# WHAT MUST EXPERT FORMATION PRESERVE? FORMATION–DEPLOYMENT COMPATIBILITY IN CHANNEL-COMPOSED TWO-EXPERT MODELS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Mixture-of-experts (MoE) systems often couple two distinct choices: how experts are formed during training and how the resulting experts are combined at inference. We separate these factors in a controlled channel-composed expert model and study how training-time formation affects effective post-training deployment. Matched formation interventions vary assignment information, exact mixing values, low-order statistics, online adaptation, reference dependence, and stochasticity; the resulting frozen checkpoints are then evaluated under multiple deployment rules. Across ESC-50, Speech Commands V2, FSD50K, CIFAR-100, and Tiny ImageNet, no single formation–deployment pair dominates. To test whether learned coefficient-to-channel correspondence matters beyond the coefficient values themselves, we perform a post-hoc exploratory assignment intervention that preserves the exact learned coefficient multiset while permuting only its correspondence with representation channels. The aligned assignment outperforms its permutation mean on ESC-50, Speech Commands, CIFAR-100, and Tiny ImageNet, but not on FSD50K. On ESC-50, the aligned-minus-permuted effect is $+2.901$ Robust-B points with a descriptive fold-level 95% interval $[2.363,\ 3.438]$; the corresponding learned-channel versus learned-scalar difference is $+2.056$ points $[1.528,\ 2.583]$. Thus channel-assignment sensitivity is itself regime dependent. For the studied linear head, every fixed channel-wise composition compiles exactly into a single affine classifier. A separate reference-free $N \in \{2, 4, 8\}$ stress test shows that increasing expert count can help or hurt, with no monotonic trend across task, formation, and deployment. Together, these results identify formation–deployment compatibility as a distinct design problem.

## 1 INTRODUCTION

Mixture-of-experts (MoE) models increase capacity by combining parameterized experts under a routing or mixing rule (Jacobs et al., 1991; Shazeer et al., 2017; Fedus et al., 2022). In many modern MoE systems, the mechanism that shapes expert use during training is also retained at inference (Riquelme et al., 2021; Zhou et al., 2022; Puigcerver et al., 2024). This coupling makes two effects difficult to distinguish: a routing mechanism may matter because of the expert parameters it helps form, because of how those parameters are combined at inference, or because of both. End-to-end performance therefore does not by itself reveal which properties of training-time expert formation matter for a particular deployment rule.

We separate *expert formation* from *expert deployment* in a controlled two-expert channel-composition model. A shared representation is processed by two linear expert matrices whose contributions are mixed channel-wise. Conditional parameter mixtures and channel-wise gating already have clear precedents (Yang et al., 2019; Wang et al., 2020); our goal is therefore not to introduce a new MoE architecture, but to use this controlled model to separate expert formation from deployment. The training-time composition rule can be changed while holding the backbone and expert parameterization fixed, after which the resulting checkpoint is frozen and evaluated under alternative deployments. If $F$ denotes formation and $D$ deployment, we study

$$P(F, D) = \mathrm{Eval}(\Theta_F, D)$$



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.

where $\Theta_F$ is the checkpoint produced under $F$.

To probe formation, we construct matched controls that retain different information from one fixed bounded two-view reference field. ReferenceGate, where evaluated, retains the reference field together with its original sample/channel assignment; GlobalPool preserves the current reference-value multiset while destroying assignment; MomentBeta retains only the current empirical mean and variance; Retrospective FixedBeta removes online reference adaptation; and Uniform, IID-Beta, and FixedHalf remove reference dependence, with FixedHalf also removing stochasticity.

Deployment is varied only after $\Theta_F$ is frozen. We evaluate native inference when defined, fixed static composition, and learned scalar or channel-wise static composition. This enables same-checkpoint deployment comparisons without retraining the backbone or experts. For the studied linear head, every fixed channel-wise composition additionally compiles exactly into a single affine classifier, so static deployment requires no input-dependent router at inference. This distinguishes our question from prior work on alternative routing, expert merging, or modular deployment (Roller et al., 2021; Zuo et al., 2022; Honda et al., 2024; He et al., 2023; Muqeeth et al., 2024; Chen et al., 2023; Shen et al., 2023): we ask how information retained during formation affects which deployments remain effective afterward. Figure 1 summarizes the formation–deployment factorization and the information retained by each formation control.

Across ESC-50, Speech Commands V2, FSD50K, CIFAR-100, and Tiny ImageNet, neither formation nor deployment admits a task-independent ranking. Same-checkpoint deployment can reverse sign across regimes: post-hoc static deployment of LearnedMoE improves Speech Commands but degrades ESC-50, FSD50K, CIFAR-100, and Tiny ImageNet relative to native deployment. A complementary post-hoc assignment intervention preserves each learned MomentBeta coefficient multiset exactly while permuting only coefficient-to-channel correspondence. The aligned assignment outperforms its permutation mean on ESC-50, Speech Commands, CIFAR-100, and Tiny ImageNet, but not FSD50K. Thus even the relevance of channel identity is regime dependent. Together, these results show that formation and deployment should be studied jointly rather than in isolation.

A separate reference-free stress test reruns $N \in \{2, 4, 8\}$ experts under FixedEqual and symmetric Dirichlet formation. Increasing expert count can improve, saturate, or reduce performance depending on task, formation, and deployment. Learned channel-simplex deployment has the highest observed mean in 24 of 30 dataset–formation–expert-count conditions, with all six exceptions on FSD50K. Expert multiplicity therefore adds another compatibility axis rather than yielding a universal "more experts is better" trend.

Our contributions are:

- an experimental factorization of expert formation and deployment through $P(F, D) = \mathrm{Eval}(\Theta_F, D)$;
- matched formation controls that vary the information retained from the training-time mixing process;
- same-checkpoint deployment interventions across five datasets, including an exact-multiset channel-assignment test and exact static compilation for the studied linear head; and
- a matched reference-free $N \in \{2, 4, 8\}$ extension showing that expert multiplicity also interacts with formation and deployment.

## 2 Related Work

**Routing in MoE.** Classical MoE jointly learns gates and experts (Jacobs et al., 1991); modern sparse variants scale this idea through conditional computation (Shazeer et al., 2017; Fedus et al., 2022). V-MoE, Expert Choice, and Soft MoE instantiate different expert-assignment mechanisms (Riquelme et al., 2021; Zhou et al., 2022; Puigcerver et al., 2024), and a controlled vision study shows systematic differences among routing formulations (Liu et al., 2024). THOR uses stochastic expert activation, whereas Hash Layers replaces learned routing with fixed hashing (Zuo et al., 2022; Roller et al., 2021). Our question is different: rather than ranking routers, we ask how training-time formation affects which post-training deployments remain effective.

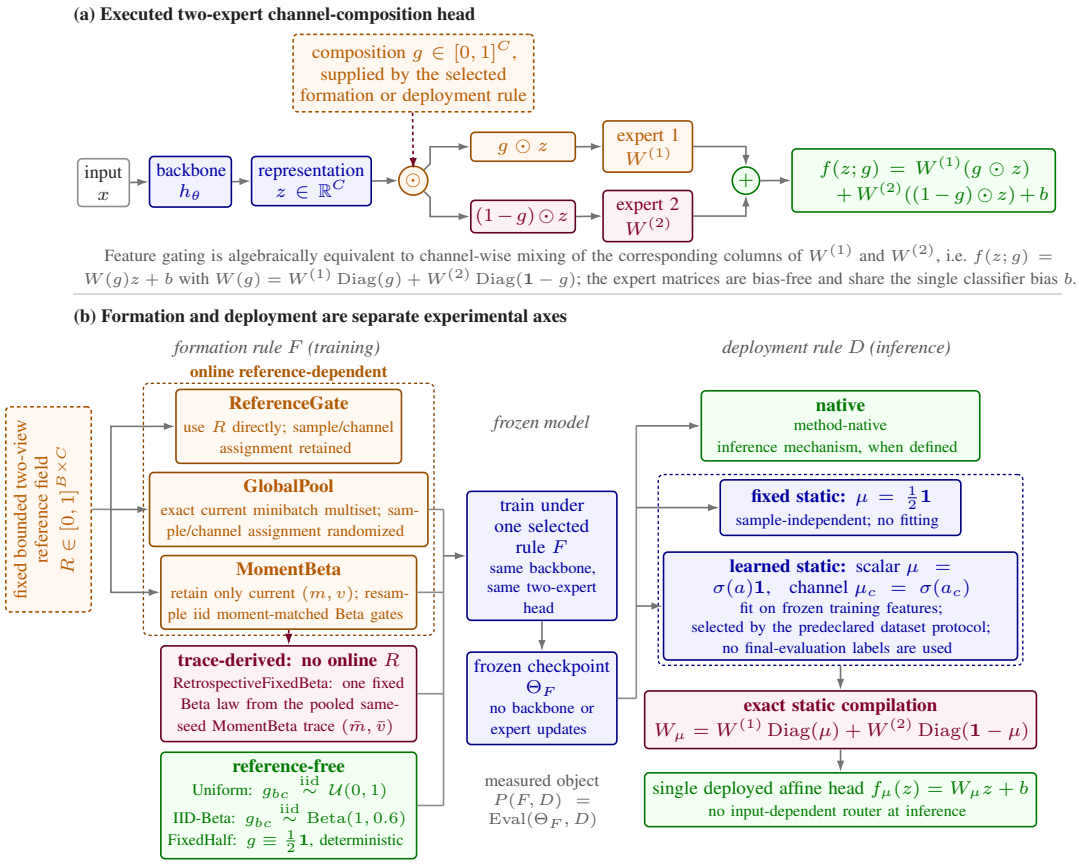


Figure 1: Formation and deployment are controlled separately. (a) A shared representation $z$ is processed by two bias-free expert matrices $W^{(1)}$ and $W^{(2)}$ with shared bias $b$, mixed channel-wise by $g$. The formation rule $F$ determines the training-time mixing process and yields a checkpoint $\Theta_F$. (b) After training, $\Theta_F$ is frozen and evaluated under alternative deployment rules $D$. Formation controls retain different information and dependence properties and are not assumed to form a nested hierarchy. Any fixed channel-wise deployment $\mu$ compiles exactly to the affine classifier $W_\mu = W^{(1)}\,\mathrm{Diag}(\mu) + W^{(2)}\,\mathrm{Diag}(1-\mu)$, $f_\mu(z) = W_\mu z + b$, requiring neither an input-dependent router nor two expert branches at inference.

**Conditional expert mixtures.** Our head is closer to conditional parameter mixtures than to capacity-limited token routing. CondConv combines expert parameters as a function of the input and establishes the corresponding linear-mixture equivalence (Yang et al., 2019); DeepMoE applies sparse channel-wise gating (Wang et al., 2020). We use a simple two-expert linear construction because it exposes formation and deployment as independently controllable variables and permits exact compilation for static composition; the architecture itself is not claimed as a new MoE scaling mechanism.

**Expert aggregation, merging, and post-training deployment.** Prior work aggregates expert outputs or parameters, merges experts for efficient deployment, or encourages modular expert structure (Honda et al., 2024; He et al., 2023; Muqeeth et al., 2024; Chen et al., 2023; Shen et al., 2023). Recent work also directly optimizes how existing experts are combined: NAMEx derives expert-merging weights through Nash bargaining (Nguyen et al., 2026). Our question is related but distinct. We do not propose a new expert-merging optimizer; instead, we intervene separately on how expert parameters are formed and how a frozen checkpoint is subsequently composed, asking which formation-time information remains useful under alternative deployments. For our linear channel-composed head, once a fixed deployment is selected, collapsing the two experts into a single classifier is lossless and requires no additional merging optimization.

# 3 FORMATION–DEPLOYMENT FRAMEWORK

## 3.1 TWO-EXPERT CHANNEL COMPOSITION

A shared backbone maps $x$ to $z = h_\theta(x) \in \mathbb{R}^C$. Two bias-free expert matrices $W^{(1)}, W^{(2)} \in \mathbb{R}^{K\times C}$ share one bias $b \in \mathbb{R}^K$. Given a channel-wise composition vector $g \in [0,1]^C$,

$$f(z; g) = W^{(1)}(g \odot z) + W^{(2)}((\mathbf{1} - g) \odot z) + b \tag{1}$$

Equivalently,

$$f(z; g) = W(g)z + b, \qquad W(g) = W^{(1)} \operatorname{Diag}(g) + W^{(2)} \operatorname{Diag}(\mathbf{1} - g) \tag{2}$$

Thus $g_c$ controls the relative contribution of the two expert-weight columns associated with representation channel $c$.

Training uses two independently augmented views of each example and the mean loss over the two views. For the matched formation controls, the same per-example channel gate is applied to both augmented views, so that the intervention concerns gate construction rather than view-specific routing. For multi-label FSD50K, cross-entropy is replaced by binary cross-entropy.

