safety-critical expert identification (Fayyaz et al., 2026; Lai et al., 2025), MoE interpretability and expert analyses (Yang et al., 2025; Herbst et al., 2026), and sparse autoencoder representations (Gao et al., 2025) provide related settings for studying internal components. Intervention studies also examine representation divergence (Grant et al., 2026), dormant pathways that mislead subspace patching (Makelov et al., 2024), and compensation by later components (Wang et al., 2023; Rushing & Nanda, 2024). In adaptive sparse models, an intervention must also specify

whether subsequent selections are fixed or recomputed. To isolate subsequent expert selection, we replay expert identities and order while recomputing continuous states and coefficients.

2.3 Importance and Decision Quality

Preserving perplexity need not preserve task performance (Jaiswal et al., 2024), and compressed models with similar accuracy can still change individual answers, more often where the margin between the top two answer probabilities is small (Dutta et al., 2024). We accordingly distinguish effect estimation, component choice, and the behavior of the selected modification, and we use ranking margins to decide where additional measurement is needed. Selection based on observed outcomes can invalidate naive inference using the same data (Berk et al., 2013; Lee et al., 2016). We use separate partitions to calibrate measurement thresholds and evaluate the resulting disagreement and regret.

3 Sparse Selection and Intervention Effects

Following Covert et al. (2021), we distinguish three elements of component attribution: the intervention, the model behavior under evaluation, and the summary of the resulting effects. We extend this decomposition to internal components of adaptive sparse models, where exclusion can alter the selected support by admitting previously unselected alternatives. We specify the intervention and scalar response before defining effect aggregation and component selection. All paired comparisons hold model parameters fixed and begin from a common incoming state.

Components and representations. Let $\mathcal{C} = \{1, \dots, M\}$ index the candidate components and $\mathcal{A} \subseteq \mathcal{C}$ denote the selected support. At a given incoming state, expert mixtures and sparse codebook representations admit the form $\boldsymbol{h} = \boldsymbol{V}\boldsymbol{a} + \boldsymbol{b}$, where $\boldsymbol{V} = [\boldsymbol{v}_1 \ \cdots \ \boldsymbol{v}_M]$ contains the component vectors, $\mathrm{supp}(\boldsymbol{a}) \subseteq \mathcal{A}$, and $\boldsymbol{b}$ collects contributions outside the mixture (Muennighoff et al., 2025; Li et al., 2025). The support determines which components participate, while the coefficients determine their contributions to the representation. For token selection, the selected vectors constitute the sequence processed by subsequent attention (Rao et al., 2021). Accordingly, token removal deletes the corresponding row from the sequence. Appendix A specifies the representation and selection mechanism for each architecture.

Counterfactual effects. Let $\mathcal{A}^0$ denote the original top-K support and $\prec^0$ the candidate score order, including a fixed rule for ties. For a target set $\mathcal{U} \subseteq \mathcal{C}$ satisfying $|\mathcal{C} \setminus \mathcal{U}| \geq K$, define

$$\mathcal{A}^{\mathrm{sup}} = \mathcal{A}^0 \setminus \mathcal{U}, \qquad \mathcal{A}^{\mathrm{exc}} = \mathrm{TopK}_{\prec^0}(\mathcal{C} \setminus \mathcal{U}). \tag{1}$$

Suppression removes targeted contributions from the original support, whereas exclusion removes targets from the eligible candidate set before selection. Under the original score order, exclusion preserves every surviving member of $\mathcal{A}^0$ and fills the vacated positions with eligible alternatives. When scores depend on candidate availability, we additionally consider rescored exclusion, denoted res. Recomputing scores after target removal can change the candidate order and displace originally selected survivors. Appendix A.1 formalizes these operators and establishes the conditions for survivor preservation.

Coefficients and recursive computation. For nonnegative native weights $p_{t,e}$, retained coefficients are $a_{t,e} = p_{t,e}\mathbf{1}\{e \in \mathcal{A}_t\}$. Under separate normalization, these coefficients are divided by the support mass $m_t(\mathcal{A}_t) = \sum_{j \in \mathcal{A}_t} p_{t,j} > 0$. Normalization therefore modifies surviving contributions even when their component identities remain unchanged. We retain the original MoE weights at the intervention site to separate reselection from renormalization. For sparse fields, we compare both conventions against their respective native baselines, restricting normalization to positive support mass (Appendix A.2).

For input i at site t, let $\boldsymbol{z}_{it}^q$ denote the incoming state, $\mathcal{C}_{it}^q$ the available candidates, and $\mathcal{U}_t$ the targets. Using selection operator $\mathcal{S}_t^q$ and coefficient convention ν, we propagate each branch as

$$\mathcal{A}_{it}^q = \mathcal{S}_t^q(\boldsymbol{z}_{it}^q, \mathcal{C}_{it}^q; \mathcal{U}_t), \qquad \boldsymbol{z}_{i,t+1}^q = F_t(\boldsymbol{z}_{it}^q, \mathcal{C}_{it}^q, \mathcal{A}_{it}^q; \nu). \tag{2}$$

Here F_t constructs the coefficients and updates the state. We recompute subsequent scores and selections within each branch unless a controlled continuation fixes the selections (Appendix A.3). Because branch states can diverge, the support inclusion established in Proposition 1 need not hold between branches at subsequent sites. For $q \in \{\mathrm{sup}, \mathrm{exc}, \mathrm{res}\}$, we define the signed effect $\delta_i^q = \ell_i^q - \ell_i^0$, where ℓ_i^q and ℓ_i^0 are the final scalar responses under intervention q and native computation, respectively.

4 Computational Structure and Component Attribution

We investigate four structural sources of attribution disagreement under suppression and exclusion. For language MoEs (Qwen Team, 2024; Muennighoff et al., 2025), we retain the original candidate order and router probabilities at the intervention site and quantify the contribution of subsequent expert selection through controlled replay in OLMoE. For Time-MoE (Shi et al., 2025), we decompose forecast loss differences into residual alignment and squared displacement to establish the role of response geometry. In DynamicViT (Rao et al., 2021), we examine the dependence of selection scores on candidate availability through pooled token representations. For fitted LangSplatV2 fields (Li et al., 2025), we assess the sensitivity of component attribution to coefficient normalization in the absence of subsequent adaptive selection.

4.1 Expert Attribution and Routing Dependence

We evaluate every routed expert in every layer of Qwen1.5-MoE (Qwen Team, 2024) and OLMoE (Muennighoff et al., 2025) on the same 100 MMLU questions (Hendrycks et al., 2021). Each intervention targets one expert at all prompt positions in a single layer, and we measure its effect on the negative log likelihood of the correct answer. Figure 1a shows that suppression and exclusion rank experts similarly by local perturbation magnitude, with a median Spearman correlation of 0.997 in Qwen, but agree much less on signed loss effects, with a median of 0.290. Knowing which experts produce the largest representation changes therefore does not establish which experts contribute most to the measured response.

Table 1 shows how this distinction affects attribution. Removing the same expert can increase loss under one intervention and decrease it under the other, as occurs in 27.2% of eligible Qwen conditions and 23.8% of OLMoE conditions. The interventions select different experts with the largest loss increase in 56.8% and 36.4% of groups, respectively. Allowing effects within 10^{-3} nats of each maximum still leaves 52.3% and 33.8% of groups without a common maximizing expert. The disagreement thus persists when we treat nearly equal effects as equivalent, supporting a dependence of expert importance on the removal rule (Appendices B and B.1).

Table 1: **Expert attribution under suppression and exclusion.** Rates (%) of opposite loss-effect signs, different maximizing experts, and disjoint τ-optimal sets. Sign disagreement is evaluated among expert conditions where both effect magnitudes exceed 10^{-3} nats. Selection comparisons include all question–layer groups. Each τ-optimal set contains experts whose signed effects are within τ of the maximum under the corresponding intervention.

Model	Sign disagreement (both $\lvert\delta\rvert > 10^{-3}$ nats)	Top-1 disagreement (all groups)	Disjoint τ-optimal sets ($\tau = 10^{-3}$ nats)
Qwen1.5-MoE	27.2	56.8	52.3
OLMoE	23.8	36.4	33.8

To identify the contribution of subsequent routing, we treat later expert assignments as a mediator (Pearl, 2001; VanderWeele, 2013). We impose the expert identities and order recorded under suppression on the exclusion branch, while recomputing continuous states and coefficients. Denoting the resulting loss by $\ell^{\mathrm{exc,fix}}$ gives

$$\ell^{\mathrm{exc}} - \ell^{\mathrm{sup}} = \underbrace{\ell^{\mathrm{exc,fix}} - \ell^{\mathrm{sup}}}_{\text{fixed assignments}} + \underbrace{\ell^{\mathrm{exc}} - \ell^{\mathrm{exc,fix}}}_{\text{subsequent routing}}. \tag{3}$$

Figure 1b,c shows that recomputing subsequent assignments reverses the loss ordering between exclusion and suppression in 44 of 128 conditions at layer 0 and 17 of 128 at layer 6. Because the target exclusion remains unchanged, these reversals show that later expert selections can determine which intervention produces the larger loss. At the final expert layer, the two contrasts coincide in all 128 conditions, as expected when no subsequent routing remains (Appendix B.2).

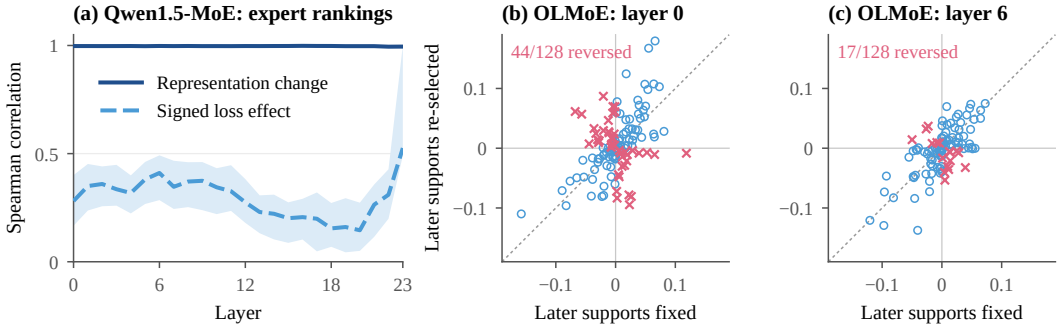

Figure 1: **Expert rankings and routing mediation.** (a) Agreement on rankings by local perturbation magnitude does not extend to signed loss effects. Lines and bands show medians and interquartile ranges. (b,c) Recomputing subsequent expert assignments can reverse the loss ordering between exclusion and suppression. Red crosses mark opposite contrast signs with both magnitudes above 10^{-3} nats.