A *formation rule* $F$ specifies how $g$ is generated during training, yielding $\Theta_F = \mathrm{Train}(F)$. A *deployment rule* $D$ specifies how the already-trained experts are used at inference. For compatible pairs we study

$$P(F, D) = \mathrm{Eval}(\Theta_F, D) \tag{3}$$

## 3.2 FORMATION CONTROLS: WHAT INFORMATION IS RETAINED

For a minibatch, let $G = [g_{ic}] \in [0,1]^{B\times C}$ denote the gate matrix, with row $g_i \in [0,1]^C$ applied to example $i$. ReferenceGate, GlobalPool, and MomentBeta use a bounded training-time reference field $R = [r_{ic}] \in [0,1]^{B\times C}$ computed from the two augmented views. The reference field is a fixed experimental anchor: its formula is not selected per formation rule or dataset, its values are detached from gradient computation, and it is not treated as semantic ground truth. Its executed construction is given in Appendix A.

Uniform, IID-Beta, and FixedHalf never use $R$. Retrospective FixedBeta uses no online $R$ during its own optimization, but its fixed distribution is calibrated from the same-seed MomentBeta training trace. The controls therefore probe both reference-dependent and reference-independent formation. These controls form an information-retention suite rather than a strictly nested stochastic family: GlobalPool preserves an exact finite multiset with dependence induced by sampling without replacement, whereas MomentBeta generates iid values from a moment-matched parametric law. A compact summary of the controls is given in Appendix A, Table 5.

**ReferenceGate.** ReferenceGate uses the reference field directly as the sample- and channel-dependent training composition,

$$G = R$$

It therefore retains the reference values together with their original sample/channel assignment. Under native deployment, the same reference-gate construction is evaluated on the inference views.

**GlobalPool.** Let $s = \mathrm{vec}(R) \in [0,1]^M$, where $M = BC$. For a uniformly random permutation $\Pi$,

$$\mathrm{vec}(G)_j = s_{\Pi(j)}$$

GlobalPool therefore preserves the exact current minibatch multiset while destroying sample/channel assignment identity.

**MomentBeta.** Let

$$m = \frac{1}{M}\sum_j s_j, \qquad v = \frac{1}{M}\sum_j (s_j - m)^2$$

MomentBeta samples

$$G_{ic} \overset{\text{iid}}{\sim} \mathrm{Beta}(\alpha, \beta)$$

where $(\alpha, \beta)$ are parameterized from the current empirical mean $m$ and variance $v$; Appendix B gives the interior moment inversion.

**Retrospective FixedBeta and reference-free controls.** Retrospective FixedBeta (RetroFixedBeta) pools the same-seed MomentBeta training trace into $(\bar{m}, \bar{v})$ and trains a separate checkpoint using one fixed Beta law. It therefore removes online reference adaptation while remaining trace-derived. Uniform samples iid $U(0,1)$ gates, IID-Beta uses the fixed $\mathrm{Beta}(1, 0.6)$ law, and FixedHalf sets $g = \frac{1}{2}\mathbf{1}$. The latter three are reference-free, and FixedHalf additionally removes stochasticity. The RetroFixedBeta trace aggregation and the formal MomentBeta/GlobalPool properties are given in Appendix B.

### 3.3 Static deployment and exact compilation

After formation, the neural checkpoint is frozen. A deterministic $\mu \in [0,1]^C$ defines

$$f_\mu(z) = W^{(1)}(\mu \odot z) + W^{(2)}((\mathbf{1} - \mu) \odot z) + b \tag{4}$$

We compare fixed-half deployment $\mu = \frac{1}{2}\mathbf{1}$, learned scalar deployment $\mu = \mathrm{sigmoid}(a)\mathbf{1}$, and learned channel-wise deployment $\mu_c = \mathrm{sigmoid}(a_c)$.

For learned static deployment, only the static composition parameters ($a$ for scalar deployment or $\{a_c\}_{c=1}^C$ for channel-wise deployment) are optimized on frozen training features; the backbone and expert parameters remain fixed. When a validation split exists, the predeclared dataset-specific validation criterion selects the iterate. The held-out final-evaluation split is never used to fit $\mu$. ESC-50 instead follows its predeclared fixed static schedule because its final protocol trains on four official folds and tests on the fifth. Full dataset-specific fitting and selection details are given in Appendix C.

For any fixed $\mu$, define

$$W_\mu = W^{(1)}\,\mathrm{Diag}(\mu) + W^{(2)}\,\mathrm{Diag}(\mathbf{1} - \mu)$$

Then

$$f_\mu(z) = W_\mu z + b$$

Thus every fixed static composition compiles exactly into a single affine classifier: no input-dependent router or two expert branches are required in the deployed head. The formal statement and proof are given in Theorem 1 in Appendix B.

### 3.4 Reference-free $N$-expert stress test

The principal formation-information study uses two experts, for which the reference-dependent interventions have a direct complementary-gate interpretation. We test expert multiplicity separately using a reference-free $N \in \{2, 4, 8\}$ extension with FixedEqual and symmetric Dirichlet formation. Each frozen checkpoint is evaluated under FixedEqual, learned global-simplex, and learned channel-simplex deployment; $N = 2$ is rerun within the same implementation and serves as the matched anchor for $N = 4$ and $N = 8$.

For fixed deployment weights $\mu_{ec}$ satisfying $\sum_{e=1}^N \mu_{ec} = 1$ for each channel $c$, the compiled classifier satisfies

$$W_\mu[:, c] = \sum_{e=1}^{N} \mu_{ec} W^{(e)}[:, c], \qquad f_\mu(z) = W_\mu z + b$$

Thus exact static compilation extends directly to the $N$-expert head. Training-head capacity grows with $N$, whereas the compiled classifier remains a single $K(C+1)$-parameter affine head. Full notation and compilation algebra are given in Appendix B, and the execution protocol and parameter counts in Appendix C.

## 4 Experimental Design

### 4.1 Datasets and protocols

We evaluate five heterogeneous datasets using the primary metrics and splits summarized in Table 1. All principal experiments use five scientific seeds. ESC-50 uses all five official folds and reports Robust-B, the mean accuracy over 20 predeclared waveform-corruption conditions formed by five corruption families crossed with severities $\{0.25, 0.50, 0.75, 1.00\}$. In Mode B, both views use the same family and severity but independent corruption randomness. Complete preprocessing, dataset splits, and selection rules are given in Appendix C.

Table 1: Five evaluation regimes. Robust-B is the predeclared ESC robustness aggregate; Speech, CIFAR, and Tiny use clean top-1 accuracy.

| Dataset | Task | Final metric | Scale/split |
|---|---|---|---|
| ESC-50 Piczak (2015) | audio, single-label | Robust-B acc. | 5 official folds |
| Speech Cmds V2 Warden (2018) | speech | clean acc. | official split |
| FSD50K Fonseca et al. (2022) | audio, multi-label | mAP | official eval |
| CIFAR-100 Krizhevsky (2009) | vision | clean acc. | 45k/5k/10k |
| Tiny ImageNet Le and Yang (2015) | vision | clean acc. | 90k/10k/10k |

### 4.2 Matched model family and comparators

The controlled family uses the same three-block convolutional backbone and a $C = 128$ representation, with the two expert matrices in Eq. (1). Dense uses the same backbone with one affine classifier; therefore the backbone is matched but the classifier parameter count is not identical. We report the head-parameter formulas in Appendix C instead of describing the experiment as fully parameter matched. We additionally evaluate an architecture-matched learned two-way soft router (Learned-MoE), THOR-style stochastic dispatch with consistency (Zuo et al., 2022), Shake-Shake-style branch mixing (Gastaldi, 2017), and SMEAR-style adaptive parameter merging (Muqeeth et al., 2024). The latter are labeled "style" because they isolate mechanisms under our matched backbone rather than reproduce the original full architectures. For vision, fixed reference-field computation is transferred unchanged to RGB feature maps only for GlobalPool/MomentBeta; we make no claim that the statistic itself is modality neutral. This strict-transfer choice is a limitation and a stress test, not a new vision router.

### 4.3 Reporting, statistical scope, and stress tests

All principal experiments use five scientific seeds. For ESC-50, the five seeds are first averaged within each official held-out fold, and variation is reported across the five fold means. The 25 seed–fold cells are retained for paired diagnostics but are not treated as 25 independent inferential replicates; central ESC comparisons therefore distinguish descriptive cell-wise win counts from fold-level summaries and intervals. The principal cross-dataset study is descriptive because metrics and selection protocols differ across tasks. Small differences between empirical means are not used to support universal superiority claims.

**Post-hoc channel-assignment audit.** For each principal MomentBeta checkpoint, we additionally evaluate 100 deterministic permutations of the already learned channel-composition vector. Each permutation preserves the exact finite coefficient multiset $\{\mu_c\}_{c=1}^{C}$ and changes only its correspondence with representation channels. No backbone, expert weight, learned coefficient value, or probe is refit. Within each replicate, the aligned composition and all permutations are evaluated under identical dataset-specific replay conditions using the same principal metric. ESC-50 is summarized fold-first after averaging scientific seeds within each official fold, whereas Speech Commands, FSD50K, CIFAR-100, and Tiny ImageNet are summarized across their five scientific seeds. This audit is post-hoc and exploratory, is not used for model, checkpoint, or hyperparameter selection, and is interpreted separately within each dataset's principal metric defined in Table 1.

The reference-free expert-count study of Section 3.4 is treated as an exploratory factorial stress test. For Speech Commands, FSD50K, CIFAR-100, and Tiny ImageNet, the $N = 4$ and $N = 8$ effects are paired by scientific seed against the rerun $N = 2$ anchor defined in Section 3.4; ESC additionally reports seed–fold directions descriptively, while uncertainty remains based on the five official fold aggregates. We emphasize paired effect sizes, direction counts, and interactions rather than isolated unadjusted $p$-values or claims of an "optimal" expert count. The complete formation–deployment matrix is reported in Appendix H. Reference strength is treated separately as a nuisance-parameter sensitivity analysis rather than a model-selection axis. The principal reference-dependent experiments use the pre-specified $\lambda = 0.2$; GlobalPool, MomentBeta, and Retrospective FixedBeta are additionally rerun at $\lambda \in \{0.05, 0.10, 0.50, 1.00\}$ with matched scientific seeds and all five ESC folds. These alternative settings are reported only as sensitivity analyses and are not used to choose the operating point; the $\lambda = 0.2$ runs remain the basis of the main formation–deployment results.

Table 2: Representative principal two-expert results. Values are percentages, with FSD50K mAP reported on the same 0–100 scale. For methods labeled -STATIC, the suffix denotes learned channel-wise static deployment of the frozen checkpoint. Bold indicates the highest observed mean in the principal two-expert experiment for each dataset; complete principal results are in Appendix D.

| Method | ESC-50 ↑ Robust-B acc. | Speech ↑ clean acc. | FSD50K ↑ mAP | CIFAR-100 ↑ clean acc. | Tiny IN ↑ clean acc. |
|---|---|---|---|---|---|
| Dense | 61.712 | 89.405 | 33.847 | 50.940 | 34.436 |
| FixedHalf-Static | 64.379 | **89.903** | 33.929 | **51.772** | 33.578 |
| GlobalPool-Static | 64.112 | 89.423 | 33.886 | 51.648 | **34.792** |
| MomentBeta-Static | **64.566** | 89.525 | 34.141 | 51.558 | 33.734 |
| Uniform-Static | 63.418 | 89.617 | 33.755 | 51.664 | 34.228 |
| LearnedMoE-Native | 59.392 | 88.798 | 33.933 | 51.010 | 33.912 |
| LearnedMoE-Static | 53.129 | 89.458 | 27.017 | 47.918 | 30.212 |
| SMEAR-style-Native | 50.872 | 85.767 | **35.783** | 46.590 | 32.366 |

# 5 RESULTS

## 5.1 NO UNIVERSAL FORMATION/DEPLOYMENT WINNER

Table 2 gives representative methods; complete tables are in Appendix D. Four different mechanisms win across the five datasets. MomentBeta-Static leads ESC-50, FixedHalf-Static leads Speech Commands and CIFAR-100, SMEAR-style-Native leads FSD50K, and GlobalPool-Static leads Tiny ImageNet. No formation or deployment mechanism is best across all five datasets.

The formation controls also do not collapse to "random mixing works." On ESC-50, Uniform-Static trails MomentBeta by 1.1485 Robust-B points; all five fold means favor MomentBeta and the descriptive fold-level 95% interval is $[0.125, 2.172]$. Yet on Speech Commands, Uniform slightly exceeds MomentBeta, and IID-Beta remains comparatively close to MomentBeta across several datasets, including ESC-50, Speech Commands, FSD50K, and CIFAR-100. Thus neither generic stochasticity nor moment-matched Beta resampling provides a task-independent explanation. Figure 2 in Appendix D summarizes these cross-dataset differences relative to Dense. Tiny ImageNet is particularly informative. GlobalPool preserves the exact anonymous reference-value multiset but destroys its assignment and reaches 34.792, while Retrospective FixedBeta reaches 34.534, both above Dense (34.436). In contrast, FixedHalf (33.578), IID-Beta (33.704), and MomentBeta (33.734) are below Dense; the first two are reference-free, whereas MomentBeta retains only current reference-field moments. From means alone we do not claim a significant GlobalPool advantage; rather, the ordering is consistent with regime-dependent information sufficiency. Because GlobalPool differs from MomentBeta in both the exact finite marginal values and the without-replacement dependence structure, this comparison does not identify either factor alone.