4.2 Response Geometry

We evaluate Time-MoE to determine how the direction of a prediction change affects its loss. For the suppression residual $\boldsymbol{r}^{\text{sup}} = \hat{\boldsymbol{y}}^{\text{sup}} - \boldsymbol{y}$ and forecast difference $\boldsymbol{d} = \hat{\boldsymbol{y}}^{\text{exc}} - \hat{\boldsymbol{y}}^{\text{sup}}$ over d_y coordinates, squared error gives

$$\ell^{\text{exc}} - \ell^{\text{sup}} = \underbrace{\frac{2}{d_y}(\boldsymbol{r}^{\text{sup}})^\top \boldsymbol{d}}_{\text{residual alignment}} + \underbrace{\frac{1}{d_y}\|\boldsymbol{d}\|_2^2}_{\text{squared displacement}} . \tag{4}$$

Table 2 shows that the absolute alignment term exceeds squared displacement in at least 99.7% of conditions with $|\ell^{\text{exc}} - \ell^{\text{sup}}| > 10^{-4}$ MSE, thereby determining the loss ordering in Equation 4. Suppression and exclusion also identify disjoint maximizing sets in 4.7–21.4% of groups. These results show that displacement magnitude alone cannot determine loss effects or the resulting expert choice (Appendices A.4 and C).

Table 2: **Response direction and expert selection in Time-MoE.** Suppression and exclusion across 64 test windows and three layers per series. Set disagreement denotes disjoint extremizing sets with a 10^{-4} MSE tolerance. Opposite-sign rates use all expert conditions and require both effect magnitudes to exceed 10^{-4} MSE. Native MSE uses context normalization; other entries are percentages. Blue highlights alignment dominance.

Metric	ETTh1	ETTm1	Weather	ECL	Traffic
Native MSE	0.7699	0.7039	0.5427	0.1143	0.0203
Maximizer set disagreement	15.6	20.8	21.4	19.8	4.7
Minimizer set disagreement	13.5	16.7	17.2	16.1	5.7
Opposite effect signs	9.9	8.7	8.5	4.5	1.4
Alignment dominance	99.9	100.0	100.0	99.9	99.7

4.3 Candidate Dependence of Selection Scores

We evaluate DynamicViT on 3,893 Imagenette images. Figure 2a compares rankings by signed loss effects δ and effect magnitudes $|\delta|$, rather than the representation changes examined in Section 4.1.

At keep ratio 0.7, signed rank correlations reach 0.834–0.998, yet opposite signs occur in 12.31–20.74% of selected-target conditions (Figure 2b). Agreement in relative ordering therefore does not establish agreement on whether removal increases or decreases loss. Excluding an originally unselected token and recomputing scores also changes the selected set in 4.84–6.54% of conditions (Figure 2c). A token can thus influence selection through candidate scoring despite having zero suppression effect. Appendix D provides the protocol and controls.

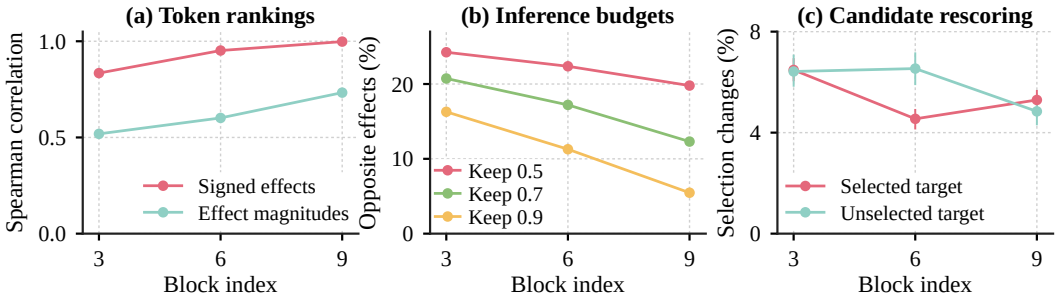

Figure 2: **Token attribution and candidate scoring in DynamicViT.** (a) Rank correlations between suppression and fixed-score exclusion for signed effects δ and magnitudes $|\delta|$. (b) Opposite-sign rates across inference budgets. Both effect magnitudes must exceed 10^{-4} nats, and the denominator includes all selected-target conditions. (c) Changes in selected sets between fixed-score and rescored exclusion. Panels (a,c) use keep ratio 0.7.

4.4 Coefficient Normalization

We evaluate fitted LangSplatV2 fields to examine coefficient normalization with geometry and compositing fixed and no subsequent adaptive selection. For a common surviving support and separate normalization in each branch, let $m_{\mathrm{S}} > 0$ and $m_{\mathrm{N}} > 0$ denote surviving and replacement weight masses, and let $\bar{\boldsymbol{v}}_{\mathrm{N}}$ denote the weighted mean replacement vector. Then

$$\boldsymbol{h}^{\mathrm{exc}} - \boldsymbol{b} = (1-\lambda)(\boldsymbol{h}^{\mathrm{sup}} - \boldsymbol{b}) + \lambda \bar{\boldsymbol{v}}_{\mathrm{N}}, \qquad \lambda = \frac{m_{\mathrm{N}}}{m_{\mathrm{S}} + m_{\mathrm{N}}}. \tag{5}$$

Replacement mass therefore rescales every surviving contribution. Across all evaluated views, attribution disagreements persist without subsequent adaptive selection and increase under normalized coefficients (Table 3). Set disagreement denotes disjoint sets of codewords whose effects lie within 10^{-3} relevance of each intervention's maximum. Each convention uses its own native baseline, so this comparison measures sensitivity to the coefficient convention (Appendices A.2 and E).

Table 3: **Component attribution in fitted sparse fields.** We compare suppression and exclusion under retained and normalized coefficients, evaluating 64 codewords individually in each of 915 3D-OVS and 279 LERF groups. IoU difference reports the percentage of conditions with an absolute IoU difference above 0.01. Purple and blue highlight opposite signs and maximizer disagreement, respectively, under normalized coefficients.

Benchmark	Coefficients	Opposite signs (%)	Maximizer disagreement (%)	Set disagreement (%)	IoU difference (%)
3D-OVS	Retained	0.79	8.52	1.31	3.21
	Normalized	2.45	14.43	2.95	5.05
LERF	Retained	3.00	3.94	0.72	5.20
	Normalized	3.92	5.02	1.08	6.77

5 From Importance Estimates to Component Selection

We evaluate how effect aggregation influences the transfer of expert rankings and how attribution criteria determine joint modifications under a common exclusion operator.

5.1 Effect Aggregation and Temporal Transfer

For a candidate modification $\mathcal{U}$, we define

$$\widehat{J}_q^{\text{signed}}(\mathcal{U}) = \frac{1}{n}\sum_{i=1}^{n}\delta_i^q(\mathcal{U}), \qquad \widehat{J}_q^{\text{abs}}(\mathcal{U}) = \frac{1}{n}\sum_{i=1}^{n}|\delta_i^q(\mathcal{U})|. \tag{6}$$

For $r \in \{\text{signed}, \text{abs}\}$, we bound the difference between interventions using the triangle and reverse triangle inequalities:

$$\left|\widehat{J}_{\text{exc}}^r(\mathcal{U}) - \widehat{J}_{\text{sup}}^r(\mathcal{U})\right| \le \frac{1}{n}\sum_{i=1}^{n}\left|\delta_i^{\text{exc}}(\mathcal{U}) - \delta_i^{\text{sup}}(\mathcal{U})\right|. \tag{7}$$

If the right-hand side is at most ϵ uniformly over a common candidate family, a unique suppression optimum with a gap greater than 2ϵ remains optimal under exclusion. The guarantee applies to maximizing mean signed effects or minimizing mean absolute effects on the same inputs. It establishes neither temporal transfer nor joint optimality of sets assembled from singleton rankings (Appendix A.5).

Maximizing the signed objective identifies the largest mean loss increase, whereas minimizing the absolute objective limits deviation from the native loss without cancellation across inputs. We assess temporal transfer separately by comparing Time-MoE expert rankings under exclusion on 32 calibration windows and 64 later test windows per series. Across five series and three layers, rankings by mean absolute effect achieve a median temporal correlation of 0.929, compared with 0.024 for mean signed effect (Figure 3). Experts therefore retain a more consistent ordering by the magnitude of their removal effects than by their average signed contribution to loss. Stable rankings of effect magnitude alone cannot justify transferring expert choices based on mean loss increase (Appendix C).

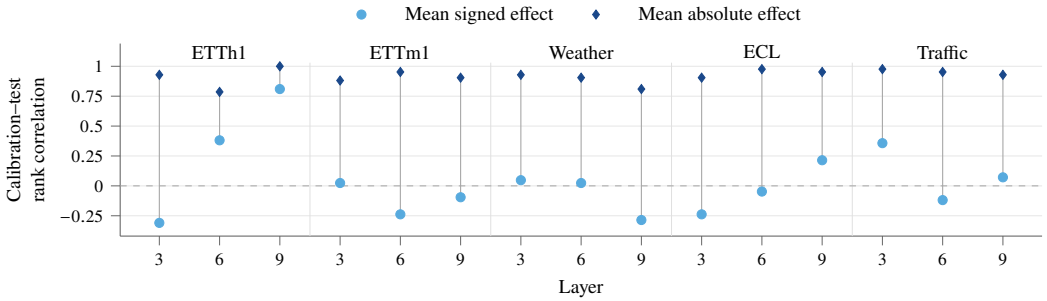

Figure 3: **Temporal transfer of expert rankings.** We compare Time-MoE expert rankings under exclusion on 32 calibration windows and 64 later test windows per series. Circles and diamonds show rank correlations for mean signed and mean absolute effects, respectively, with each segment pairing the two summaries for the same series and layer.

5.2 Joint Expert Exclusion

We retrospectively compare six criteria using 50 calibration questions and 1,000 disjoint, previously examined evaluation questions per model. The effect criteria select the k experts per layer with the smallest mean absolute singleton effects under suppression or exclusion. We compare these choices with REAP (Lasby et al., 2026), REAP with normalized calibration weights, routing frequency, and expert output norm, then freeze all selected sets before evaluation under common exclusion. Calibration costs differ across criteria (Appendix F). Suppression calibration yields lower loss deviation than REAP in all four model and budget settings under the adjusted paired comparisons, without establishing an accuracy advantage (Appendix F). Table 4 also shows that the criterion with the lowest loss deviation need not minimize prediction disagreement. Preservation of the measured loss therefore does not establish preservation of other predictive properties. The supplementary criticality comparison also favors different calibration criteria in the two models (Appendix F).

Table 4: **Expert selection under common exclusion.** We compare how six selection criteria preserve model behavior when $k = 8$ or 16 experts per layer are jointly excluded in OLMoE and Qwen1.5-MoE. Loss deviation and KL divergence are reported in nats; accuracy, prediction disagreement (PD), and correctness flips are percentages. Purple and blue highlight the lowest loss deviation and KL divergence, respectively, for each model and value of k.

OLMoE	$k = 8$					$k = 16$				
Criterion	Loss dev.↓	Acc.↑	KL↓	PD↓	Flips↓	Loss dev.↓	Acc.↑	KL↓	PD↓	Flips↓
Suppression	0.1163	54.2	0.0173	9.8	6.5	0.2288	50.0	0.0665	18.0	11.7
Exclusion	0.1343	52.6	0.0224	11.3	7.5	0.2337	51.5	0.0675	19.1	13.2
REAP	0.1431	53.1	0.0213	12.3	7.2	0.2626	51.8	0.0662	21.8	14.7
Routing frequency	0.1266	52.7	0.0241	10.3	6.4	0.2773	50.4	0.0838	19.7	13.5
Output norm	0.2069	51.9	0.0450	19.2	12.2	0.3614	50.2	0.1169	28.9	20.9
REAP (normalized)	0.1479	52.9	0.0241	11.6	7.2	0.3932	48.0	0.1123	31.1	20.5
Qwen1.5-MoE	$k = 8$					$k = 16$				
Criterion	Loss dev.↓	Acc.↑	KL↓	PD↓	Flips↓	Loss dev.↓	Acc.↑	KL↓	PD↓	Flips↓
Suppression	0.1394	63.9	0.0264	9.2	7.7	0.2464	61.9	0.0821	14.6	11.1
Exclusion	0.1572	65.3	0.0318	10.0	6.9	0.2453	62.8	0.0758	15.1	11.6
REAP	0.1696	64.7	0.0396	8.8	6.5	0.2974	62.9	0.0986	15.7	11.9
Routing frequency	0.2007	63.7	0.0465	12.1	9.3	0.2721	60.7	0.0891	20.0	14.7
Output norm	0.1933	64.2	0.0481	9.1	6.8	0.3428	61.0	0.1218	18.3	13.6
REAP (normalized)	0.1853	64.1	0.0511	10.3	8.5	0.3335	62.7	0.1079	16.8	12.1

6 Selective Measurement for Component Choice

We define the reference choice as the component with the largest signed effect under intervention q. We use exclusion for the language MoEs and sparse fields, and rescored exclusion for DynamicViT. For each group g, defined by an input and a selection site, let $\mathcal{C}_g$ denote the evaluated candidates and let $e_q(g)$ maximize $\delta^q_{g,e}$ over $\mathcal{C}_g$, with a fixed rule for ties. We define the suppression margin as

$$m_g = \delta^{\sup}_{g,e_{\sup}(g)} - \max_{e \in \mathcal{C}_g \setminus \{e_{\sup}(g)\}} \delta^{\sup}_{g,e}. \tag{8}$$

The largest discrepancy between candidate effects gives a sufficient condition for agreement:

$$\epsilon_g = \max_{e \in \mathcal{C}_g} \left| \delta^q_{g,e} - \delta^{\sup}_{g,e} \right|, \qquad m_g > 2\epsilon_g \implies e_{\sup}(g) = e_q(g). \tag{9}$$

This condition preserves the maximizing component because no pairwise effect difference can change by more than $2\epsilon_g$ (Appendix A.5). We cannot certify an unmeasured choice without a bound on ϵ_g, but small suppression margins identify frequent disagreements in the language models (Figure 4a). We therefore use these margins to prioritize additional measurements under the reference intervention. For each system, we calibrate τ to recover 90% of calibration disagreements and measure the reference effects for all candidates in a test group when $m_g \le \tau$:

$$\widehat{e}(g) = \begin{cases} e_q(g), & m_g \le \tau, \\ e_{\sup}(g), & m_g > \tau. \end{cases} \tag{10}$$

We evaluate the rule on test groups separate from threshold calibration and compare it with the expected outcome of uniform random measurement at the same group count. The budget counts additional group measurements and excludes initial suppression measurements and calibration. We report remaining disagreement with the reference choice and mean selection regret, the average gap between the largest reference effect and the reference effect of the chosen component.

For each language model, we average results over 500 random partitions of the same 100 questions, with 50 questions for calibration and 50 for evaluation in each partition. For sparse fields, we use the planned partition of frames within the recorded scenes. The DynamicViT analysis uses a separate retrospective sample of 128 calibration and 512 test images at keep ratio 0.9. Appendix G provides the protocols and uncertainty estimates.

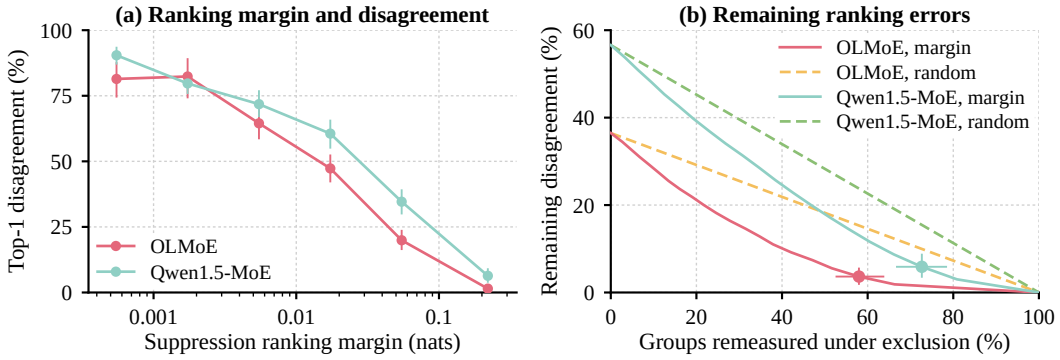

Figure 4: **Ranking margins and selective measurement.** (a) Disagreement between suppression and exclusion in the expert with the largest signed effect, grouped by suppression margin. (b) Remaining disagreement under margin allocation (solid) and uniform random allocation (dashed, expectation) at matched measurement counts. Markers indicate the 90% calibration recovery target. Error bars show question bootstrap variability in (a) and central 95% ranges across 500 question partitions in (b).

Table 5 shows that small suppression margins identify choices that change under the reference intervention, reducing disagreement relative to random allocation at the same measurement count. For language MoEs and sparse fields, lower mean regret also indicates that margin allocation resolves disagreements with larger combined effect gaps. In DynamicViT, however, it leaves fewer disagreements (1.6% versus 3.3%) but greater mean regret (0.1772 versus 0.1103, in units of 10^{-3} nats). Because calibration targets disagreement recovery without weighting effect gaps, improved agreement does not guarantee lower selection regret (Appendix G).

Table 5: **Selective measurement for component choice.** We compare two strategies at the same measurement count: (i) *Random*: uniform allocation across test groups, with results reported in expectation; (ii) *Margin*: allocation to groups whose suppression margins are at or below a calibrated threshold. Rates are percentages of all test groups, and regret is measured in nats (query relevance for LangSplatV2). Purple and blue highlight the lower remaining disagreement and mean regret, respectively, within each row.

	Initial	Measured	Remaining disagreement		Mean regret ($\times 10^{-3}$)	
System	disagr.	groups	Random	Margin	Random	Margin
OLMoE (Muennighoff et al., 2025)	36.5	58.0	15.3	3.6	5.6645	1.4702
Qwen1.5-MoE (Qwen Team, 2024)	56.6	72.7	15.5	5.9	8.2680	3.7354
DynamicViT (Rao et al., 2021)	13.0	74.7	3.3	1.6	0.1103	0.1772
LangSplatV2 (Li et al., 2025), retained	6.8	28.8	4.8	0.6	0.0646	0.0376
LangSplatV2 (Li et al., 2025), normalized	12.2	30.9	8.4	1.7	0.1199	0.0319

7 Conclusion

Suppression and exclusion can assign opposite effect signs and select different components, even when, as in language MoEs, they rank components almost identically by representation change. We identify how subsequent routing, response geometry, candidate scoring, and coefficient normalization contribute to these differences. Suppression margins guide additional measurements that recover more reference choices than random allocation at the same measurement count, but improved agreement need not reduce mean selection regret. Calibration objectives that weigh the effect gap of each remaining error are therefore a natural extension for verifying component attribution in adaptive sparse models.

AI USE STATEMENT

We used large language models to assist with literature search, editing and polishing the manuscript text, writing and revising code for experiments, analysis scripts, and figures generated from our recorded experimental outputs, drafting descriptions and interpretations of results, and checking mathematical derivations for errors. All AI-suggested content was reviewed and verified by the authors against the experimental records and the original publications. The authors designed and ran all experiments and take full responsibility for the content of this paper.

REPRODUCIBILITY STATEMENT

We specify the intervention operators and coefficient conventions in Section 3 and provide the assumptions and proofs for selection stability in Appendix A.5. Appendices B, C, D, and E document the model configurations, data preparation, intervention protocols, and evaluation metrics for the four model families. Appendices F and G describe the joint selection and selective measurement experiments, including calibration and evaluation partitions, measurement budgets, and uncertainty estimation.

REFERENCES

Sikai Bai, Haoxi Li, Jie Zhang, Zicong Hong, and Song Guo. DiEP: Adaptive mixture-of-experts compression through differentiable expert pruning. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, pp. 56090–56115. Curran Associates, Inc., 2025. doi: 10.52202/085713-1878. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/511c7fd69db9f1ce7492a57285975849-Paper-Conference.pdf.

Richard Berk, Lawrence Brown, Andreas Buja, Kai Zhang, and Linda Zhao. Valid post-selection inference. *The Annals of Statistics*, 41(2):802–837, 2013. doi: 10.1214/12-AOS1077.

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your ViT but faster. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=JroZRaRw7Eu.

Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 782–791, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Chefer_Transformer_Interpretability_Beyond_Attention_Visualization_CVPR_2021_paper.html.

I-Chun Chen, Hsu-Shen Liu, Wei-Fang Sun, Chen-Hao Chao, Yen-Chang Hsu, and Chun-Yi Lee. Retraining-free merging of sparse MoE via hierarchical clustering. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 8594–8620. PMLR, 2025. URL https://proceedings.mlr.press/v267/chen25aq.html.

Ian C. Covert, Scott Lundberg, and Su-In Lee. Explaining by removing: A unified framework for model explanation. *Journal of Machine Learning Research*, 22(209):1–90, 2021. ISSN 1532-4435. URL https://jmlr.org/papers/v22/20-1316.html.

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 Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1280–1297, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.70. URL https://aclanthology.org/2024.acl-long.70/.

Abhinav Dutta, Sanjeev Krishnan, Nipun Kwatra, and Ramachandran Ramjee. Accuracy is not all you need. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/e0e956681b04ac126679e8c7dd706b2e-Abstract-Conference.html.