## 5.2 DEPLOYMENT COMPATIBILITY DEPENDS ON FORMATION

Static deployment is not intrinsically beneficial. For the same LearnedMoE formation family (Table 2), post-hoc static deployment improves Speech clean accuracy by $+0.660$ points but changes performance by $-6.263$ Robust-B points on ESC-50, $-6.916$ mAP points on FSD50K, $-3.092$ accuracy points on CIFAR-100, and $-3.700$ points on Tiny ImageNet. Thus, across tasks, the sign of the same deployment intervention can reverse even within the same formation family. A contrasting pattern appears for ReferenceGate in the three audio datasets where both native and static deployments were evaluated: ReferenceGate-Static exceeds ReferenceGate-Native by $+4.402$ Robust-B points on ESC-50, $+1.499$ clean-accuracy points on Speech Commands, and $+0.385$ mAP points on FSD50K. Together, these comparisons show that neither "static is always better" nor "native routing is always necessary" holds across the evaluated formation–task pairs.

## 5.3 CHANNEL-ASSIGNMENT SENSITIVITY IS REGIME DEPENDENT

A learned channel vector can outperform simpler static compositions, but this alone does not establish that the specific coefficient-to-channel correspondence matters: relative to a scalar composition, a

channel vector has greater fitting freedom and may also benefit from the marginal distribution of its learned coefficients. We therefore perform a post-hoc assignment intervention on the principal MomentBeta checkpoints. For each checkpoint, we preserve the exact learned coefficient multiset $\{\mu_c\}_{c=1}^{C}$ and apply 100 permutations generated from fixed random seeds only to its assignment across representation channels. No backbone, expert weight, learned coefficient value, or probe is refit. Within each paired replicate, the aligned vector and all permutations are evaluated under identical replay conditions. The resulting paired effects are summarized in Table 3.

Table 3: Post-hoc channel-assignment intervention. For each principal MomentBeta checkpoint, we compare the learned channel assignment with 100 permutations preserving the exact coefficient multiset. Effects are aligned minus the checkpoint-specific permutation mean in the dataset's primary metric; no parameters are refit. Descriptive 95% $t$-intervals use five folds for ESC-50 and five scientific seeds otherwise.

| Dataset | Aligned − permuted | 95% interval | Positive paired units |
|---|---|---|---|
| ESC-50 | $+2.901$ | $[2.363, 3.438]$ | 25/25 cells; 5/5 folds |
| Speech Commands | $+0.464$ | $[0.290, 0.639]$ | 5/5 seeds |
| FSD50K | $-0.012$ | $[-0.098, 0.075]$ | 2/5 seeds |
| CIFAR-100 | $+0.415$ | $[0.106, 0.724]$ | 5/5 seeds |
| Tiny ImageNet | $+0.276$ | $[0.125, 0.426]$ | 5/5 seeds |

The assignment intervention reveals a clear regime-dependent pattern. The aligned composition outperforms its mean permuted counterpart on ESC-50, Speech Commands, CIFAR-100, and Tiny ImageNet, with positive paired effects in every reporting unit used for those four datasets. The effect is largest on ESC-50 ($+2.901$ points) and smaller but consistent on Speech Commands ($+0.464$), CIFAR-100 ($+0.415$), and Tiny ImageNet ($+0.276$). In contrast, FSD50K shows essentially no assignment advantage ($-0.012$ points, 2/5 positive seeds), with its interval spanning zero. Thus, channel-wise parameterization alone does not imply that channel identity is consequential. In four of the five evaluated regimes, the specific learned coefficient-to-channel correspondence carries performance-relevant structure: randomizing that correspondence while preserving the exact coefficient values reduces mean performance. FSD50K provides the complementary case in which the same intervention has no detectable benefit for the aligned assignment. Channel-assignment sensitivity is therefore itself checkpoint- and regime-dependent.

### 5.4 Expert multiplicity is a compatibility axis, not a monotonic scaling law

Table 4 summarizes the reference-free $N$-expert stress test under learned channel-simplex deployment; the complete formation–deployment matrices and paired effects are reported in Appendix H. No common expert count gives the highest observed mean across datasets: the maximum occurs at $N = 4$ on ESC-50, $N = 2$ on FSD50K, $N = 8$ on Speech Commands, $N = 4$ on CIFAR-100, and $N = 8$ on Tiny ImageNet when all evaluated formations and deployments are considered.

Table 4: Reference-free expert-count stress test under learned channel-simplex deployment. Values are on the 0–100 scale; FSD50K reports mAP and the other datasets report accuracy. Bold indicates the highest observed mean within each dataset row.

| | Dirichlet formation | | | FixedEqual formation | | |
|---|---|---|---|---|---|---|
| Dataset / primary metric | $N = 2$ | $N = 4$ | $N = 8$ | $N = 2$ | $N = 4$ | $N = 8$ |
| ESC-50 Robust-B | 64.225 | **65.558** | 64.216 | 64.674 | 65.144 | 64.994 |
| FSD50K mAP | 34.407 | 34.157 | 34.269 | **34.412** | 34.191 | 33.739 |
| Speech Commands | 89.877 | 89.363 | 90.228 | 90.152 | 89.868 | **90.637** |
| CIFAR-100 | 51.208 | 51.808 | 51.868 | 51.692 | **52.294** | 51.956 |
| Tiny ImageNet | 34.128 | 33.928 | 33.388 | 33.598 | 33.986 | **34.310** |

The paired effects make the interaction clearer. Under FixedEqual formation and learned channel deployment, Speech Commands improves by $+0.485$ points from $N = 2$ to $N = 8$ with 5/5 seed-wise wins, while CIFAR-100 improves by $+0.602$ points from $N = 2$ to $N = 4$ with 4/5 wins.

FSD50K instead has negative mean $N = 4$ and $N = 8$ effects in all 12 formation–deployment comparisons against the rerun $N = 2$ anchor. Tiny ImageNet provides a direct formation interaction: under channel deployment, the $N = 8 - N = 2$ effect is $-0.740$ points for Dirichlet formation but $+0.712$ for FixedEqual formation. Deployment also interacts with expert count. Learned channel simplex has the highest observed deployment mean in 24 of 30 dataset–formation–$N$ conditions, with all six exceptions on FSD50K. Greater expert utilization is not sufficient either: for FSD50K with FixedEqual formation and channel deployment, $N_{\text{eff}}$ rises from $1.85$ to $3.38$ to $6.04$ as $N$ increases from 2 to 4 to 8, while mAP falls from $34.412$ to $34.191$ to $33.739$. Expert multiplicity therefore changes the formation–deployment compatibility structure rather than providing a monotonic scaling benefit. Reference-strength sensitivity is reported separately in Appendix G. The principal analysis remains fixed at the pre-specified $\lambda = 0.2$; no sensitivity setting is used for model selection.

## 6 Discussion and Limitations

The formation results do not support a task-independent hierarchy of retained mixing information. Exact sample/channel assignment, an anonymous value multiset, current mean and variance, and reference-free formation are competitive in different regimes. On ESC-50, generic stochastic mixing does not explain the MomentBeta result: Uniform trails FixedHalf, whereas MomentBeta returns to approximately the FixedHalf level. Tiny ImageNet gives a different ordering, with GlobalPool above Dense but MomentBeta below it. Because GlobalPool and MomentBeta differ in both retained information and dependence structure, these results support regime-dependent information sufficiency rather than a universally necessary reference property.

Deployment is similarly checkpoint dependent. Static deployment improves LearnedMoE on Speech Commands but degrades it on the other four datasets, whereas ReferenceGate shows positive static-minus-native differences on the three audio evaluations where both deployments were measured. The channel-assignment intervention further shows that preserving learned coefficient values while permuting only their correspondence with representation channels reduces performance on ESC-50, Speech Commands, CIFAR-100, and Tiny ImageNet, but not FSD50K. Thus channel-wise parameterization does not by itself make channel identity consequential; its relevance depends on the formed checkpoint and evaluation regime. For the studied linear head, every fixed composition compiles exactly into a single affine classifier and therefore requires no input-dependent inference router.

The reference-free $N$-expert study extends this compatibility picture to expert multiplicity. The expert count with the highest observed mean differs across datasets, and Tiny ImageNet shows that even the sign of the $N = 8$ versus $N = 2$ effect can reverse across formation rules. FSD50K further shows that a larger effective expert count can coincide with lower performance. The $\lambda$ sensitivity analysis provides a related boundary: no common alternative coefficient or formation rule dominates across datasets. Together, these results argue against treating expert count, effective expert count, or the reference-generator coefficient as independently monotonic design axes.

These conclusions are bounded by the experimental design. Cross-dataset comparisons are descriptive; non-ESC datasets use five scientific seeds, and ESC-50 uses fold-first uncertainty over five official folds. Exact compilation applies only to the studied linear channel-composed head. The $N$-expert extension is reference-free and increases training-head capacity with $N$, so it is not a parameter-matched scaling experiment. The vision experiments strictly transfer the fixed reference construction to RGB feature maps, and the sensitivity analysis varies only $\lambda$ within one reference-generator family.

## 7 Conclusion

We separated how experts are formed during training from how the resulting checkpoint is deployed. Across five datasets, neither retained formation information nor deployment strategy has a task-independent ordering, and an exact-multiset intervention shows that learned coefficient-to-channel correspondence is consequential in some regimes but not others. Every fixed composition in the studied head compiles exactly into one affine classifier. A reference-free $N \in \{2, 4, 8\}$ extension further shows that expert count interacts with both formation and deployment rather than producing a monotonic scaling benefit. The resulting design problem is therefore one of formation–deployment compatibility, with expert multiplicity adding a third interaction axis.

**Reproducibility statement.** Section 4 and Appendix C specify dataset splits, training and checkpoint-selection rules, static-deployment fitting, comparator implementations, and reporting conventions. All principal experiments use five scientific seeds; ESC-50 additionally uses all five official folds with fold-first uncertainty. Principal reference-dependent runs fix $\lambda = 0.2$, the $N$-expert study reruns its matched $N = 2$ anchors, and the post-hoc assignment audit uses 100 fixed-seed permutations without refitting the backbone, experts, learned coefficients, or probes. Complete results and sensitivity analyses are reported in Appendices D–H.

## AI USE STATEMENT

Generative AI tools were used for language editing, including grammar and clarity improvements, and for assistance with LaTeX formatting and manuscript organization. Experimental measurements and numerical results reported in the paper were produced by executed project code rather than generated by an AI system. The authors reviewed the final manuscript, experimental results, mathematical content, and citations and retain responsibility for their correctness.

## REFERENCES


Zitian Chen, Yikang Shen, Mingyu Ding, Zhenfang Chen, Hengshuang Zhao, Erik G. Learned-Miller, and Chuang Gan. Mod-Squad: Designing mixtures of experts as modular multi-task learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11828–11837, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Chen_Mod-Squad_Designing_Mixtures_of_Experts_As_Modular_Multi-Task_Learners_CVPR_2023_paper.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.

Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. FSD50K: An open dataset of human-labeled sound events. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:829–852, 2022. doi: 10.1109/TASLP.2021.3133208.

Xavier Gastaldi. Shake-Shake regularization of 3-branch residual networks. In *International Conference on Learning Representations Workshop*, 2017. URL https://openreview.net/forum?id=HkO-PCmYl.

Shwai He, Run-Ze Fan, Liang Ding, Li Shen, Tianyi Zhou, and Dacheng Tao. Merging experts into one: Improving computational efficiency of mixture of experts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14685–14691. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.907. URL https://aclanthology.org/2023.emnlp-main.907/.

Ukyo Honda, Tatsushi Oka, Peinan Zhang, and Masato Mita. Not eliminate but aggregate: Post-hoc control over mixture-of-experts to address shortcut shifts in natural language understanding. *Transactions of the Association for Computational Linguistics*, 12:1268–1289, 2024. doi: 10.1162/tacl_a_00701. URL https://aclanthology.org/2024.tacl-1.70/.

Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991. doi: 10.1162/neco.1991.3.1.79.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Ya Le and Xuan Yang. Tiny ImageNet visual recognition challenge. CS231N course project report, Stanford University, 2015. URL https://cs231n.stanford.edu/reports/2015/pdfs/yle_project.pdf.

Tianlin Liu, Mathieu Blondel, Carlos Riquelme Ruiz, and Joan Puigcerver. Routers in vision mixture of experts: An empirical study. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=aHk3vctnf1.

Mohammed Muqeeth, Haokun Liu, and Colin Raffel. Soft merging of experts with adaptive routing. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=7I199lc54z.

Dung V. Nguyen, Anh T. Nguyen, Minh H. Nguyen, Luc Q. Nguyen, Shiqi Jiang, Ethan Fetaya, Linh Duy Tran, Gal Chechik, and Tan M. Nguyen. Expert merging in sparse mixture of experts with nash bargaining. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/ef9f16cd0cb8d7797ea35c0ee7964cdf-Abstract-Conference.html.

Karol J. Piczak. ESC: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM International Conference on Multimedia*, pages 1015–1018, 2015. doi: 10.1145/2733373.2806390.

Joan Puigcerver, Carlos Riquelme Ruiz, Basil Mustafa, and Neil Houlsby. From sparse to soft mixtures of experts. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/79fea214543ba263952ac3f4e5452b14-Abstract-Conference.html.

Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/48237d9f2dea8c74c2a72126cf63d933-Abstract.html.

Stephen Roller, Sainbayar Sukhbaatar, Arthur Szlam, and Jason Weston. Hash layers for large sparse models. In *Advances in Neural Information Processing Systems*, volume 34, pages 17555–17566, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/92bf5e6240737e0326ea59846a83e076-Abstract.html.

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

Yikang Shen, Zheyu Zhang, Tianyou Cao, Shawn Tan, Zhenfang Chen, and Chuang Gan. ModuleFormer: Modularity emerges from mixture-of-experts. *arXiv preprint arXiv:2306.04640*, 2023. URL https://arxiv.org/abs/2306.04640. Preprint.

Xin Wang, Fisher Yu, Lisa Dunlap, Yi-An Ma, Ruth Wang, Azalia Mirhoseini, Trevor Darrell, and Joseph E. Gonzalez. Deep mixture of experts via shallow embedding. In *Proceedings of the 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, pages 552–562. PMLR, 2020. URL https://proceedings.mlr.press/v115/wang20d.html.

Pete Warden. Speech commands: A dataset for limited-vocabulary speech recognition. *arXiv preprint arXiv:1804.03209*, 2018. URL https://arxiv.org/abs/1804.03209.

Brandon Yang, Gabriel Bender, Quoc V. Le, and Jiquan Ngiam. CondConv: Conditionally parameterized convolutions for efficient inference. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/f2201f5191c4e92cc5af043eebfd0946-Abstract.html.

Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Y. Zhao, Andrew M. Dai, Zhifeng Chen, Quoc V. Le, and James Laudon. Mixture-of-experts with Expert Choice routing. In *Advances in Neural Information Processing Systems*, volume 35, pages 7103–7114, 2022. doi: 10.52202/068431-0515. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/2f00ecd787b432c1d36f3de9800728eb-Abstract-Conference.html.

Simiao Zuo, Xiaodong Liu, Jian Jiao, Young Jin Kim, Hany Hassan, Ruofei Zhang, Tuo Zhao, and Jianfeng Gao. Taming sparsely activated transformer with stochastic experts. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2110.04260.

## A Exact Reference-Field Construction

To instantiate ReferenceGate, GlobalPool, and MomentBeta, we use one fixed bounded two-view reference field $R$. This appendix records its executed construction for exact reproducibility. The reference-field formula is not tuned per dataset or formation method, and no optimality or semantic-ground-truth claim is attached to it.

For context, Table 5 summarizes how the formation controls retain or avoid reference-derived information.

Table 5: Formation-control summary. ReferenceGate uses the reference field directly and therefore retains its sample/channel assignment. GlobalPool preserves its exact current values while randomizing assignment, and MomentBeta retains only current empirical mean and variance. RetroFixedBeta is trace-derived; Uniform, IID-Beta, and FixedHalf are reference-free. "Adaptive" means minibatch-dependent.

| Formation | retained information | adaptive? |
|---|---|---|
| ReferenceGate | exact $R$, including sample/channel assignment | yes |
| GlobalPool | exact minibatch multiset | yes |
| MomentBeta | current $(m, v)$ | yes |
| RetroFixedBeta | pooled trace $(\bar{m}, \bar{v})$ | no |
| Uniform | none | no |
| IID-Beta(1,0.6) | none | no |
| FixedHalf | none | no |

For one example, let the final convolutional feature maps of the two augmented views be

$$A^{(1)}, A^{(2)} \in \mathbb{R}^{C\times U\times V}$$

We suppress the example index below. The implementation uses $\epsilon = 10^{-6}$ and defines the numerically stabilized normalization

$$\mathrm{norm}_\epsilon(x) = \frac{x}{\max(\|x\|_2, \epsilon)}$$

For channel $c$ and index $u$, it computes the clipped directional agreement

$$a_{cu} = \mathrm{clip}_{[0,1]}\left(\left\langle \mathrm{norm}_\epsilon(A^{(1)}_{cu:}), \mathrm{norm}_\epsilon(A^{(2)}_{cu:})\right\rangle\right) \tag{5}$$

with view-specific energies

$$e^{(\ell)}_{cu} = \|A^{(\ell)}_{cu:}\|_2^2, \qquad \ell \in \{1, 2\}$$

With energy floor $\delta = 10^{-8}$, the joint weight is

$$w_{cu} = \sqrt{(e^{(1)}_{cu} + \epsilon)(e^{(2)}_{cu} + \epsilon)}\, \mathbf{1}\left\{e^{(1)}_{cu} > \delta,\ e^{(2)}_{cu} > \delta\right\} \tag{6}$$

A channel has measured shared support when $\sum_u w_{cu} \geq \delta$. When measured support exists,

$$\pi_c = \mathrm{clip}_{[10^{-4},\, 1-10^{-4}]}\left(\frac{\sum_u w_{cu} a_{cu}}{\max(\sum_u w_{cu}, \epsilon)}\right) \tag{7}$$

The channel-RMS branch is

$$q^{(\ell)}_c = \sqrt{\frac{1}{UV}\sum_{u,v}\left(A^{(\ell)}_{cuv}\right)^2}, \qquad \ell \in \{1, 2\} \tag{8}$$

$$d_c = \frac{|q^{(1)}_c - q^{(2)}_c|}{q^{(1)}_c + q^{(2)}_c + \epsilon} \tag{9}$$

$$\rho_c = \mathrm{clip}_{[0,20]}\left(\frac{d_c}{\tau + \epsilon}\right) \tag{10}$$

$$n_c = \mathrm{clip}_{[10^{-4},\, 1-10^{-4}]}\left(e^{-\rho_c}\right) \tag{11}$$

The scale $\tau$ is a train-only exponential moving average of the minibatch-mean discrepancy. Restoring the example index for this update, define

$$\widehat{\tau} = \max\left(\frac{1}{BC}\sum_{i=1}^{B}\sum_{c=1}^{C} d_{ic}, 10^{-4}\right)$$

With EMA momentum $0.99$, training initializes the stored scale to $\widehat{\tau}$ on its first update and subsequently applies

$$\tau \leftarrow 0.99\,\tau + 0.01\,\widehat{\tau}$$

The stored value is lower-bounded by $10^{-4}$ when used. Evaluation does not update this state; if an uninitialized model is evaluated directly, the implementation falls back to $\widehat{\tau}$.

The coefficient $\lambda = 0.2$ is held fixed in all principal reference-dependent runs. It is treated as a nuisance parameter of this fixed reference generator and is not selected per dataset, seed, or final-evaluation result. Define

$$r_c = \lambda\pi_c^\star + (1-\lambda)n_c, \qquad \pi_c^\star = \begin{cases}\pi_c, & \text{measured shared support,}\\ n_c, & \text{otherwise.}\end{cases} \tag{12}$$

Thus, when directional agreement has no measured shared support, the routing construction falls back to the RMS-based stability value $n_c$ rather than using an unsupported directional estimate.

The equations above are applied independently to each example. Restoring the suppressed example index $i$ gives

$$r_i = (r_{i1}, \ldots, r_{iC}) \in [0,1]^C$$

and stacking the minibatch yields

$$R = [r_{ic}] \in [0,1]^{B\times C}$$

consistent with the notation in Section 3.2. The reference field is detached before it is used to construct a formation gate. In vision, the same computation is applied to the two spatial feature-map axes as a strict transfer; we do not interpret those axes as having audio semantics.

# B Proofs and Algebraic Details

## B.1 Mean/contrast decomposition

Define

$$W_0 = \frac{W^{(1)} + W^{(2)}}{2}, \qquad \Delta_W = \frac{W^{(1)} - W^{(2)}}{2} \tag{13}$$

Substituting $W^{(1)} = W_0 + \Delta_W$ and $W^{(2)} = W_0 - \Delta_W$ into Eq. (1) gives the exact identity

$$f(z; g) = W_0 z + \Delta_W((2g - \mathbf{1}) \odot z) + b \tag{14}$$

Thus stochastic formation acts through the expert-contrast term $\Delta_W$; this is an algebraic statement, not a semantic interpretation of the experts. For FixedHalf, $g = \frac{1}{2}\mathbf{1}$ and therefore $f(z; g) = W_0 z + b$. The forward predictor is dense-equivalent even though $W^{(1)}$ and $W^{(2)}$ remain separately parameterized during optimization.

## B.2 Beta moment inversion

**Lemma 1.** *If $0 < m < 1$, $0 < v < m(1-m)$, $\kappa = m(1-m)/v - 1$, $\alpha = m\kappa$, and $\beta = (1-m)\kappa$, then $G \sim \mathrm{Beta}(\alpha, \beta)$ satisfies $\mathbb{E}G = m$ and $\mathrm{Var}(G) = v$.*

*Proof.* For a Beta variable, $\mathbb{E}G = \alpha/(\alpha+\beta)$ and $\mathrm{Var}(G) = \alpha\beta/[(\alpha+\beta)^2(\alpha+\beta+1)]$. Since $\alpha + \beta = \kappa$, the mean is $m$, and

$$\mathrm{Var}(G) = \frac{m(1-m)}{\kappa+1} = v \tag{15}$$

□

### B.3 RETROSPECTIVE FIXEDBETA TRACE AGGREGATION

Let $(m_t, v_t)$ denote the MomentBeta minibatch moments recorded over the same-seed training trace, $t = 1, \dots, T$. Retrospective FixedBeta uses

$$\bar{m} = \frac{1}{T}\sum_{t=1}^{T} m_t, \qquad \bar{v} = \frac{1}{T}\sum_{t=1}^{T}(v_t + m_t^2) - \bar{m}^2$$

A separate checkpoint is then trained using the fixed Beta law whose mean and variance are $(\bar{m}, \bar{v})$. Thus RetroFixedBeta uses no online reference field during its own optimization, although its distribution remains derived from the same-seed MomentBeta training trace.

### B.4 MOMENTBETA ERASES REFERENCE-FIELD ASSIGNMENT IDENTITY

**Lemma 2.** *Let $R, R' \in [0,1]^{B\times C}$ have the same empirical mean and variance. The conditional MomentBeta gate distributions are identical: $G \mid R \overset{d}{=} G' \mid R'$.*

*Proof.* MomentBeta maps $R$ to Beta parameters only through $(m(R), v(R))$. Equal moments produce equal $(\alpha, \beta)$, and the $BC$ generated entries are iid from that same law. Therefore their joint conditional distributions coincide. □

This establishes invariance of the generator to sample/channel permutation, rank order, and individual values beyond information contained in $(m, v)$.

### B.5 GLOBALPOOL MULTISET AND COVARIANCE

Let $s = (s_1, \dots, s_M)$ and $G_j = s_{\Pi(j)}$ for a uniformly random permutation $\Pi$. Define

$$\bar{s} = M^{-1}\sum_i s_i, \qquad v_s = M^{-1}\sum_i (s_i - \bar{s})^2$$

GlobalPool preserves the multiset exactly, and for each $j$, $\mathbb{E}[G_j \mid s] = \bar{s}$ and $\mathrm{Var}(G_j \mid s) = v_s$. For $j \neq k$, exchangeability gives a common covariance $c$. Since $\sum_j G_j = \sum_j s_j$ deterministically,

$$0 = \mathrm{Var}\left(\sum_j G_j \mid s\right) = Mv_s + M(M-1)c$$

so

$$\mathrm{Cov}(G_j, G_k \mid s) = -\frac{v_s}{M-1}$$

Hence GlobalPool and MomentBeta differ not only in retained marginal information but also in their dependence structure.

### B.6 CONDITIONAL STOCHASTIC LOGIT MOMENTS

Fix one representation $z \in \mathbb{R}^C$ and let $g = (g_1, \dots, g_C) \in [0,1]^C$ be a random channel gate whose coordinates are conditionally independent. Define the conditional coordinate-wise moments

$$\boldsymbol{m} = (m_1, \dots, m_C), \qquad \boldsymbol{v} = (v_1, \dots, v_C)$$

where

$$\mathbb{E}[g_c \mid z, \boldsymbol{m}, \boldsymbol{v}] = m_c, \qquad \mathrm{Var}(g_c \mid z, \boldsymbol{m}, \boldsymbol{v}) = v_c$$

Using the mean/contrast decomposition from Section B.1,

$$\mathbb{E}[f(z;g) \mid z, \boldsymbol{m}, \boldsymbol{v}] = W_0 z + \Delta_W\left((2\boldsymbol{m} - \mathbf{1}) \odot z\right) + b$$

and

$$\mathrm{Cov}(f(z;g) \mid z, \boldsymbol{m}, \boldsymbol{v}) = 4\Delta_W\,\mathrm{Diag}(\boldsymbol{v} \odot (z \odot z))\,\Delta_W^\top$$

For iid MomentBeta coordinates with common conditional mean $m_0$ and variance $v_0$,

$$\mathbb{E}[f(z;g) \mid z, m_0, v_0] = W_0 z + (2m_0 - 1)\Delta_W z + b$$

and

$$\mathrm{Cov}(f(z;g) \mid z, m_0, v_0) = 4v_0 \Delta_W \,\mathrm{Diag}(z \odot z)\Delta_W^\top$$

For the squared-logit surrogate,

$$\mathbb{E}\left[\frac{1}{2}\|f - y\|_2^2 \mid z, m_0, v_0\right] = \frac{1}{2}\,\|\mathbb{E}[f \mid z, m_0, v_0] - y\|_2^2 + 2v_0\,\|\Delta_W \,\mathrm{Diag}(z)\|_F^2$$

This last identity is exact for the squared-logit surrogate, not for cross-entropy. A local Hessian expansion can provide an interpretation for smooth losses but is not used as an exact training theorem.