Zhiwen Fan, Kevin Wang, Kairun Wen, Zehao Zhu, Dejia Xu, and Zhangyang Wang. LightGaussian: Unbounded 3D gaussian compression with 15x reduction and 200+ FPS. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.nips.cc/paper_files/paper/2024/hash/fd881d3b625437354d4421818f81058f-Abstract-Conference.html.

Mohsen Fayyaz, Seyed MohammadAli Modarressi, Hanieh Deilamsalehy, Franck Dernoncourt, Ryan Rossi, Trung Bui, Hinrich Schuetze, and Nanyun Peng. Steering MoE LLMs via expert (de)activation. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 7803–7826, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/0d61c5f5ef91e7e8a091b7b8f72b853c-Paper-Conference.pdf.

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

Leo Gao, Tom Dupre la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=tcsZt9ZNKD.

Satchel Grant, Simon Jerome Han, Alexa R. Tartaglini, and Christopher Potts. Addressing divergent representations from causal interventions on neural networks. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=cZrTMqYVL6.

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. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Jeremy Herbst, Stefan Wermter, and Jae Hee Lee. The expert strikes back: Interpreting mixture-of-experts language models at expert level. In *Mechanistic Interpretability Workshop at ICML 2026*, 2026. URL https://openreview.net/forum?id=ZCdnWNavOF.

Weizhong Huang, Yuxin Zhang, Xiawu Zheng, Fei Chao, Rongrong Ji, and Liujuan Cao. Discovering important experts for mixture-of-experts models pruning through a theoretical perspective. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, pp. 135973–136003. Curran Associates, Inc., 2025. doi: 10.52202/085713-4538. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/c66a9db149261435664284a20b6f1d42-Paper-Conference.pdf.

Ajay Jaiswal, Zhe Gan, Xianzhi Du, Bowen Zhang, Zhangyang Wang, and Yinfei Yang. Compressing LLMs: The truth is rarely pure and never simple. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=B9klVS7Ddk.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023. URL https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/.

Zhenglin Lai, Mengyao Liao, Bingzhe Wu, Dong Xu, Zebin Zhao, Zhihang Yuan, Chao Fan, and Jianqiang Li. SAFEx: Analyzing vulnerabilities of MoE-based LLMs via stable safety-critical expert identification. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, pp. 130380–130404. Curran Associates, Inc., 2025. doi: 10.52202/

085713-4344. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/bd127877149d4965ad834c75a65b3052-Paper-Conference.pdf.

Mike Lasby, Ivan Lazarevich, Nish Sinnadurai, Sean Lie, Yani Ioannou, and Vithursan Thangarasa. REAP the experts: Why pruning prevails for one-shot MoE compression. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=ukGxWd2aDG.

Jason D. Lee, Dennis L. Sun, Yuekai Sun, and Jonathan E. Taylor. Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927, 2016. ISSN 00905364. URL http://www.jstor.org/stable/43818915.

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.

Maximilian Li and Lucas Janson. Optimal ablation for interpretability. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=opt72TYzwZ.

Wanhua Li, Yujie Zhao, Minghan Qin, Yang Liu, Yuanhao Cai, Chuang Gan, and Hanspeter Pfister. Langsplatv2: High-dimensional 3d language gaussian splatting with 450+ FPS. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=XR5y4nvTfz.

Youwei Liang, Chongjian Ge, Zhan Tong, Yibing Song, Jue Wang, and Pengtao Xie. Not all patches are what you need: Expediting vision transformers via token reorganizations. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=BjyvwnXXVn_.

Xudong Lu, Qi Liu, Yuhui Xu, Aojun Zhou, Siyuan Huang, Bo Zhang, Junchi Yan, and Hongsheng Li. Not all experts are equal: Efficient expert pruning and skipping for mixture-of-experts large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6159–6172, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.334. URL https://aclanthology.org/2024.acl-long.334/.

Aleksandar Makelov, Georg Lange, Atticus Geiger, and Neel Nanda. Is this the subspace you are looking for? an interpretability illusion for subspace activation patching. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Ebt7JgMHv1.

Niklas Muennighoff, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Jacob Morrison, Sewon Min, Weijia Shi, Evan Pete Walsh, Oyvind Tafjord, Nathan Lambert, Yuling Gu, Shane Arora, Akshita Bhagia, Dustin Schwenk, David Wadden, Alexander Wettig, Binyuan Hui, Tim Dettmers, Douwe Kiela, Ali Farhadi, Noah A. Smith, Pang Wei Koh, Amanpreet Singh, and Hannaneh Hajishirzi. OLMoE: Open mixture-of-experts language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=xXTkbTBmqq.

Judea Pearl. Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, UAI'01, pp. 411–420, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. ISBN 1558608001.

Minghan Qin, Wanhua Li, Jiawei Zhou, Haoqian Wang, and Hanspeter Pfister. LangSplat: 3D language gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20051–20060, 2024.

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

Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=kR95DuwwXHZ.

Cody Rushing and Neel Nanda. Explorations of self-repair in language models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 42836–42855. PMLR, 2024. URL https://proceedings.mlr.press/v235/rushing24a.html.

Harshay Shah, Andrew Ilyas, and Aleksander Madry. Decomposing and editing predictions by modeling model computation. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 44244–44292. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/shah24a.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.

Xiaoming Shi, Shiyu Wang, Yuqi Nie, Dianqi Li, Zhou Ye, Qingsong Wen, and Ming Jin. Time-MoE: Billion-scale time series foundation models with mixture of experts. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=e1wDDFmlVu.

Yuanbo Tang, Naifan Zhang, Yan Tang, Meixuan Chen, Shuhan Huang, Tingyu Cao, and Yang Li. Specialization through collaboration: Understanding expert interaction in mixture-of-expert large language models. In Vera Demberg, Kentaro Inui, and Lluís Marquez (eds.), *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2326–2339, Rabat, Morocco, March 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.eacl-long.104. URL https://aclanthology.org/2026.eacl-long.104/.

Tyler J. VanderWeele. A three-way decomposition of a total effect into direct, indirect, and interactive effects. *Epidemiology*, 24(2):224–232, 2013. doi: 10.1097/EDE.0b013e318281a64e.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=NpsVSN6o4ul.

Xingyi Yang, Constantin Venhoff, Ashkan Khakzar, Christian Schroeder De Witt, Puneet K. Dokania, Adel Bibi, and Philip Torr. Mixture of experts made intrinsically interpretable. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 71231–71248, 2025. URL https://proceedings.mlr.press/v267/yang25ag.html.

Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, 2014. URL https://cs.nyu.edu/~fergus/papers/zeilerECCV2014.pdf.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Hf17y6u9BC.

APPENDIX

A MATHEMATICAL FOUNDATIONS

We use calligraphic symbols for sets, bold lowercase symbols for vectors, and bold uppercase symbols for matrices. Local identities compare branches at the same incoming state. The recursion below specifies their subsequent execution.

A.1 SELECTION AND COEFFICIENTS

For candidates $\mathcal{C} = \{1, \dots, M\}$, incoming state $\boldsymbol{z}$, and availability $\mathcal{B} \subseteq \mathcal{C}$ with $|\mathcal{B}| \geq K$, define

$$\mathcal{A}(\boldsymbol{z}, \mathcal{B}) \in \operatorname*{arg\,max}_{\substack{\mathcal{S} \subseteq \mathcal{B} \\ |\mathcal{S}| = K}} \sum_{e \in \mathcal{S}} s_e(\boldsymbol{z}, \mathcal{B}). \tag{A.1}$$

A fixed candidate-ID order resolves ties. Exchanging a selected candidate with a higher-scoring eligible candidate proves that this rule selects the first K candidates in the resulting total order $\prec^0$. For $|\mathcal{C} \setminus \mathcal{U}| \geq K$, suppression uses $\mathcal{A}^{\mathrm{sup}} = \mathcal{A}^0 \setminus \mathcal{U}$ and fixed-score exclusion uses $\mathcal{A}^{\mathrm{exc}} = \mathrm{TopK}_{\prec^0}(\mathcal{C} \setminus \mathcal{U})$. Rescored exclusion recomputes $s_e(\boldsymbol{z}, \mathcal{C} \setminus \mathcal{U})$.

Proposition 1 (Support preservation)**.** *Under a fixed order,* $\mathcal{A}^{\mathrm{sup}} \subseteq \mathcal{A}^{\mathrm{exc}}$ *and* $|\mathcal{A}^{\mathrm{exc}} \setminus \mathcal{A}^{\mathrm{sup}}| = |\mathcal{A}^0 \cap \mathcal{U}|$.

Proof. Every surviving native member has at most $K - 1$ predecessors. Exclusion cannot increase its eligible predecessors, so it remains selected. The second claim follows by subtracting the support sizes. □

An inactive target therefore leaves both supports unchanged, but output invariance also requires unchanged coefficients, component values, execution order, and continuation. Rescoring need not preserve the support. Suppression of an active target has fewer than K entries, although redundant vectors can produce the same representation from different supports.

Expert mixtures and fitted codebooks use $\boldsymbol{h}^q = \boldsymbol{V}\boldsymbol{a}^q + \boldsymbol{b}$. For fixed nonnegative weights p_e and $m(\mathcal{A}) = \sum_{e \in \mathcal{A}} p_e$, retained coefficients are $a_e = p_e \mathbf{1}\{e \in \mathcal{A}\}$; separate normalization instead uses $a_e = p_e \mathbf{1}\{e \in \mathcal{A}\}/m(\mathcal{A})$, which requires $m(\mathcal{A}) > 0$. For $\mathcal{N} = \mathcal{A}^{\mathrm{exc}} \setminus \mathcal{A}^{\mathrm{sup}}$, subtraction gives

$$\boldsymbol{h}^{\mathrm{exc}} - \boldsymbol{h}^{\mathrm{sup}} = \sum_{e \in \mathcal{N}} a_e^{\mathrm{exc}} \boldsymbol{v}_e + \sum_{e \in \mathcal{A}^{\mathrm{sup}}} (a_e^{\mathrm{exc}} - a_e^{\mathrm{sup}}) \boldsymbol{v}_e. \tag{A.2}$$

Retained weights eliminate the second sum. If rescoring removes original survivors, the same subtraction includes their negative contributions and restricts the coefficient-difference sum to the shared support.