### B.7 Nested static deployment families

Let

$$\begin{aligned}\mathcal{M}_{1/2} &= \{\tfrac{1}{2}\mathbf{1}\}\\ \mathcal{M}_{\text{scalar}} &= \{\mathrm{sigmoid}(a)\mathbf{1} : a \in \mathbb{R}\}\\ \mathcal{M}_{\text{channel}} &= \{\mathrm{sigmoid}(a) : a \in \mathbb{R}^C\}\end{aligned}$$

Then

$$\mathcal{M}_{1/2} \subseteq \mathcal{M}_{\text{scalar}} \subseteq \mathcal{M}_{\text{channel}}$$

Hence, for any fixed checkpoint and static-fitting objective $\mathcal{L}_{\text{fit}}(\mu)$,

$$\inf_{\mu \in \mathcal{M}_{\text{channel}}} \mathcal{L}_{\text{fit}}(\mu) \le \inf_{\mu \in \mathcal{M}_{\text{scalar}}} \mathcal{L}_{\text{fit}}(\mu) \le \inf_{\mu \in \mathcal{M}_{1/2}} \mathcal{L}_{\text{fit}}(\mu)$$

This follows only from nesting of the feasible deployment families. It does not imply that optimization reaches these infima, nor does it imply monotone validation or test performance.

### B.8 Exact static compilation

**Theorem 1** (Exact static compilation)**.** *For any fixed $\mu \in [0,1]^C$, define*

$$W_\mu = W^{(1)}\,\mathrm{Diag}(\mu) + W^{(2)}\,\mathrm{Diag}(\mathbf{1} - \mu)$$

*Then*

$$f_\mu(z) = W_\mu z + b$$

*for every $z$.*

*Proof.* Using $\mathrm{Diag}(\mu)z = \mu \odot z$ and $\mathrm{Diag}(\mathbf{1} - \mu)z = (\mathbf{1} - \mu) \odot z$ in Eq. (4) gives

$$\begin{aligned}f_\mu(z) &= W^{(1)}\,\mathrm{Diag}(\mu)z + W^{(2)}\,\mathrm{Diag}(\mathbf{1} - \mu)z + b\\ &= W_\mu z + b\end{aligned}$$

□

### B.9 $N$-expert simplex composition and static compilation

For the reference-free $N$-expert extension, let $W^{(e)} \in \mathbb{R}^{K \times C}$, $e \in \{1, \dots, N\}$, be bias-free expert matrices sharing one bias $b \in \mathbb{R}^K$. For example $i$ and channel $c$, define the expert-composition vector

$$\boldsymbol{\gamma}_{ic} = (\gamma_{i1c}, \dots, \gamma_{iNc}) \in \Delta^{N-1}, \qquad \gamma_{iec} \ge 0, \qquad \sum_{e=1}^{N} \gamma_{iec} = 1$$

For expert $e$, collect its channel-wise coefficients as

$$\boldsymbol{\gamma}_{ie} = (\gamma_{ie1}, \dots, \gamma_{ieC}) \in [0,1]^C$$

The resulting classifier is

$$f(z_i;\Gamma_i)=\sum_{e=1}^{N}W^{(e)}\big(\boldsymbol{\gamma}_{ie}\odot z_i\big)+b,\qquad \Gamma_i=\{\boldsymbol{\gamma}_{ic}\}_{c=1}^{C}$$

FixedEqual formation sets

$$\gamma_{iec}=\frac{1}{N}$$

for every example, expert, and channel. Symmetric Dirichlet formation samples independently

$$\boldsymbol{\gamma}_{ic}\sim\mathrm{Dirichlet}(\mathbf{1}_N)$$

For static deployment, let

$$\mu=[\mu_{ec}]\in[0,1]^{N\times C},\qquad \mu_{ec}\geq 0,\qquad \sum_{e=1}^{N}\mu_{ec}=1\quad\text{for every } c$$

FixedEqual deployment uses $\mu_{ec}=1/N$. Learned global-simplex deployment uses one $\pi\in\Delta^{N-1}$ shared across channels, so $\mu_{ec}=\pi_e$. Learned channel-simplex deployment permits a separate $\mu_{\cdot c}\in\Delta^{N-1}$ for each channel.

For expert $e$, define

$$\boldsymbol{\mu}_e=(\mu_{e1},\ldots,\mu_{eC})\in[0,1]^C$$

A fixed deployment therefore gives

$$f_\mu(z)=\sum_{e=1}^{N}W^{(e)}\big(\boldsymbol{\mu}_e\odot z\big)+b$$

Define the compiled matrix $W_\mu\in\mathbb{R}^{K\times C}$ column-wise by

$$W_\mu[:,c]=\sum_{e=1}^{N}\mu_{ec}W^{(e)}[:,c]$$

Then, for every $z\in\mathbb{R}^C$,

$$f_\mu(z)=W_\mu z+b$$

Hence every fixed $N$-expert simplex composition compiles exactly into a single affine classifier.

## C Detailed Experimental Protocol

**Shared architecture.** The controlled network uses three Conv2d–ReLU blocks with pooling after the first two blocks and global spatial averaging before classification. The representation width is $C=128$. There is no batch normalization. The standard two-expert head has two bias-free $K\times C$ matrices and one shared $K$-dimensional bias. Dense replaces that head by one $K\times C$ affine classifier. Thus all controlled methods share the backbone, while classifier parameter counts differ by design. Excluding the shared backbone, the Dense head has $K(C+1)$ trainable parameters and the standard two-expert head has $K(2C+1)$. LearnedMoE adds a two-way affine router with $2(C+1)$ parameters. The SMEAR-style hidden width is chosen algorithmically to approximate the standard two-expert head+router budget; exact values are saved in each run's checkpoint metadata.

**ESC-50.** We use the five official folds. For each held-out fold, the other four folds (1600 clips) are training data and the held-out 400 clips are final test data. The final neural schedule is fixed at 100 epochs/final checkpoint; static composition uses its predeclared 30-epoch final-iterate schedule. No outer validation/test selection is performed. Five scientific seeds are run for each fold. Robust-B averages the 20 Mode-B conditions formed by five waveform-domain families (Gaussian noise, synthetic reverberation, band-limiting, clipping, quantization) crossed with severities $0.25, 0.50, 0.75, 1.00$. Mode B independently corrupts both views with the same family/severity. Clean accuracy is also reported.

**Speech Commands V2.** We use the official 35-class fixed split: 84,843 train, 9,981 validation, and 11,005 test examples. Five scientific seeds are run. Clean top-1 accuracy is primary and Robust-B is secondary. Neural checkpoint selection uses the official validation split and the predeclared matched validation score: mean accuracy over clean data and the five corruption families at severity $0.60$ in Mode B, with mean validation loss breaking ties within the selection tolerance. For learned static deployment, the neural checkpoint is frozen and the composition parameters are optimized on a deterministic heterogeneous training cache. Static selection uses a frozen validation cache containing clean data and the same five corruption families at severity $0.60$ in Mode B; validation cross-entropy is the primary selection criterion, with validation accuracy breaking ties. The official test split is used only after neural and static selection.

**FSD50K.** We use the official development train/validation partition and reserve the official evaluation split for final evaluation. Targets are multi-hot and training uses binary cross-entropy. Neural checkpoints are selected by clean validation mAP. For learned static deployment, only the composition parameters are optimized on frozen clean training representations using BCE; the iterate with the highest clean validation mAP is retained. The official evaluation split is not used for neural or static selection. Final official-evaluation mAP is averaged over five scientific seeds.

**CIFAR-100.** The official 50k training set is split with a fixed stratified seed into 45k train and 5k validation (450/50 per class). The official 10k test set is used only after all neural/static selection. Training uses RandomCrop(32,padding=4) and RandomHorizontalFlip. Augmentation and sample order are deterministic functions of scientific seed, epoch, sample index, and view. Five seeds are run.

**Tiny ImageNet-200.** From the official 100k training set, 50 images/class are held out, producing 90k train and 10k internal validation. The labeled official 10k validation set is reserved as final evaluation. Training uses RandomCrop(64,padding=4) and RandomHorizontalFlip under the same deterministic two-view contract as CIFAR. Five seeds are run.

**Checkpoint selection.** For CIFAR/Tiny formation controls, the neural checkpoint is selected by clean validation accuracy under the same router-free $0.5$ two-expert validation path, with validation cross-entropy breaking ties within the predeclared tolerance. Dense uses native dense validation; LearnedMoE uses native learned-router validation; THOR-style uses deterministic ensemble validation; Shake-Shake-style uses its $0.5$ merge; SMEAR-style uses native parameter-merged validation. Learned static $\mu$ is optimized on frozen clean training features and selected by the same accuracy-first, cross-entropy-tie-break rule on internal validation before the final held-out evaluation split is touched.

**Optimization.** The principal two-expert audio runs use AdamW with learning rate $10^{-3}$, weight decay $10^{-4}$, batch size 64, and gradient clipping at 5. ESC-50 uses its fixed 100-epoch final-checkpoint schedule. Speech Commands V2 and FSD50K train for up to 100 epochs with patience 20 and minimum improvement $10^{-4}$. Audio static composition is optimized with Adam at learning rate $0.05$ for up to 30 epochs. Speech Commands and FSD50K use patience 6 with their dataset-specific validation criteria, whereas ESC-50 follows its predeclared 30-epoch final-iterate schedule without outer validation selection. CIFAR-100 and Tiny ImageNet use AdamW with learning rate $10^{-3}$, weight decay $10^{-4}$, batch size 256, gradient clipping at 5, and up to 100 epochs with patience 15 and minimum improvement $10^{-4}$. Their static compositions are optimized with Adam at learning rate $10^{-2}$ for up to 100 epochs with patience 15.

**Comparator implementation.** LearnedMoE routes each augmented view independently with a learned two-way soft router and a small predeclared load-balance penalty. THOR-style uses per-example random expert dispatch, two stochastic passes, and symmetric KL consistency, then reports expected dispatch, deterministic ensemble, and static deployment. Shake-Shake-style uses independent forward/backward affine branch coefficients, with 0.5 canonical and matched-static evaluation. SMEAR-style uses a two-way router to merge expert parameters per example before a nonlinearity; hidden width is chosen to approximately match the standard two-expert head+router parameter budget. These are mechanism adaptations, not exact replications of the original architectures.

**$N$-expert stress-test protocol.** The reference-free expert-count study uses the same dataset splits and primary metrics as the principal experiments. We evaluate $N \in \{2, 4, 8\}$ under FixedEqual and

symmetric $\mathrm{Dirichlet}(\mathbf{1}_N)$ formation with five scientific seeds; ESC-50 additionally evaluates all five official held-out folds. For every trained checkpoint, the neural parameters are frozen before evaluating FixedEqual, learned global-simplex, and learned channel-simplex static deployment. The $N = 2$ checkpoints are rerun within the same $N$-expert implementation and are the only matched anchors used for paired $N = 4$ and $N = 8$ effects; principal two-expert checkpoints are not substituted.

The training head contains $N$ bias-free $K \times C$ expert matrices and one shared $K$-dimensional bias, for

$$K(NC + 1)$$

trainable head parameters. After any fixed simplex deployment is selected, its exactly compiled classifier contains

$$K(C + 1)$$

parameters, independent of $N$. Consequently, this experiment varies training-head capacity with expert count and is not a parameter-matched training-capacity comparison.

**Reference-strength sensitivity protocol.** The main reference-dependent experiment always uses $\lambda = 0.2$. For sensitivity only, GlobalPool, MomentBeta, and Retrospective FixedBeta are rerun at $\lambda \in \{0.05, 0.10, 0.50, 1.00\}$ using five scientific seeds and the same final evaluation protocol; ESC uses all five official folds. Retrospective FixedBeta is recalibrated from the matched same-seed MomentBeta trace at the corresponding $\lambda$. These runs are not used to choose the main operating point. For the post-hoc channel-assignment audit, each already fitted MomentBeta channel vector is evaluated together with 100 permutations generated from fixed random seeds. Every permutation preserves the exact finite coefficient multiset and changes only coefficient-to-channel correspondence. Within a replicate, the aligned vector and all permutations use identical frozen representations and evaluation randomness; no backbone, expert, learned coefficient, or probe is refit. ESC-50 effects are averaged across scientific seeds within each official fold before fold-level summarization; the other datasets use their five scientific seeds. The audit is exploratory and is not used for model, checkpoint, or hyperparameter selection.

# D Complete Five-Dataset Results

This appendix reports the complete principal five-dataset results underlying Table 2. Figure 2 the formation controls under learned channel-wise static deployment relative to Dense, while the following tables report the complete per-dataset results using the metrics and aggregation rules defined in Section 4. ESC-50 additionally reports clean accuracy, and Speech Commands additionally reports Robust-B as a secondary metric. Tables 6–10 report the complete per-dataset results.

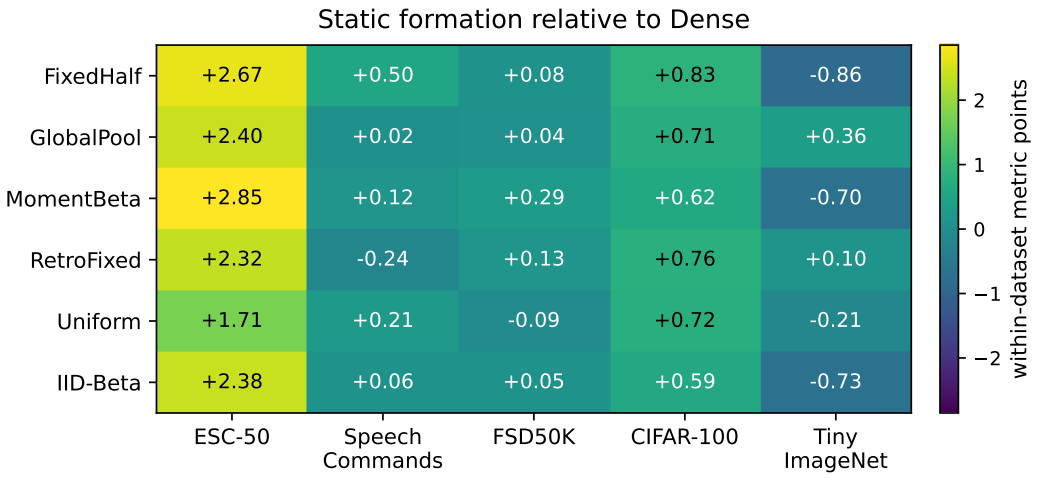


Figure 2: Formation-control performance under learned channel-wise static deployment varies across regimes. Each cell is the method's mean metric minus Dense on the corresponding dataset (ESC: Robust-B accuracy; Speech/CIFAR/Tiny: clean accuracy; FSD50K: mAP). Values are within-dataset points; magnitudes should not be compared as if the five metrics were identical.

Table 6: ESC-50 final 4-train/1-test protocol. Values are mean ± standard deviation across the five official test-fold means, where each fold mean is first averaged over five scientific seeds.

| ESC-50 method | Robust-B acc. | Clean acc. |
|---|---|---|
| MomentBeta-Static | $64.566 \pm 1.687$ | $68.070 \pm 1.308$ |
| ReferenceGate-Static | $64.477 \pm 2.359$ | $67.950 \pm 2.027$ |
| FixedHalf-Static | $64.379 \pm 1.607$ | $67.760 \pm 1.148$ |
| GlobalPool-Static | $64.112 \pm 1.924$ | $67.650 \pm 1.860$ |
| IID-Beta(1,0.6)-Static | $64.091 \pm 2.035$ | $67.560 \pm 1.735$ |
| RetrospectiveFixedBeta-Static | $64.032 \pm 1.623$ | $67.190 \pm 1.417$ |
| Uniform-Static | $63.418 \pm 1.326$ | $66.930 \pm 1.079$ |
| THOR-style-Static | $61.781 \pm 1.598$ | $65.480 \pm 1.582$ |
| Dense | $61.712 \pm 1.844$ | $64.990 \pm 1.872$ |
| ReferenceGate-Native | $60.075 \pm 1.972$ | $63.280 \pm 2.141$ |
| LearnedMoE-Native | $59.392 \pm 2.285$ | $62.580 \pm 1.996$ |
| THOR-style-Ensemble | $58.961 \pm 1.676$ | $62.390 \pm 2.017$ |
| THOR-style-ExpectedDispatch | $58.573 \pm 1.757$ | $61.870 \pm 2.088$ |
| LearnedMoE-Static | $53.129 \pm 1.626$ | $56.280 \pm 1.496$ |
| Shake-Shake-style-matched-static | $52.799 \pm 1.872$ | $56.240 \pm 2.054$ |
| SMEAR-style-Native | $50.872 \pm 2.549$ | $53.490 \pm 2.419$ |
| Shake-Shake-style-canonical-0.5 | $46.839 \pm 2.326$ | $49.680 \pm 2.614$ |

Table 7: Speech Commands V2 official fixed-split results over five scientific seeds. Values are mean ± seed standard deviation.

| Speech Commands V2 method | Clean acc. | Robust-B acc. |
|---|---|---|
| FixedHalf-Static | $89.903 \pm 0.384$ | $87.423 \pm 0.410$ |
| Uniform-Static | $89.617 \pm 0.728$ | $87.108 \pm 0.785$ |
| MomentBeta-Static | $89.525 \pm 0.551$ | $87.019 \pm 0.586$ |
| IID-Beta-Static | $89.463 \pm 0.770$ | $87.034 \pm 1.041$ |
| LearnedMoE-Static | $89.458 \pm 0.478$ | $87.137 \pm 0.511$ |
| GlobalPool-Static | $89.423 \pm 0.309$ | $86.924 \pm 0.418$ |
| Dense | $89.405 \pm 0.678$ | $87.001 \pm 0.704$ |
| RetrospectiveFixedBeta-Static | $89.169 \pm 0.764$ | $86.820 \pm 0.979$ |
| THOR-style-Static | $89.010 \pm 0.455$ | $86.448 \pm 0.644$ |
| THOR-style-Ensemble | $88.858 \pm 0.538$ | $86.248 \pm 0.707$ |
| ReferenceGate-Static | $88.818 \pm 0.383$ | $86.292 \pm 0.499$ |
| THOR-style-ExpectedDispatch | $88.814 \pm 0.526$ | $86.215 \pm 0.702$ |
| LearnedMoE-Native | $88.798 \pm 0.385$ | $86.143 \pm 0.305$ |
| ReferenceGate-Native | $87.319 \pm 0.608$ | $84.608 \pm 0.724$ |
| SMEAR-style-Native | $85.767 \pm 2.172$ | $83.049 \pm 2.318$ |
| Shake-Shake-style-matched-static | $84.792 \pm 1.488$ | $81.919 \pm 1.476$ |
| Shake-Shake-style-canonical-0.5 | $83.095 \pm 1.293$ | $80.365 \pm 1.294$ |

Table 8: FSD50K five-seed multi-label results. Values are mean ± seed standard deviation in mAP points.

| FSD50K method | mAP |
|---|---|
| SMEAR-style-Native | $35.783 \pm 0.556$ |
| MomentBeta-Static | $34.141 \pm 0.461$ |
| RetrospectiveFixedBeta-Static | $33.978 \pm 0.428$ |
| LearnedMoE-Native | $33.933 \pm 0.273$ |
| FixedHalf-Static | $33.929 \pm 0.320$ |
| IID-Beta-Static | $33.895 \pm 0.327$ |
| GlobalPool-Static | $33.886 \pm 0.493$ |
| Dense | $33.847 \pm 0.368$ |
| Uniform-Static | $33.755 \pm 0.384$ |
| ReferenceGate-Static | $33.697 \pm 0.727$ |
| THOR-style-Static | $33.618 \pm 0.544$ |
| THOR-style-Ensemble | $33.604 \pm 0.550$ |
| THOR-style-ExpectedDispatch | $33.596 \pm 0.549$ |
| ReferenceGate-Native | $33.312 \pm 0.661$ |
| LearnedMoE-Static | $27.017 \pm 0.942$ |
| Shake-Shake-style-matched-static | $17.654 \pm 1.991$ |
| Shake-Shake-style-canonical-0.5 | $17.607 \pm 1.857$ |

Table 9: CIFAR-100 five-seed official-test accuracy. Values are mean ± seed standard deviation in percentage.

| CIFAR-100 method | Test acc. |
|---|---|
| FixedHalf-Static | $51.772 \pm 0.813$ |
| RetrospectiveFixedBeta-Static | $51.704 \pm 1.084$ |
| Uniform-Static | $51.664 \pm 0.742$ |
| GlobalPool-Static | $51.648 \pm 0.520$ |
| MomentBeta-Static | $51.558 \pm 0.821$ |
| IID-Beta(1,0.6)-Static | $51.528 \pm 0.771$ |
| LearnedMoE-Native | $51.010 \pm 0.463$ |
| Dense | $50.940 \pm 1.266$ |
| THOR-style-Static | $49.516 \pm 1.467$ |
| THOR-style-Ensemble | $49.364 \pm 1.447$ |
| THOR-style-ExpectedDispatch | $49.361 \pm 1.496$ |
| LearnedMoE-Static | $47.918 \pm 0.688$ |
| SMEAR-style-Native | $46.590 \pm 1.257$ |
| Shake-Shake-style-matched-static | $40.088 \pm 1.096$ |
| Shake-Shake-style-canonical-0.5 | $39.220 \pm 1.172$ |

Table 10: Tiny ImageNet-200 five-seed held-out final-evaluation accuracy. Values are mean ± seed standard deviation in percentage.

| Tiny ImageNet-200 method | Final acc. |
|---|---|
| GlobalPool-Static | $34.792 \pm 0.429$ |
| RetrospectiveFixedBeta-Static | $34.534 \pm 1.299$ |
| Dense | $34.436 \pm 0.553$ |
| Uniform-Static | $34.228 \pm 0.446$ |
| LearnedMoE-Native | $33.912 \pm 0.763$ |
| MomentBeta-Static | $33.734 \pm 1.278$ |
| IID-Beta(1,0.6)-Static | $33.704 \pm 0.851$ |
| FixedHalf-Static | $33.578 \pm 0.827$ |
| SMEAR-style-Native | $32.366 \pm 1.074$ |
| THOR-style-Static | $31.408 \pm 1.209$ |
| THOR-style-Ensemble | $31.376 \pm 1.210$ |
| THOR-style-ExpectedDispatch | $31.332 \pm 1.227$ |
| LearnedMoE-Static | $30.212 \pm 1.150$ |
| Shake-Shake-style-matched-static | $24.572 \pm 1.769$ |
| Shake-Shake-style-canonical-0.5 | $24.324 \pm 1.965$ |

# E Paired Deployment-Mechanism Analyses

## E.1 ESC-50 paired comparisons

The final ESC-50 protocol yields 25 paired seed–fold cells. MomentBeta-Static exceeds Dense by 2.8545 Robust-B points on average, wins all five fold means and 24/25 cells, with clean accuracy $+3.08$ points and CE $-0.16715$. Relative to the ReferenceGate-Native model, the mean Robust-B difference is $+4.4915$ points with 25/25 cell wins. Relative to the ReferenceGate-Static model the difference is only $+0.0895$ points (13/25 cells), so those methods are a tie at the resolution of this experiment.

Among formation controls, MomentBeta versus FixedHalf is $+0.187$ points (14/25 cells) and versus GlobalPool $+0.4545$ points (15/25); neither is decisive. MomentBeta versus Uniform is the clearest reference-dependent versus reference-free contrast: $+1.1485$ points, all five fold means, 17/25 cells, and a descriptive fold-level 95% interval $[0.1246, 2.1724]$. Retrospective FixedBeta trails by $0.5345$ points (18/25 cells), suggestive but not decisive. Uniform itself trails FixedHalf by $0.9615$ points in the same matched analysis, showing that generic stochasticity is not sufficient as a universal explanation.

### E.2 CROSS-DATASET CHANNEL-ASSIGNMENT AUDIT

The learned channel composition raises a separate deployment question: whether performance depends on the specific correspondence between learned coefficients and representation channels, rather than only on channel-wise fitting freedom or on the marginal distribution of the learned coefficients. We test this post hoc on the MomentBeta checkpoint families used for the mechanism audit. For each replicate, we retain the exact finite learned multiset $\{\mu_c\}_{c=1}^{C}$ and evaluate 100 deterministic random permutations that alter only its correspondence with representation channels. No backbone, expert weight, learned coefficient value, or probe is refit. Within each replicate, the aligned vector and all permutations are evaluated under identical dataset-specific replay conditions.

Table 11 separates two questions. The channel–scalar comparison asks whether channel-wise fitting is useful at all; the assignment intervention asks whether the particular learned coefficient-to-channel correspondence matters after the coefficient values themselves are held fixed.