A.2 NORMALIZATION GEOMETRY

Set $m_{\mathrm{S}} = m(\mathcal{A}^{\mathrm{sup}}) > 0$, $m_{\mathrm{N}} = m(\mathcal{N}) > 0$, $\bar{\boldsymbol{v}}_{\mathrm{N}} = \sum_{e \in \mathcal{N}} p_e \boldsymbol{v}_e / m_{\mathrm{N}}$, and $\lambda = m_{\mathrm{N}}/(m_{\mathrm{S}} + m_{\mathrm{N}})$. Splitting the normalized numerator gives

$$\boldsymbol{h}^{\mathrm{exc}} - \boldsymbol{b} = (1 - \lambda)(\boldsymbol{h}^{\mathrm{sup}} - \boldsymbol{b}) + \lambda \bar{\boldsymbol{v}}_{\mathrm{N}}, \qquad \|\boldsymbol{h}^{\mathrm{exc}} - \boldsymbol{h}^{\mathrm{sup}}\| \leq \lambda D_V, \tag{A.3}$$

where $D_V = \max_{e,f \in \mathcal{A}^{\mathrm{exc}}} \|\boldsymbol{v}_e - \boldsymbol{v}_f\|$. To prove the bound, express the difference between the two weighted averages as $\sum_{e,f} \alpha_e \beta_f (\boldsymbol{v}_e - \boldsymbol{v}_f)$ and apply the triangle inequality. If $m_{\mathrm{N}} = 0 < m_{\mathrm{S}}$, the mixtures coincide. If $m_{\mathrm{S}} = 0$, normalized suppression is undefined; our normalized measurements exclude this case. Defining an empty mixture as zero would extend the operator, but we do not use that extension in the experiments.

Physical router deletion requires the same coefficient convention. With unchanged surviving logits, a softmax over $\mathcal{B}$ gives $p'_e = p_e / \sum_{j \in \mathcal{B}} p_j$. Subsequent support normalization cancels this denominator because $p'_e / \sum_{j \in \mathcal{A}} p'_j = p_e / \sum_{j \in \mathcal{A}} p_j$. Retaining full-bank weights does not.

A.3 Recursive Computation

The structural equations are $\boldsymbol{s} = f_s(\boldsymbol{z}, \mathcal{B})$, $\boldsymbol{p} = f_p(\boldsymbol{z})$, $\mathcal{A} = f_{\mathrm{sel}}(\boldsymbol{s}, \mathcal{B})$, $\boldsymbol{a} = f_a(\boldsymbol{p}, \mathcal{A})$, and $\boldsymbol{h} = f_V(\boldsymbol{z})\boldsymbol{a} + f_b(\boldsymbol{z})$. Thus weights affect the representation even when membership remains fixed. These equations describe model execution, not the process generating its inputs. For branch q, let $\mathcal{C}_t^q$ be the candidates reaching site t, $\mathcal{U}_t$ the targeted identities, and $\widetilde{\boldsymbol{s}}_t^q = f_{s,t}(\boldsymbol{z}_t^q, \mathcal{C}_t^q)$. Then

$$\begin{aligned} \mathcal{A}_t^{\mathrm{sup}} &= \mathrm{TopK}(\widetilde{\boldsymbol{s}}_t^{\mathrm{sup}}, \mathcal{C}_t^{\mathrm{sup}}) \setminus \mathcal{U}_t, \\ \mathcal{A}_t^{\mathrm{exc}} &= \mathrm{TopK}(\widetilde{\boldsymbol{s}}_t^{\mathrm{exc}}, \mathcal{C}_t^{\mathrm{exc}} \setminus \mathcal{U}_t), \\ \mathcal{A}_t^{\mathrm{res}} &= \mathrm{TopK}\big(f_{s,t}(\boldsymbol{z}_t^{\mathrm{res}}, \mathcal{C}_t^{\mathrm{res}} \setminus \mathcal{U}_t), \mathcal{C}_t^{\mathrm{res}} \setminus \mathcal{U}_t\big), \\ \boldsymbol{z}_{t+1}^q &= F_t(\boldsymbol{z}_t^q, \mathcal{A}_t^q; \nu). \end{aligned} \tag{A.4}$$

Here F_t includes coefficient convention ν and subsequent computation. Each exclusion site must retain at least K_t eligible candidates. Fixed-score exclusion freezes scores against availability changes at the current state, not against changes in earlier states. Local support preservation therefore does not imply inclusion across later branch supports.

A.4 Response Geometry

For $\boldsymbol{d}_h = \boldsymbol{h}^{\mathrm{exc}} - \boldsymbol{h}^{\mathrm{sup}}$, affine loss gives $\ell(\boldsymbol{h}^{\mathrm{exc}}) - \ell(\boldsymbol{h}^{\mathrm{sup}}) = \boldsymbol{w}^\top \boldsymbol{d}_h$. Opposite vectors have equal norms but opposite affine loss changes. If the continuation has an L_∇-Lipschitz gradient along the segment, integrating its gradient yields

$$\ell(\boldsymbol{h}^{\mathrm{exc}}) - \ell(\boldsymbol{h}^{\mathrm{sup}}) = \nabla \ell(\boldsymbol{h}^{\mathrm{sup}})^\top \boldsymbol{d}_h + R, \qquad |R| \le \tfrac{1}{2} L_\nabla \|\boldsymbol{d}_h\|_2^2. \tag{A.5}$$

The remainder bound integrates $L_\nabla \alpha \|\boldsymbol{d}_h\|^2$ over $\alpha \in [0,1]$. Hard selection can violate this smoothness assumption. At the output, full-vocabulary cross entropy instead has the exact identity

$$\ell^{\mathrm{exc}} - \ell^{\mathrm{sup}} = (\boldsymbol{p}_{\mathrm{out}}^{\mathrm{sup}} - \boldsymbol{e}_y)^\top \boldsymbol{d}_o + D_{\mathrm{KL}}(\boldsymbol{p}_{\mathrm{out}}^{\mathrm{sup}} \| \boldsymbol{p}_{\mathrm{out}}^{\mathrm{exc}}), \quad \boldsymbol{d}_o = \boldsymbol{o}^{\mathrm{exc}} - \boldsymbol{o}^{\mathrm{sup}}, \tag{A.6}$$

where $\boldsymbol{p}_{\mathrm{out}}^q = \mathrm{softmax}(\boldsymbol{o}^q)$. Writing $F(\boldsymbol{o}) = \log \sum_j \exp(o_j)$, the KL term equals $F(\boldsymbol{o}^{\mathrm{exc}}) - F(\boldsymbol{o}^{\mathrm{sup}}) - (\boldsymbol{p}_{\mathrm{out}}^{\mathrm{sup}})^\top \boldsymbol{d}_o$, which proves the identity by substitution.

For d_y forecast coordinates, set $\boldsymbol{r} = \hat{\boldsymbol{y}}^{\mathrm{sup}} - \boldsymbol{y}$ and $\boldsymbol{d} = \hat{\boldsymbol{y}}^{\mathrm{exc}} - \hat{\boldsymbol{y}}^{\mathrm{sup}}$. Expanding the square gives

$$\ell^{\mathrm{exc}} - \ell^{\mathrm{sup}} = \big(2\boldsymbol{r}^\top \boldsymbol{d} + \|\boldsymbol{d}\|_2^2\big)/d_y. \tag{A.7}$$

This proves Equation 4 for nonlinear forecasts as well. The alignment term determines the sign whenever $\boldsymbol{z}_{t+1}^q = F_t(\boldsymbol{z}_t^q, \mathcal{C}_t^q, \mathcal{A}_t^q; \nu)$.

A.5 Decision Stability

On a fixed input distribution, let $J_q^{\mathrm{signed}}(\mathcal{U}) = \mathbb{E}[\delta_i^q(\mathcal{U})]$ and $J_q^{\mathrm{abs}}(\mathcal{U}) = \mathbb{E}[|\delta_i^q(\mathcal{U})|]$, with finite expectations. The triangle and reverse triangle inequalities give

$$\big|J_{\mathrm{exc}}^r(\mathcal{U}) - J_{\mathrm{sup}}^r(\mathcal{U})\big| \le \mathbb{E}[|\delta_i^{\mathrm{exc}}(\mathcal{U}) - \delta_i^{\mathrm{sup}}(\mathcal{U})|], \qquad r \in \{\mathrm{signed}, \mathrm{abs}\}. \tag{A.8}$$

The absolute case uses $\big||a| - |b|\big| \le |a - b|$ before expectation. The same argument applies to empirical means.

Proposition 2 (Objective stability)**.** *If $\sup_{\mathcal{U} \in \mathcal{F}} |J_{\mathrm{exc}}(\mathcal{U}) - J_{\mathrm{sup}}(\mathcal{U})| \le \epsilon$ on a common candidate family, optimizing suppression incurs at most 2ϵ exclusion regret. A unique suppression optimum with gap $m > 2\epsilon$ remains the unique exclusion optimum, for either minimization or maximization.*

Proof. For minimizers $\mathcal{U}_{\mathrm{sup}}$ and $\mathcal{U}_{\mathrm{exc}}$,

$$J_{\mathrm{exc}}(\mathcal{U}_{\mathrm{sup}}) \le J_{\mathrm{sup}}(\mathcal{U}_{\mathrm{sup}}) + \epsilon \le J_{\mathrm{sup}}(\mathcal{U}_{\mathrm{exc}}) + \epsilon \le J_{\mathrm{exc}}(\mathcal{U}_{\mathrm{exc}}) + 2\epsilon.$$

For maximizers, reversing the optimality comparison gives $J_{\mathrm{exc}}(\mathcal{U}_{\mathrm{exc}}) \le J_{\mathrm{sup}}(\mathcal{U}_{\mathrm{exc}}) + \epsilon \le J_{\mathrm{sup}}(\mathcal{U}_{\mathrm{sup}}) + \epsilon \le J_{\mathrm{exc}}(\mathcal{U}_{\mathrm{sup}}) + 2\epsilon$. For any competitor $\mathcal{U}$, a suppression maximizer $\mathcal{U}_\star$ with gap m satisfies $J_{\mathrm{exc}}(\mathcal{U}_\star) - J_{\mathrm{exc}}(\mathcal{U}) \ge m - 2\epsilon > 0$. For a minimizer, use $J_{\mathrm{exc}}(\mathcal{U}) - J_{\mathrm{exc}}(\mathcal{U}_\star)$ instead. □

This proves the margin condition for individual effects and for absolute response objectives. It does not bound sampling error or justify composing singleton choices into an optimal joint intervention.

Information required for exclusion. Suppression alone cannot bound the unknown exclusion discrepancy. For $K = 1$, weights $(0.6, 0.4)$, $v_1 = y = 1$, and squared loss, compare $v_2 = 0$ with $v_2 = 2.5$. Both models have native loss 0.16 and suppression effects $(0.84, 0)$, but exclusion effects are $(0.84, 0)$ and $(-0.16, 0)$, respectively. Their optimal exclusion choices differ despite identical native and suppression information.