Table 11: Post-hoc deployment-mechanism diagnostics for MomentBeta checkpoint families. Channel–scalar compares learned channel-wise and learned scalar static composition. Aligned–permuted preserves the exact learned coefficient multiset and changes only coefficient-to-channel correspondence; each reported permutation comparator is the mean over 100 permutations of the same replicate. Effects are within-dataset metric points with descriptive 95% $t$-intervals in brackets. ESC-50 uses Robust-B accuracy and fold-first aggregation; Speech Commands, CIFAR-100, and Tiny ImageNet use clean accuracy over five scientific seeds; FSD50K uses official-evaluation mAP over five seeds. The final column reports $\mathbb{E}\|\Delta_W((2\mu-1)\odot z)\|_2/\mathbb{E}\|W_0 z\|_2$. Metric-point magnitudes are not compared across datasets.

| Dataset | Channel − scalar | Aligned − permuted | Functional contrast |
|---|---|---|---|
| ESC-50 | $+2.056$ [1.528, 2.583] | $+2.901$ [2.363, 3.438] | 0.0548 |
| Speech Commands | $+0.300$ [0.145, 0.455] | $+0.464$ [0.290, 0.639] | 0.0272 |
| FSD50K | $-0.071$ [$-0.159$, 0.016] | $-0.012$ [$-0.098$, 0.075] | 0.0287 |
| CIFAR-100 | $+0.276$ [$-0.032$, 0.584] | $+0.415$ [0.106, 0.724] | 0.0253 |
| Tiny ImageNet | $+0.188$ [$-0.032$, 0.408] | $+0.276$ [0.125, 0.426] | 0.0342 |

The assignment intervention gives a sharper result than the channel–scalar comparison alone. On ESC-50, the aligned assignment exceeds its corresponding permutation mean by $2.901$ points, with positive differences in all 25 seed–fold cells and all five official fold aggregates. Speech Commands, CIFAR-100, and Tiny ImageNet show smaller but consistent aligned-assignment advantages of $+0.464$, $+0.415$, and $+0.276$ points, respectively, with positive differences in all five scientific seeds for each dataset. In contrast, FSD50K shows essentially no assignment advantage ($-0.012$ points; 2/5 positive seeds), and its descriptive interval spans zero.

This distinction is particularly informative on the vision datasets. On CIFAR-100 and Tiny ImageNet, the channel–scalar intervals include zero, whereas the aligned–permuted intervals are entirely positive. Thus the assignment intervention identifies sensitivity to coefficient-to-channel correspondence that is not captured by comparing a $C$-coefficient composition with a one-coefficient scalar composition. Conversely, FSD50K shows neither a clear channel–scalar advantage nor detectable sensitivity to coefficient-to-channel assignment. The performance relevance of channel identity is therefore regime dependent rather than an automatic consequence of channel-wise parameterization.

**Functional expert contrast.** To avoid overloading the deployment symbol $D$, define

$$W_0 = \frac{W^{(1)} + W^{(2)}}{2}, \qquad \Delta_W = \frac{W^{(1)} - W^{(2)}}{2}$$

The contrast expressed by the learned deployment is summarized by

$$\rho_\Delta = \frac{\mathbb{E}_z \left\| \Delta_W \big( (2\mu - 1) \odot z \big) \right\|_2}{\mathbb{E}_z \left\| W_0 z \right\|_2}$$

This quantity is descriptive rather than an intervention. In particular, Speech Commands and FSD50K have similar mean values ($0.0272$ and $0.0287$) despite showing clearly different assignment

sensitivity. A nonzero expert-contrast magnitude is therefore not sufficient to establish that a specific coefficient-to-channel correspondence is useful; the direct assignment intervention provides the stronger mechanism test.

**Unrestricted-probe provenance.** For transparency, we also report the primary-metric channel–probe comparisons. On ESC-50 and Speech Commands, where the saved probe artifacts are directly verified, the channel-minus-probe differences are $+0.454$ Robust-B points $[-0.317,\ 1.225]$ and $+0.222$ clean-accuracy points $[-0.007,\ 0.451]$, respectively. The ESC difference is positive in 3/5 official fold means (16/25 seed–fold cells), and the Speech difference is positive in 4/5 seeds. For the recorded-only historical probes, FSD50K gives $+3.609$ mAP points $[3.230,\ 3.988]$, whereas CIFAR-100 and Tiny ImageNet give $-1.440$ $[-1.913,\ -0.967]$ and $-0.846$ $[-1.260,\ -0.432]$ clean-accuracy points, respectively. Because the probe weights were not retained for FSD50K, CIFAR-100, or Tiny ImageNet, these three are metric comparisons only and are not used to make artifact-level parameterization claims. The channel-assignment intervention, rather than probe superiority, is the direct mechanism test. Table 12 gives the detailed ESC-50 same-checkpoint deployment decomposition.

Table 12: MomentBeta ESC-50 deployment decomposition. The unrestricted probe is not decisively better than the compiled channel head; it is slightly lower in the aggregate. In the final column, "channel 25/25" indicates that the learned-channel deployment outperforms the listed comparator in all 25 paired seed–fold cells; cell and fold win counts are shown for the unrestricted probe.

| Deployment | Robust-B | Clean | vs. channel | paired note |
|---|---|---|---|---|
| learned channel | 64.566 | 68.070 | – | reference |
| learned scalar | 62.511 | 65.850 | $-2.056$ | channel 25/25 |
| $\mu = 0.5$ | 62.423 | 65.890 | $-2.143$ | channel 25/25 |
| $\mu = \bar{m}$ | 62.447 | 65.800 | $-2.119$ | channel 25/25 |
| $\mu = 0$ | 60.278 | 63.420 | $-4.289$ | – |
| $\mu = 1$ | 61.669 | 65.130 | $-2.898$ | – |
| unrestricted linear probe | 64.112 | 68.010 | $-0.454$ | 16/25; 3/5 folds |

## F Interpretation guide and claim boundaries

For clarity, the following are *not* claims of this paper: (i) MomentBeta is universally superior or necessary; (ii) random gating is sufficient; (iii) static deployment always beats dynamic/native routing; (iv) channel-wise static composition always beats parameter merging; (v) the fixed two-view reference field has universal semantic meaning; (vi) the controlled CNN vision results establish state-of-the-art vision MoE performance; (vii) increasing the number of experts monotonically improves performance; (viii) $N = 4$ or $N = 8$ is a universally optimal expert count; (ix) effective use of more experts necessarily predicts higher performance; (x) learned channel-simplex deployment is universally best, since FSD50K is a systematic counterexample; or (xi) the reference coefficient $\lambda$ has been tuned separately for each dataset; or (xii) coefficient-to-channel assignment is universally consequential across datasets.

The supported claim is narrower. Formation and deployment can be manipulated as separate experimental variables, and their interaction is empirically nontrivial. The information-retention controls show that the training-time mixing information required for a useful checkpoint varies by regime. Same-checkpoint deployment comparisons show that simple deterministic deployment can exploit some formed checkpoints well but fail badly for others. The reference-free $N$-expert extension further shows that expert multiplicity changes this compatibility structure rather than providing a universal scaling benefit. The exact compilation results give these observations a concrete deployment consequence for the studied channel-composed head.

## G Reference-Strength Sensitivity

The principal reference-dependent experiments use the pre-specified $\lambda = 0.20$ reference generator. To assess sensitivity, we additionally evaluate $\lambda \in \{0.05, 0.10, 0.50, 1.00\}$ for GlobalPool, MomentBeta,

and Retrospective FixedBeta using the same matched evaluation protocol. These alternative settings are not used for model selection. In Tables 13–17 bold $\lambda$ denotes the pre-specified principal setting; the remaining rows are sensitivity perturbations.

Table 13: ESC-50 Robust-B reference-strength sensitivity. Values are fold-aggregated percentages. Bold $\lambda$ denotes the pre-specified principal setting; the other values are sensitivity perturbations and are not used for model selection.

| $\lambda$ | RetroFixedBeta | GlobalPool | MomentBeta |
|---|---|---|---|
| 0.05 | 63.869 | 63.786 | 64.018 |
| 0.10 | 63.416 | 63.889 | 63.967 |
| **0.20** | 64.032 | 64.112 | 64.566 |
| 0.50 | 64.076 | 63.998 | 64.265 |
| 1.00 | 64.365 | 64.387 | 64.248 |

Table 14: FSD50K reference-strength sensitivity. Values are mAP on the 0–100 scale. Bold $\lambda$ denotes the pre-specified principal setting; the other values are sensitivity perturbations and are not used for model selection.

| $\lambda$ | RetroFixedBeta | GlobalPool | MomentBeta |
|---|---|---|---|
| 0.05 | 33.855 | 33.778 | 33.968 |
| 0.10 | 33.980 | 34.097 | 33.818 |
| **0.20** | 33.978 | 33.886 | 34.141 |
| 0.50 | 34.074 | 33.974 | 33.824 |
| 1.00 | 33.937 | 33.816 | 33.818 |

Table 15: Speech Commands V2 clean-accuracy reference-strength sensitivity. Values are percentages. Bold $\lambda$ denotes the pre-specified principal setting; the other values are sensitivity perturbations and are not used for model selection.

| $\lambda$ | RetroFixedBeta | GlobalPool | MomentBeta |
|---|---|---|---|
| 0.05 | 89.201 | 89.141 | 89.594 |
| 0.10 | 89.548 | 89.359 | 89.788 |
| **0.20** | 89.169 | 89.423 | 89.525 |
| 0.50 | 89.227 | 89.159 | 89.319 |
| 1.00 | 89.225 | 89.668 | 89.577 |

Table 16: CIFAR-100 clean-accuracy reference-strength sensitivity. Values are percentages. Bold $\lambda$ denotes the pre-specified principal setting; the other values are sensitivity perturbations and are not used for model selection.

| $\lambda$ | RetroFixedBeta | GlobalPool | MomentBeta |
|---|---|---|---|
| 0.05 | 51.946 | 51.180 | 50.862 |
| 0.10 | 51.568 | 52.022 | 50.910 |
| **0.20** | 51.704 | 51.648 | 51.558 |
| 0.50 | 51.392 | 51.422 | 52.340 |
| 1.00 | 51.430 | 51.618 | 51.856 |

Table 17: Tiny ImageNet clean-accuracy reference-strength sensitivity. Values are percentages. Bold $\lambda$ denotes the pre-specified principal setting; the other values are sensitivity perturbations and are not used for model selection.

| $\lambda$ | RetroFixedBeta | GlobalPool | MomentBeta |
|---|---|---|---|
| 0.05 | 34.062 | 34.110 | 33.462 |
| 0.10 | 34.482 | 34.496 | 34.256 |
| **0.20** | 34.534 | 34.792 | 33.734 |
| 0.50 | 34.246 | 34.524 | 33.826 |
| 1.00 | 33.730 | 34.494 | 34.294 |

Restricting attention to the four alternative sensitivity settings $\lambda \in \{0.05, 0.10, 0.50, 1.00\}$, no common coefficient or formation rule dominates across all five datasets. MomentBeta has the highest observed value in 9 of the 20 dataset–$\lambda$ cells, GlobalPool in 8, and RetroFixedBeta in 3. Observed sensitivity is heterogeneous rather than uniformly small: over the four alternative settings, MomentBeta spans about 0.30 points on ESC-50 and 0.15 mAP points on FSD50K, but 1.48 accuracy points on CIFAR-100. These ranges are descriptive and are not interpreted as formal evidence of invariance or instability. Most importantly, the sensitivity analysis is not a tuning procedure: the principal $\lambda = 0.20$ operating point was fixed before these comparisons.

# H Reference-free expert-count stress test

This appendix reports the complete $N \in \{2, 4, 8\}$ reference-free stress test. FixedEqual and Dirichlet denote *formation* rules; FixedEqual, learned global simplex, and learned channel simplex denote *deployment* rules. In particular, learned global simplex is not GlobalPool: GlobalPool is a reference-information formation intervention in the principal experiment.

For non-ESC datasets, paired differences are computed within scientific seed against the rerun $N = 2$ anchor. ESC uses five seeds on each of five official folds; 25 cell-wise direction counts are descriptive, while reported uncertainty follows the fold-first protocol. Because this is an exploratory factorial stress test, the tables should not be read as 60 independent confirmatory hypothesis tests.

## H.1 Complete formation–deployment matrix

In the tables below, Equal denotes FixedEqual deployment, Global denotes learned-global-simplex deployment, and Channel denotes learned-channel-simplex deployment. For a deployment matrix $\mu = [\mu_{ec}] \in [0, 1]^{N \times C}$, define the per-channel expert entropy

$$H_c = -\sum_{e=1}^{N} \mu_{ec} \log \mu_{ec}$$

The reported mean effective expert count is

$$N_{\text{eff}} = \frac{1}{C} \sum_{c=1}^{C} \exp(H_c),$$

with probabilities clamped below $10^{-12}$ before the logarithm for numerical stability. Thus $N_{\text{eff}} = 1$ when every channel uses a one-hot expert composition and $N_{\text{eff}} = N$ when every channel uses the uniform composition. The columns $N_{\text{eff}}^{G}$ and $N_{\text{eff}}^{C}$ report this quantity for learned-global-simplex and learned-channel-simplex deployment, respectively. Performance is reported on the same 0–100 primary-metric scale used in the main text. The complete dataset-specific deployment matrices are reported in Tables 18–22.