B Language Expert Attribution

We evaluate the same 100 MMLU questions in Qwen1.5-MoE-A2.7B (24 layers, 60 experts, $K = 4$) and OLMoE-1B-7B-0125 (16 layers, 64 experts, $K = 8$) (Qwen Team, 2024; Muennighoff et al., 2025; Hendrycks et al., 2021). Each intervention targets one expert at all prompt positions in one layer, retains original router probabilities, and preserves Qwen's shared branch. Subsequent computation proceeds normally except during replay. For vocabulary $\mathcal{V}$, answer options $\mathcal{O}$, and correct token y_i,

$$\ell^q_{ite} = \log \sum_{v \in \mathcal{V}} \exp(o^q_{ite,v}) - o^q_{ite,y_i}, \quad \delta^q_{ite} = \ell^q_{ite} - \ell^0_i, \quad \widehat{y}^q_{ite} = \arg\max_{a \in \mathcal{O}} o^q_{ite,a}. \tag{B.1}$$

Loss uses the full vocabulary, whereas prediction compares four options. Thus identical option logits need not imply identical losses. Local norms concatenate expert-mixture outputs across prompt positions. An active expert participates at least once in the prompt; inactive experts remain in the bank with zero effects. We use average ranks for correlation ties, a fixed ID order for component choice, and questions as resampling units. The study measures one answer token, not autoregressive generation.

B.1 Attribution Agreement

Within each question and layer, define $\mathcal{M}^q_{it}(\tau) = \{e : \delta^q_{ite} \geq \max_j \delta^q_{itj} - \tau\}$. The main table compares maximizing IDs and disjoint sets at $\tau = 10^{-3}$. Their denominators are all 2,400 Qwen and 1,600 OLMoE groups. Sign disagreement instead conditions on both magnitudes exceeding 10^{-3}: 27,821/102,156 for Qwen and 17,194/72,353 for OLMoE. Active-target answer differences are 5,212/130,819 and 1,957/94,887. Restricting Qwen rank comparisons to active experts yields median correlations 0.996 for local norms and 0.256 for signed effects. Thus inactive zeros do not explain the difference. Figures B.1 and B.2 show tolerance sensitivity and layerwise effects. The latter includes final-layer sign reversals, which cannot require subsequent expert selection.

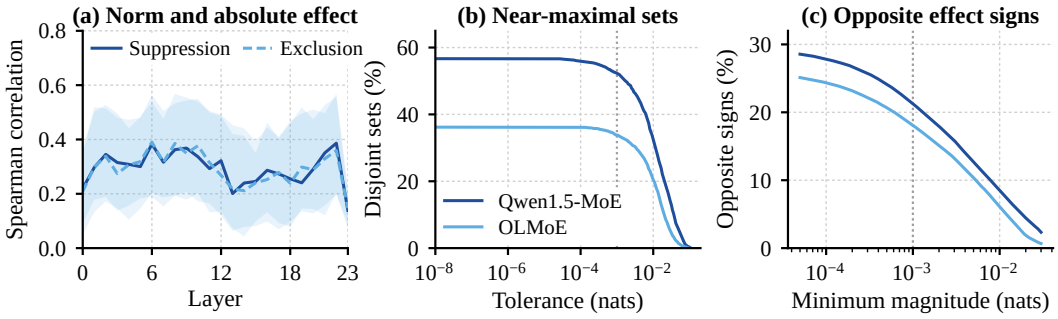

Figure B.1: **Sensitivity of attribution comparisons.** (a) Local norms correlate weakly with absolute loss effects in Qwen; lines and bands show medians and quartiles across questions. (b) Increasing the effect tolerance reduces disjoint maximizing sets. (c) Opposite signs persist above the indicated magnitude thresholds. Panel (c) uses all active conditions as its denominator, whereas Table 1 conditions on both magnitudes exceeding 10^{-3} nats. Vertical lines mark that threshold.

B.2 Routing Mediation and Prediction

The OLMoE control uses 32 previously examined questions and four targets per question at layers 0, 6, and 15. We record subsequent expert identities and order under suppression and replay

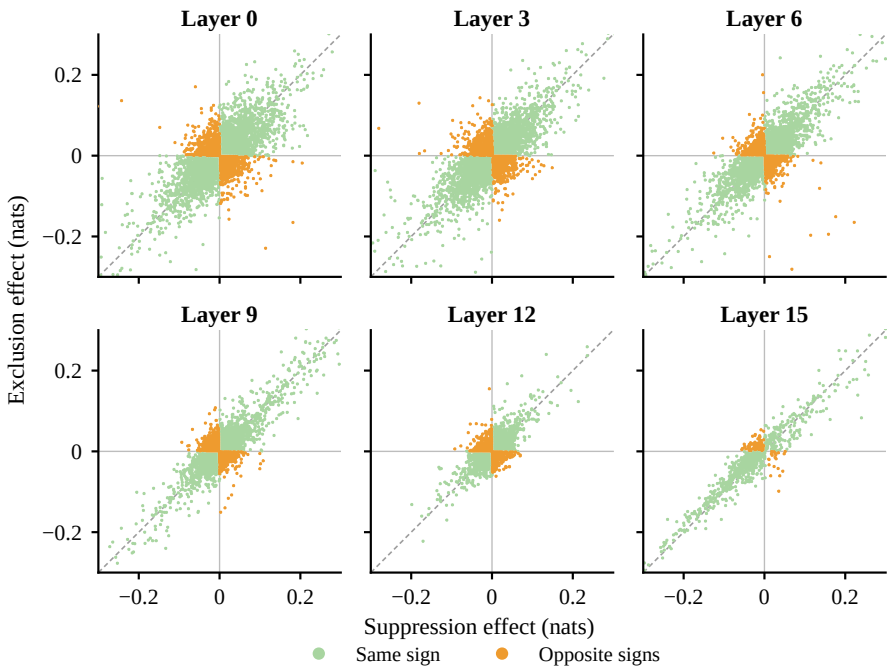

Figure B.2: **Intervention effects across OLMoE layers.** Suppression and exclusion can assign opposite signs even at the final expert layer. Axes show δ^{sup} and δ^{exc}; red marks require both magnitudes above 10^{-3} nats. Each panel contains 6,400 conditions before cropping to $[-0.3, 0.3]^2$. Counts and eligible denominators use the uncropped data.

them during exclusion, while recomputing hidden states, probabilities, attention, and expert outputs. Writing $L(q, \boldsymbol{m}; \boldsymbol{x})$ for the executable loss with assignments $\boldsymbol{m}$, we compare $L(\text{sup}, \boldsymbol{m}^{\text{sup}}; \boldsymbol{x})$, $L(\text{exc}, \boldsymbol{m}^{\text{sup}}; \boldsymbol{x})$, and $L(\text{exc}, \boldsymbol{m}^{\text{exc}}; \boldsymbol{x})$. Adding and subtracting the middle term gives Equation 3. The decomposition depends on the reference assignments (Pearl, 2001); its fixed-assignment term still includes continuous downstream responses.

Self-replay reproduces stored logits, inactive controls reproduce native outputs, and reconstructed losses agree to floating-point precision. Figure B.3 includes all 128 conditions per layer. Fixed and recomputed exclusion predict different answers in 12, 2, and 0 conditions, respectively. For question 2062, layer 0, expert 48, fixed exclusion retains the correct option C and recomputed exclusion selects D. Each winning margin is 0.0625, one bfloat16 spacing at the recorded logits. This case does not establish robustness to higher precision.

Effect estimation. An auxiliary OLMoE ridge predictor uses 32 fitting, 32 calibration, and 64 test questions at three layers, with fitting-only standardization and penalty 1. Suppression effects, router statistics, and local vector norms and alignments predict the exclusion correction. MAE falls from 0.02248 to 0.02198 nats, but maximizer errors rise from 81 to 84/192 groups. We therefore make no selection claim from this predictor and do not use it for margin allocation.

C Temporal Effects and Rank Transfer

Time-MoE-200M (Shi et al., 2025) uses the OT column of ETTh1, ETTm1, Weather, ECL, and Traffic. We normalize using each context's mean and standard deviation, supply 512 context points, and evaluate a direct 32-step forecast at the final context position. Interventions target each of eight routed experts at all context positions in layers 3, 6, and 9, retaining the shared branch and original probabilities ($K = 2$). Each series uses 32 fitting, 32 calibration, and 64 test windows. Table C.1 fixes their ranges. Test targets do not overlap, although contexts can overlap and temporal dependence remains. Saved forecasts reproduce all 15,360 effects and the squared-error identity to maximum residual 1.7×10^{-15}.

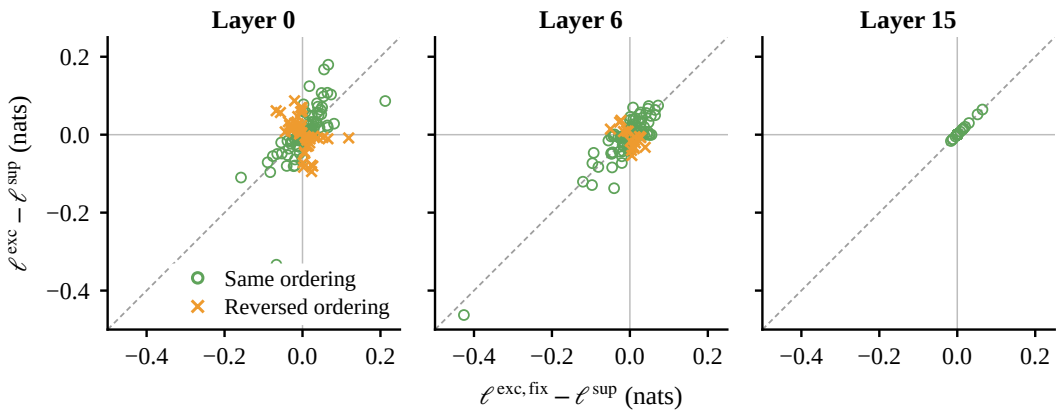

Figure B.3: **Complete routing contrasts in OLMoE.** The horizontal axis gives $\ell^{\text{exc,fix}} - \ell^{\text{sup}}$ and the vertical axis gives $\ell^{\text{exc}} - \ell^{\text{sup}}$. Red crosses mark opposite signs with both magnitudes above 10^{-3} nats. The final layer provides a control with no subsequent expert selection.

Table C.1: **Temporal partitions.** Half-open index ranges include both context and target values. The three partitions contain 32, 32, and 64 windows, respectively.

Series	Predictor fitting	Calibration	Test
ETTh1	$[4320, 8640)$	$[8640, 11520)$	$[11520, 14400)$
ETTm1	$[17280, 34560)$	$[34560, 46080)$	$[46080, 57600)$
Weather	$[26348, 36876)$	$[36887, 42135)$	$[42156, 52684)$
ECL	$[13152, 18400)$	$[18412, 21036)$	$[21043, 26291)$
Traffic	$[8772, 12260)$	$[12280, 14008)$	$[14035, 17523)$

In the series order above, disjoint maximizing sets occur in $(30, 40, 41, 38, 9)/192$ groups and disjoint minimizing sets in $(26, 32, 33, 31, 11)/192$, at tolerance 10^{-4} MSE. Opposite signs occur in $(152, 134, 131, 69, 22)/1536$ conditions, requiring both magnitudes above 10^{-4}. Among contrasts exceeding 10^{-4} in magnitude, alignment dominates in $1089/1090$, $1095/1095$, $1102/1102$, $783/784$, and $344/345$ cases.