Table 18: ESC-50 complete $N$-expert deployment matrix. Entries are Robust-B accuracy mean $\pm$ reported standard deviation in percentage points. Following the main reporting protocol, the reported standard deviation is across the five official fold means after averaging scientific seeds within fold. Bold marks the highest observed mean in this matrix, not a claim of statistically optimal $N$.

| Formation | $N$ | Equal | Global | $N_{\text{eff}}^{G}$ | Channel | $N_{\text{eff}}^{C}$ |
|---|---|---|---|---|---|---|
| Dirichlet | 2 | $61.977 \pm 1.942$ | $61.957 \pm 1.853$ | 1.958 | $64.225 \pm 1.956$ | 1.399 |
| Dirichlet | 4 | $62.107 \pm 1.016$ | $62.163 \pm 0.848$ | 3.715 | $\mathbf{65.558 \pm 1.423}$ | 2.503 |
| Dirichlet | 8 | $60.679 \pm 1.788$ | $60.939 \pm 1.691$ | 7.057 | $64.216 \pm 1.864$ | 4.997 |
| FixedEqual | 2 | $60.868 \pm 1.432$ | $60.956 \pm 1.485$ | 1.994 | $64.674 \pm 1.632$ | 1.749 |
| FixedEqual | 4 | $60.302 \pm 1.512$ | $60.637 \pm 1.380$ | 3.956 | $65.144 \pm 1.874$ | 3.244 |
| FixedEqual | 8 | $60.315 \pm 1.201$ | $60.688 \pm 1.247$ | 7.630 | $64.994 \pm 1.265$ | 6.166 |

Table 19: FSD50K complete $N$-expert deployment matrix. Entries are mAP mean $\pm$ seed standard deviation on the 0–100 scale over five scientific seeds. FSD50K is the systematic deployment exception: learned channel simplex is not the highest observed deployment in any of its six formation–$N$ conditions.

| Formation | $N$ | Equal | Global | $N_{\text{eff}}^{G}$ | Channel | $N_{\text{eff}}^{C}$ |
|---|---|---|---|---|---|---|
| Dirichlet | 2 | $\mathbf{34.557 \pm 0.378}$ | $34.553 \pm 0.373$ | 1.980 | $34.407 \pm 0.410$ | 1.585 |
| Dirichlet | 4 | $34.325 \pm 0.230$ | $34.307 \pm 0.227$ | 3.833 | $34.157 \pm 0.308$ | 2.889 |
| Dirichlet | 8 | $34.422 \pm 0.619$ | $34.381 \pm 0.602$ | 7.404 | $34.269 \pm 0.572$ | 5.547 |
| FixedEqual | 2 | $34.525 \pm 0.323$ | $34.519 \pm 0.331$ | 1.998 | $34.412 \pm 0.361$ | 1.849 |
| FixedEqual | 4 | $34.346 \pm 0.108$ | $34.332 \pm 0.107$ | 3.960 | $34.191 \pm 0.082$ | 3.378 |
| FixedEqual | 8 | $33.830 \pm 0.754$ | $33.814 \pm 0.765$ | 7.718 | $33.739 \pm 0.691$ | 6.039 |

Table 20: Speech Commands V2 complete $N$-expert deployment matrix. Entries are clean accuracy mean $\pm$ seed standard deviation in percentage points over five scientific seeds.

| Formation | $N$ | Equal | Global | $N_{\text{eff}}^{G}$ | Channel | $N_{\text{eff}}^{C}$ |
|---|---|---|---|---|---|---|
| Dirichlet | 2 | $89.532 \pm 0.636$ | $89.543 \pm 0.589$ | 1.820 | $89.877 \pm 0.635$ | 1.319 |
| Dirichlet | 4 | $88.798 \pm 0.667$ | $88.847 \pm 0.721$ | 3.275 | $89.363 \pm 0.702$ | 2.394 |
| Dirichlet | 8 | $89.661 \pm 0.580$ | $89.752 \pm 0.478$ | 6.036 | $90.228 \pm 0.503$ | 4.094 |
| FixedEqual | 2 | $89.732 \pm 0.450$ | $89.672 \pm 0.513$ | 1.981 | $90.152 \pm 0.328$ | 1.402 |
| FixedEqual | 4 | $89.361 \pm 0.882$ | $89.310 \pm 0.846$ | 3.728 | $89.868 \pm 1.015$ | 2.426 |
| FixedEqual | 8 | $89.834 \pm 0.694$ | $89.952 \pm 0.623$ | 5.966 | $\mathbf{90.637 \pm 0.500}$ | 4.202 |

Table 21: CIFAR-100 complete $N$-expert deployment matrix. Entries are test accuracy mean $\pm$ seed standard deviation in percentage points over five scientific seeds.

| Formation | $N$ | Equal | Global | $N_{\text{eff}}^{G}$ | Channel | $N_{\text{eff}}^{C}$ |
|---|---|---|---|---|---|---|
| Dirichlet | 2 | $50.770 \pm 0.796$ | $50.756 \pm 0.808$ | 1.998 | $51.208 \pm 0.746$ | 1.795 |
| Dirichlet | 4 | $51.052 \pm 1.028$ | $51.070 \pm 0.983$ | 3.924 | $51.808 \pm 0.795$ | 3.044 |
| Dirichlet | 8 | $50.814 \pm 0.881$ | $50.904 \pm 0.919$ | 7.503 | $51.868 \pm 1.191$ | 6.337 |
| FixedEqual | 2 | $51.280 \pm 0.673$ | $51.236 \pm 0.720$ | 1.999 | $51.692 \pm 0.781$ | 1.945 |
| FixedEqual | 4 | $51.452 \pm 1.125$ | $51.420 \pm 1.144$ | 3.987 | $\mathbf{52.294 \pm 1.108}$ | 3.782 |
| FixedEqual | 8 | $50.776 \pm 0.735$ | $50.856 \pm 0.814$ | 7.909 | $51.956 \pm 0.652$ | 7.186 |

Table 22: Tiny ImageNet complete $N$-expert deployment matrix. Entries are final-evaluation accuracy mean $\pm$ seed standard deviation in percentage points over five scientific seeds.

| Formation | $N$ | Equal | Global | $N_{\text{eff}}^{G}$ | Channel | $N_{\text{eff}}^{C}$ |
|---|---|---|---|---|---|---|
| Dirichlet | 2 | $33.824 \pm 1.035$ | $33.824 \pm 1.031$ | 1.998 | $34.128 \pm 0.844$ | 1.812 |
| Dirichlet | 4 | $33.432 \pm 1.161$ | $33.424 \pm 1.099$ | 3.948 | $33.928 \pm 1.026$ | 3.410 |
| Dirichlet | 8 | $32.802 \pm 0.673$ | $32.858 \pm 0.632$ | 7.861 | $33.388 \pm 0.505$ | 6.516 |
| FixedEqual | 2 | $33.342 \pm 0.779$ | $33.340 \pm 0.788$ | 1.999 | $33.598 \pm 0.791$ | 1.964 |
| FixedEqual | 4 | $33.524 \pm 1.375$ | $33.508 \pm 1.314$ | 3.994 | $33.986 \pm 1.431$ | 3.734 |
| FixedEqual | 8 | $33.162 \pm 0.665$ | $33.224 \pm 0.713$ | 7.942 | $\mathbf{34.310 \pm 0.898}$ | 7.035 |

### H.2 Paired expert-count effects relative to the rerun $N = 2$ anchor

For the four non-ESC datasets, each difference is paired by scientific seed. For ESC-50, the entries below summarize the 25 paired seed–fold cells and are descriptive diagnostics; they are not treated as 25 independent inferential replicates. $\Delta_4$ and $\Delta_8$ denote $N = 4 - N = 2$ and $N = 8 - N = 2$, respectively, in primary-metric percentage points. The complete paired effects are reported in Tables 23–27.

Table 23: ESC-50 paired expert-count diagnostics. Entries are mean paired Robust-B changes $\pm$ cell-wise SD relative to the rerun $N = 2$ condition. Win counts are descriptive across the 25 seed–fold cells; official-fold aggregation remains the uncertainty unit in the main paper.

| Formation | Deployment | $\Delta_4$ | wins | $\Delta_8$ | wins |
|---|---|---|---|---|---|
| Dirichlet | Equal | $+0.130 \pm 2.243$ | 14/25 | $-1.298 \pm 2.322$ | 8/25 |
| Dirichlet | Global | $+0.206 \pm 2.140$ | 13/25 | $-1.018 \pm 2.105$ | 10/25 |
| Dirichlet | Channel | $+1.333 \pm 1.755$ | 18/25 | $-0.009 \pm 1.889$ | 12/25 |
| FixedEqual | Equal | $-0.566 \pm 1.933$ | 10/25 | $-0.553 \pm 2.084$ | 10/25 |
| FixedEqual | Global | $-0.319 \pm 1.634$ | 9/25 | $-0.268 \pm 2.315$ | 12/25 |
| FixedEqual | Channel | $+0.470 \pm 1.603$ | 15/25 | $+0.321 \pm 1.740$ | 14/25 |

Table 24: FSD50K paired expert-count effects. Entries are paired mAP changes $\pm$ paired-seed SD relative to rerun $N = 2$. All 12 mean $N = 4$ or $N = 8$ contrasts are negative.

| Formation | Deployment | $\Delta_4$ | wins | $\Delta_8$ | wins |
|---|---|---|---|---|---|
| Dirichlet | Equal | $-0.232 \pm 0.243$ | 1/5 | $-0.135 \pm 0.612$ | 2/5 |
| Dirichlet | Global | $-0.246 \pm 0.246$ | 1/5 | $-0.172 \pm 0.594$ | 2/5 |
| Dirichlet | Channel | $-0.250 \pm 0.259$ | 1/5 | $-0.138 \pm 0.541$ | 2/5 |
| FixedEqual | Equal | $-0.179 \pm 0.397$ | 2/5 | $-0.696 \pm 0.708$ | 1/5 |
| FixedEqual | Global | $-0.187 \pm 0.400$ | 2/5 | $-0.706 \pm 0.719$ | 1/5 |
| FixedEqual | Channel | $-0.221 \pm 0.413$ | 1/5 | $-0.673 \pm 0.663$ | 1/5 |

Table 25: Speech Commands V2 paired expert-count effects. Entries are paired clean-accuracy changes $\pm$ paired-seed SD relative to rerun $N = 2$.

| Formation | Deployment | $\Delta_4$ | wins | $\Delta_8$ | wins |
|---|---|---|---|---|---|
| Dirichlet | Equal | $-0.734 \pm 0.341$ | 0/5 | $+0.129 \pm 0.970$ | 2/5 |
| Dirichlet | Global | $-0.696 \pm 0.376$ | 0/5 | $+0.209 \pm 0.877$ | 3/5 |
| Dirichlet | Channel | $-0.514 \pm 0.401$ | 0/5 | $+0.351 \pm 0.894$ | 3/5 |
| FixedEqual | Equal | $-0.371 \pm 0.727$ | 3/5 | $+0.102 \pm 0.614$ | 3/5 |
| FixedEqual | Global | $-0.362 \pm 0.680$ | 1/5 | $+0.280 \pm 0.540$ | 4/5 |
| FixedEqual | Channel | $-0.284 \pm 0.817$ | 2/5 | $\mathbf{+0.485 \pm 0.315}$ | 5/5 |

Table 26: CIFAR-100 paired expert-count effects. Entries are paired test-accuracy changes $\pm$ paired-seed SD relative to rerun $N = 2$.

| Formation | Deployment | $\Delta_4$ | wins | $\Delta_8$ | wins |
|---|---|---|---|---|---|
| Dirichlet | Equal | $+0.282 \pm 1.216$ | 3/5 | $+0.044 \pm 1.140$ | 3/5 |
| Dirichlet | Global | $+0.314 \pm 1.222$ | 3/5 | $+0.148 \pm 1.180$ | 3/5 |
| Dirichlet | Channel | $+0.600 \pm 0.929$ | 4/5 | $+0.660 \pm 1.457$ | 3/5 |
| FixedEqual | Equal | $+0.172 \pm 0.479$ | 3/5 | $-0.504 \pm 1.075$ | 2/5 |
| FixedEqual | Global | $+0.184 \pm 0.461$ | 3/5 | $-0.380 \pm 1.195$ | 2/5 |
| FixedEqual | Channel | $\mathbf{+0.602 \pm 0.469}$ | 4/5 | $+0.264 \pm 0.938$ | 3/5 |

Table 27: Tiny ImageNet paired expert-count effects. Entries are paired final-accuracy changes $\pm$ paired-seed SD relative to rerun $N = 2$. Under channel deployment, the $N = 8 - N = 2$ effect reverses from $-0.740$ points for Dirichlet formation to $+0.712$ points for FixedEqual formation.

| Formation | Deployment | $\Delta_4$ | wins | $\Delta_8$ | wins |
|---|---|---|---|---|---|
| Dirichlet | Equal | $-0.392 \pm 0.456$ | 1/5 | $-1.022 \pm 1.177$ | 1/5 |
| Dirichlet | Global | $-0.400 \pm 0.376$ | 1/5 | $-0.966 \pm 1.109$ | 1/5 |
| Dirichlet | Channel | $-0.200 \pm 0.292$ | 2/5 | $\mathbf{-0.740 \pm 0.812}$ | 0/5 |
| FixedEqual | Equal | $+0.182 \pm 1.046$ | 4/5 | $-0.180 \pm 0.445$ | 2/5 |
| FixedEqual | Global | $+0.168 \pm 1.003$ | 3/5 | $-0.116 \pm 0.386$ | 2/5 |
| FixedEqual | Channel | $+0.388 \pm 0.920$ | 4/5 | $\mathbf{+0.712 \pm 0.287}$ | 5/5 |