We compare calibration and test rankings of eight experts in each of 15 series and layer groups. Mean absolute effects use $\mathbb{E}|\delta|$, not $|\mathbb{E}\delta|$. The main figure reports every paired correlation and the medians 0.024 for signed effects and 0.929 for absolute effects. A supplementary calibration split excludes overlapping context and target ranges, leaving 3–15 windows per half. Matching sample sizes and using 2,000 dependence-block bootstrap resamples yields intervals containing zero for all 15 differences between temporal transfer and calibration stability. The observations do not separate temporal shift from sampling variability. Frozen expert choices and auxiliary regressions likewise show no consistent test-MSE or selection advantage.

Figure C.1 fixes the first test window, layer 6, and expert 0 in every series. For Figure C.2, we divide the calibration-to-test mean change by the calibration standard deviation. Routing entropy and the second-to-third expert probability gap average context positions. Replacement norms flatten the local representation difference across positions and coordinates. Router features use one value per window; norms use all eight expert conditions.

D Candidate Dependence in Token Selection

DynamicViT uses scores $s_e(\mathcal{B}) = g(\boldsymbol{l}_e, \bar{\boldsymbol{g}}(\mathcal{B}))$, where $\bar{\boldsymbol{g}}(\mathcal{B}) = |\mathcal{B}|^{-1} \sum_{j \in \mathcal{B}} \boldsymbol{g}_j$ (Rao et al., 2021). Direct subtraction gives

$$\bar{\boldsymbol{g}}(\mathcal{B} \setminus \mathcal{U}) - \bar{\boldsymbol{g}}(\mathcal{B}) = \frac{|\mathcal{U}|}{|\mathcal{B}| - |\mathcal{U}|} [\bar{\boldsymbol{g}}(\mathcal{B}) - \bar{\boldsymbol{g}}(\mathcal{U})]. \tag{D.1}$$

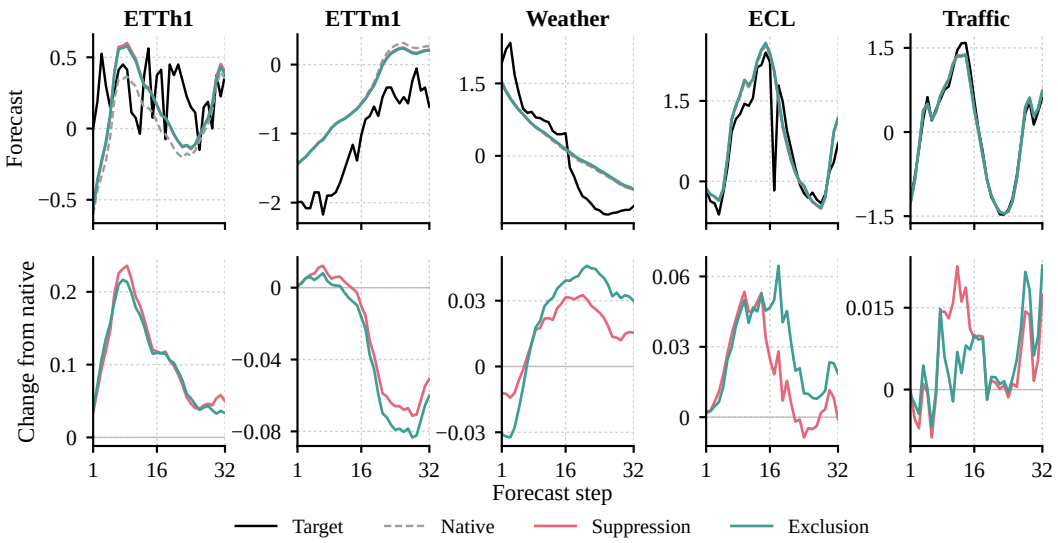

Figure C.1: **Forecasts under expert interventions.** Suppression and exclusion produce different forecast changes relative to the native prediction. The top row shows normalized targets and predictions; the bottom row shows the corresponding prediction changes. Each series uses its own labeled vertical scale.

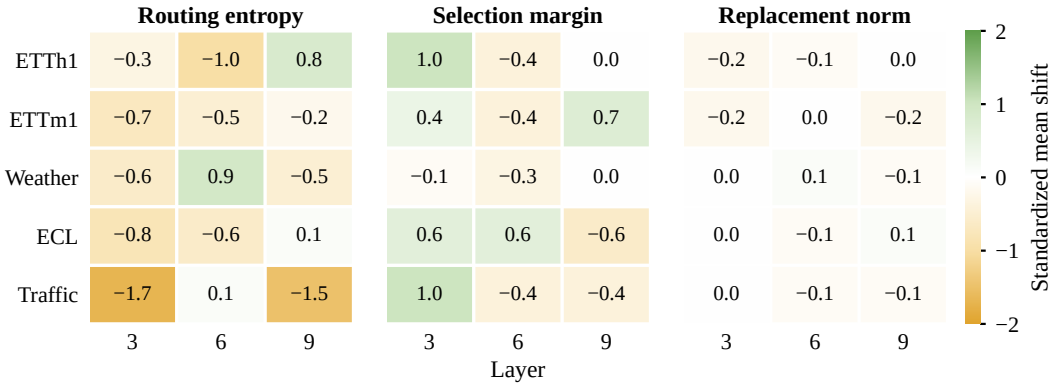

Figure C.2: **Changes in routing and local representations.** Standardized mean shifts describe changes between calibration and test periods. The measurements do not distinguish temporal distribution change from sampling variability.

If $g(\boldsymbol{l}_e, \cdot)$ is L_s-Lipschitz along this change, each surviving score changes by at most $\eta = L_s\|\bar{\boldsymbol{g}}(\mathcal{B} \setminus \mathcal{U}) - \bar{\boldsymbol{g}}(\mathcal{B})\|$. A gap greater than 2η between the Kth and $(K+1)$st eligible candidates preserves the support: every selected-minus-unselected score remains positive after perturbation. If exactly K candidates remain, availability fixes the support. Failure of the gap condition does not itself imply a support change.

We delete token rows and preserve the class token. Each intervention at block 3, 6, or 9 starts from the same native incoming state. The fixed Imagenette manifest reserves 32 of 3,925 images for engineering checks and evaluates 3,893 with the original ImageNet head. At each site, 16 selected and two unselected targets give 62,288 and 7,786 conditions. Sampling uses seeded priorities independent of labels and effects. Correlations compare targets within each image and then average images. Sign rates require both magnitudes above 10^{-4} nats but use all selected-target conditions as denominator. Bars use 2,000 image bootstrap resamples. Keep ratios 0.5, 0.7, and 0.9 change inference budgets of the same DeiT-S checkpoint trained at 0.7, not separate training runs.

At ratio 0.7, rescoring changes support for 500, 509, and 377 unselected targets across the three blocks, but changes their predicted class in only 1, 4, and 0 cases. It displaces original survivors in 329, 250, and 454 selected-target conditions. Ordering tokens by original patch ID preserves the 53, 18, and 12 class differences between the two exclusion branches; reordering fixed-score exclusion alone changes no predictions. Thus token order does not explain these class differences. DeiT-256 and DeiT-B also show unselected-target support changes (4.23–5.95% and 1.31–7.36% across blocks). Replaying subsequent selections instead changes 342, 102, and 0 predictions relative to ordinary exclusion. The large archive retains per-image summaries rather than complete token records. The selective-measurement study uses a separate sample below.

E Attribution in Fitted Semantic Fields

LangSplatV2 represents each Gaussian using four of 64 fitted codewords (Li et al., 2025). Our intervention operates offline on stored supports. To reproduce the native support despite reconstruction error, we order native members before nonmembers, then use original probabilities and ID ties within each group. Exclusion restricts this declared order. With geometry and compositing fixed,

$$\boldsymbol{F}^{\mathrm{exc}} - \boldsymbol{F}^{\mathrm{sup}} = \boldsymbol{V}(\boldsymbol{A}^{\mathrm{exc}} - \boldsymbol{A}^{\mathrm{sup}})\boldsymbol{W}, \qquad \|\boldsymbol{F}^{\mathrm{exc}}_{:,p} - \boldsymbol{F}^{\mathrm{sup}}_{:,p}\| \le \sum_i W_{i,p}\lambda_i D_{V,i}. \tag{E.1}$$

Here $\boldsymbol{A}$ contains Gaussian coefficients and $\boldsymbol{W}$ nonnegative compositing weights. The bound follows from Equation A.3 and the triangle inequality for positive surviving masses. Nonlinear query scoring and thresholding prevent extending it directly to IoU. We use negative mean relevance within a fixed annotation mask as loss, so $\delta^q_\nu = r^0_\nu - r^q_\nu$. Each coefficient convention has its own native baseline; their comparison measures convention sensitivity.

Each convention contains 915 3D-OVS and 279 LERF view conditions. The 2,388 records cover 152,832 singleton and 66,864 pair measurements: 64 singletons and 28 pairs per record, with the eight-codeword pair pool fixed by native mass on the first calibration camera. Sign comparisons condition on both magnitudes exceeding 10^{-3} relevance; choice uses all groups, fixed ID ties, and tolerance 10^{-3}. IoU comparisons use an absolute difference above 0.01. Table E.1 reports the joint-pair extension to the main singleton results.

Table E.1: **Attribution disagreement for codeword pairs.** Each group contains 28 pairs from a fixed calibration pool. Entries give numerators and denominators under the same thresholds as the singleton comparison.

Benchmark	Coefficients	Opposite signs	Maximizers differ	Sets disjoint
3D-OVS	Retained	387/18,642	216/915	15/915
	Normalized	924/18,465	309/915	84/915
LERF	Retained	146/5,200	108/279	14/279
	Normalized	134/5,120	96/279	14/279

The 35 model directories represent 14 original scenes. Identical stored arrays across variants reduce 915 3D-OVS entries to 305 and 279 LERF entries to 208 per convention. The planned statistics

retain directory weights; combining duplicates leaves 3D-OVS rates unchanged and changes normalized LERF maximizer disagreement from 5.02% to 4.81%. The records do not establish why these endpoints coincide. All records pass their sampled inactive-candidate control and preserve replacement probability order. A secondary masked-softmax control differs by more than 10^{-3} relevance in 0/58,560 3D-OVS and 12/17,856 LERF singletons. The archive lacks full pixel arrays, and small continuous baseline errors can change thresholded masks.

F Component Selection and Predictive Fidelity

Each model uses 50 calibration and 1,000 disjoint, previously examined evaluation questions. Four OLMoE evaluation questions also appeared in an earlier exploratory set. Preservation selects k experts per layer with the smallest mean absolute singleton effects; criticality selects the largest mean signed effects. We freeze the resulting sets and evaluate joint exclusion with retained router weights, without assuming additivity. REAP (Lasby et al., 2026) averages $a_e\|\boldsymbol{v}_e\|_2$ over calibration token positions that activate expert e. Output norm averages $\|\boldsymbol{v}_e\|_2$ on the same positions; routing frequency divides activation count by all calibration token positions. The normalized REAP variant renormalizes weights over the native support only for calibration. All criteria share the evaluation operator. Effect criteria need labels and interventions, whereas REAP uses 50 native passes, so acquisition costs differ.

F.1 Predictive Measures and Paired Comparisons

Let p_i^q be the vocabulary distribution, $\widetilde{p}_i^q$ its conditional distribution over the four options, and $c_i^q = \mathbf{1}\{\widehat{y}_i^q = y_i\}$. We report

$$\begin{aligned} &\mathrm{LD} = n^{-1}\sum_i |\ell_i^q - \ell_i^0|, \quad \mathrm{KL} = n^{-1}\sum_i D_{\mathrm{KL}}(\widetilde{p}_i^0 \| \widetilde{p}_i^q), \\ &\mathrm{Acc} = n^{-1}\sum_i c_i^q, \qquad \mathrm{PD} = n^{-1}\sum_i \mathbf{1}\{\widehat{y}_i^q \neq \widehat{y}_i^0\}, \qquad \mathrm{Flips} = n^{-1}\sum_i |c_i^q - c_i^0|. \end{aligned} \tag{F.1}$$

LD uses full-vocabulary loss. KL excludes changes in probability mass outside the options. PD includes changes between incorrect answers, whereas Flips counts changes in correctness (Dutta et al., 2024). Rates are percentages; LD and KL use nats. We compute these measures from the same saved outputs and frozen modifications.

Table F.1: **Loss preservation relative to REAP.** $\Delta\mathrm{LD} = \mathrm{LD}_{\mathrm{REAP}} - \mathrm{LD}_q$ in nats. Intervals use 99.375% marginal confidence for eight comparisons on 1,000 common evaluation questions per model. All eight corresponding unadjusted paired 95% accuracy intervals include zero.

Model	k	Criterion	ΔLD	Adjusted interval
OLMoE	8	Suppression	0.02684	$[0.00911, 0.04269]$
	8	Exclusion	0.00878	$[-0.00806, 0.02485]$
	16	Suppression	0.03377	$[0.00462, 0.06292]$
	16	Exclusion	0.02891	$[0.00005, 0.05805]$
Qwen1.5-MoE	8	Suppression	0.03020	$[0.00882, 0.05270]$
	8	Exclusion	0.01243	$[-0.01000, 0.03517]$
	16	Suppression	0.05092	$[0.01373, 0.08820]$
	16	Exclusion	0.05201	$[0.01534, 0.08843]$

Table F.1 supports improved loss preservation for suppression calibration in all four settings, but establishes no accuracy advantage. Intervals condition on the calibration sample and frozen sets. The two effect criteria disagree on 10.5%/16.0% of OLMoE answers and 10.1%/15.9% of Qwen answers at $k = 8/16$; these compare the two modified models, not either model with the native prediction. For criticality, the suppression-minus-exclusion set contrast is $+0.8183, +0.1329$ nats in OLMoE and $-0.1317, -0.1140$ in Qwen. The paired intervals are $[0.7604, 0.8740]$, $[0.0736, 0.1927]$ (95%), and $[-0.1686, -0.0921]$, $[-0.1561, -0.0716]$ (98.75%), respectively. Thus the preferred calibration changes with the model and objective.

Independent input evaluation. A separate Qwen study uses 128 new calibration and 512 new test questions, excluding all 4,468 previous input IDs. At each k, 12 joint candidates comprise two historical effect selections, eight shifted rank windows (offsets 1, 2, 4, 8), and two random sets (seeds 20260916/17). At 384 candidate-question measurements, uniform allocation evaluates 12 candidates on 32 inputs; priority allocation evaluates three on 128 or six on 64. A two-stage variant starts with all candidates on 16 inputs. Selection minimizes calibration error, then log loss. All six adjusted paired accuracy intervals against uniform allocation include zero. This experiment establishes no accuracy or allocation advantage on new inputs.

G Selective Measurement

G.1 Calibration and Objective Dependence

For group g, define $Z_g = \mathbf{1}\{e_{\sup}(g) \neq e_q(g)\}$. The reference is fixed-score exclusion for experts and sparse fields, and rescored exclusion for DynamicViT. Among the $d > 0$ calibration disagreements, sort suppression margins and choose $\tau = m_{(\lceil 0.9d \rceil)}$. If $d = 0$, use $\tau = -\infty$. We calibrate each system and coefficient convention separately. For $Q_g = \mathbf{1}\{m_g \leq \tau\}$, measured groups query the entire reference candidate bank; other groups retain the suppression choice. Unqueried reference effects serve only to evaluate this retrospective policy. The calibration target does not guarantee 90% test recovery.

For N test groups, let $c = N^{-1}\sum_g Q_g$, $p = N^{-1}\sum_g Z_g$, and $r_g = \max_e \delta^q_{g,e} - \delta^q_{g,e_{\sup}(g)}$. Uniform random measurement of the same number of groups gives

$$\begin{aligned} R_{\text{margin}} &= N^{-1}\sum_g (1-Q_g)Z_g, \quad \mathbb{E}R_{\text{random}} = (1-c)p, \\ \bar{r}_{\text{margin}} &= N^{-1}\sum_g (1-Q_g)r_g, \qquad \mathbb{E}\bar{r}_{\text{random}} = (1-c)\bar{r}. \end{aligned} \tag{G.1}$$

Every group has random inclusion probability c, proving both expectations conditional on the recorded groups. Subtracting gives

$$\mathbb{E}R_{\text{random}} - R_{\text{margin}} = \text{Cov}_N(Q, Z), \qquad \mathbb{E}\bar{r}_{\text{random}} - \bar{r}_{\text{margin}} = \text{Cov}_N(Q, r), \tag{G.2}$$

where $\text{Cov}_N(Q, v) = N^{-1}\sum_g Q_g v_g - c\bar{v}$. Reducing disagreement therefore need not reduce regret. DynamicViT illustrates this distinction: its regret is 0.0001772 nats under margin allocation versus 0.0001103 under random allocation, despite lower disagreement. The records establish neither rescoring as the cause of this reversal nor a validated correction using regret weights.

G.2 Evaluation and Uncertainty

Table G.1: **Evaluation design for selective measurement.** Language results average 500 partitions of the same 100 questions. The other systems use a fixed recorded partition. Spatial partitions share fitted scenes and do not test transfer to unseen scenes.

System	Candidates	Sites	Calibration	Evaluation
OLMoE	64	16	50 questions	50 questions
Qwen1.5-MoE	60	24	50 questions	50 questions
DynamicViT	8	3	128 images	512 images
LangSplatV2	64	—	499 view groups	695 view groups

All sites of an input share its partition. Language results average 500 question splits (seed 20260924); bars at the calibrated operating point show central 95% split ranges, not independent-replication intervals. Margin bins use boundaries 0.001, 0.003, 0.01, 0.03, 0.1 nats and 2,000 question bootstrap resamples. Measurement curves vary calibration recovery in steps of 0.05. We calculate rates within each split before averaging. Costs count additional reference sweeps and exclude suppression and calibration, so they do not establish total computation savings.

DynamicViT uses ratio 0.9 on the checkpoint trained at 0.7. Its retrospective bank contains six selected and two unselected tokens per site, using router scores, norms, and seeded sampling. The

experiment concerns this bank, not all tokens: 1,147/1,536 measured groups reduce 199 disagreements to 24. The older archive lacks the canonical-order control of the larger mechanism study, limiting attribution of its discrepancies to rescoring.

Spatial evaluation uses 499 calibration and 695 test groups within the same fitted scenes; 21 single-view conditions contribute only descriptive measurements. Retained and normalized thresholds are 0.00257962 and 0.00260806. They reduce 47 disagreements to 4 after 200 measurements, and 85 to 12 after 215, respectively. Figure G.1(a) uses all 1,194 groups in approximate quintiles, retaining equal margins in the same bin. Panel (b) varies test thresholds descriptively and marks the operating points chosen solely from calibration views.

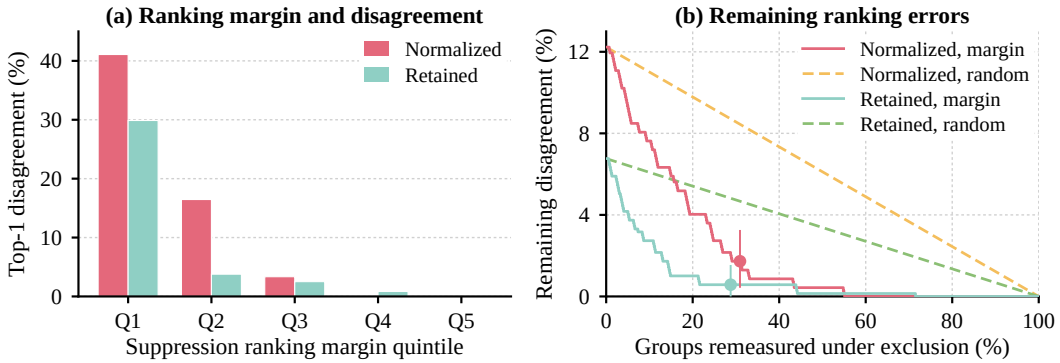

Figure G.1: **Ranking margins in fitted semantic fields.** (a) Small suppression margins identify more frequent component selection disagreements under both coefficient conventions. (b) Margin allocation reduces remaining disagreement relative to random allocation at the same measurement count. Marked points use calibrated thresholds; bars give conditional 95% intervals over 14 original scenes.

We bootstrap 14 original scenes, retaining dependent queries, views, and variants together. With 5,000 resamples and fixed fitted thresholds, remaining-disagreement 95% intervals are $[0.00, 1.59]\%$ and $[0.42, 3.19]\%$. They exclude threshold-estimation uncertainty. Treating 35 directories as independent scenes would use the wrong unit. Combining identical arrays and recalibrating on 223 groups yields 290 test groups: retained coefficients measure 151 and leave one disagreement (0.34%, random 2.98%); normalized coefficients measure 100 and leave two (0.69%, random 6.78%). Agreement benefits persist under this change in empirical weighting. The unknown effect discrepancy prevents certification of unmeasured choices. These results concern reference component identity, not accuracy, optimal joint pruning, or transfer to unseen scenes.