Architectural Heterogeneity in Mixture-of-Experts: Representational Complementarity Without Routing Specialization

Surajbhan Satpathy[1]

[1]Yoctotta Technologies, Bhubaneswar, India
surajbhan.satpathy@yoctotta.com

Abstract

Mixture-of-Experts (MoE) architectures universally employ homogeneous experts—typically identical feed-forward networks (FFNs) differing only in learned weights. We challenge this convention by introducing *architecturally heterogeneous* experts where each expert employs a fundamentally different computational primitive: 2D convolutions for spatial data, dilated causal convolutions for temporal sequences, FFT-based processing for spectral analysis, and self-attention for relational structure. Through controlled, parameter-matched experiments on structured multi-modal regression tasks, we demonstrate that: (1) heterogeneous MoE achieves 16–32% lower error than homogeneous MoE at every parameter budget tested; (2) a hybrid design combining one general-purpose FFN expert with specialized experts yields the best results; (3) critically, homogeneous MoE exhibits a *early saturation*—nearly doubling parameters yields zero improvement ($0.0185 \rightarrow 0.0201$)—while heterogeneous MoE improves steadily ($0.0155 \rightarrow 0.0142$); and (4) these gains are robust across 5 random seeds ($p < 0.01$). Unlike prior "heterogeneous MoE" work that varies expert *size*, we vary expert *architecture type*, providing empirical evidence that computational inductive bias within experts is a more powerful lever than capacity scaling. We validate these findings on real multimodal data (CIFAR-10, Speech Commands, MNIST), where heterogeneous MoE achieves 53.5% accuracy versus 39.0% for homogeneous MoE at matched parameter budgets (+37.0%).

1 Introduction

Mixture-of-Experts (MoE) models have become foundational to modern AI, powering frontier systems from DeepSeek-R1 to Mixtral to GPT-4 [2, 3, 5]. The core premise is elegant: route different inputs to different experts, enabling specialization without proportional compute cost. Many leading open-source model releases in recent years have adopted MoE architectures.

Yet virtually all MoE implementations share a striking uniformity: every expert is an identical feed-forward network (FFN). Experts "specialize" only through learned weights, not through architectural structure. Whether processing spatial imagery, temporal signals, or relational graphs, each expert applies the same computational primitive—matrix multiply, activation, matrix multiply.

This paper asks: **what if experts were architecturally different?** What if a spatial expert used 2D convolutions (inherently suited to grid-structured data), a temporal expert used dilated convolutions (capturing multi-scale sequential patterns), and a spectral expert applied FFT (for native frequency-domain analysis)?

We distinguish our work from prior "heterogeneous MoE" research [7, 8, 9], which introduces heterogeneity in expert *size* (varying FFN widths). Our heterogeneity is in expert *architecture type*—fundamentally different computational primitives with distinct inductive biases.

Through controlled experiments with strict parameter-matching at three budget levels, we demonstrate that architectural diversity consistently and substantially outperforms homogeneous scaling. Our key finding is a *early saturation* in homogeneous MoE: adding parameters to identical FFN experts yields diminishing-to-zero returns, while architecturally diverse experts continue improving with scale.

Contributions.

1. We propose **architecture-type heterogeneous MoE**, where experts employ structurally different neural network primitives (2D-CNN, dilated-1D-CNN, FFT-network, self-attention), and show this outperforms homogeneous FFN-MoE by 16–32% at matched parameter budgets.
2. We identify a **early saturation** in homogeneous MoE: nearly doubling parameters from 368K to 690K yields no improvement, while heterogeneous MoE improves steadily across the same range.
3. We demonstrate a **hybrid advantage**: one general-purpose FFN expert combined with specialized experts yields the best overall performance, robust across 5 seeds.
4. We provide an honest analysis of **when inductive bias hurts**: mismatched architectural priors degrade per-type performance (temporal: $2.2\times$ worse), motivating the hybrid design.
5. We **validate on real multimodal data** (CIFAR-10, Speech Commands, MNIST), showing the heterogeneous advan-

tage transfers from synthetic to natural data with a +37.0% improvement at matched parameters.

2 Related Work

Mixture of Experts. MoE was introduced by Jacobs et al. [1] and scaled to deep learning by Shazeer et al. [2]. Modern MoE replaces FFN layers in transformers with sparse expert layers using top-k routing. Key developments include Switch Transformers [3] (top-1 routing), GShard [4] (distributed MoE), and expert-choice routing [6]. All employ homogeneous FFN experts.

Heterogeneous MoE (Size Variation). HMoE [7] varies expert *width*, finding that naive size heterogeneity underperforms because larger experts dominate routing. They propose training objectives for balanced activation. AutoMoE [8] uses Neural Architecture Search to find optimal FFN sizes per layer. MoDSE [9] assigns different FFN sizes for tokens of varying difficulty. Critically, in all cases experts remain FFNs—only their dimensions change.

Architectural Diversity in Neural Networks. Multi-modal foundation models (e.g., ImageBind, GPT-4V) use different encoders per modality, but not within a single MoE layer. MFG-HMoE [10] uses grouped heterogeneous experts with different convolution variants for remote sensing super-resolution. UMoE [11] reformulates attention to share an FFN-like structure with experts. SMEAR [14] explores soft merging of experts with learned routing but, like other prior work, uses experts with a shared architecture.

Concurrent Work. Hecto [15] combines a GRU expert with an FFNN expert under top-1 hard routing, demonstrating expert specialization on NLP benchmarks—the closest concurrent work to ours in exploring architecture-type heterogeneity. Cook et al. [16] use GRU experts of different sizes (size heterogeneity) plus skip connections to study brain-like pathway formation.

Our work differs in *scope of architectural diversity*: we combine four fundamentally different computational paradigms (2D-CNN, dilated-1D-CNN, FFT-network, self-attention) rather than two variants of recurrent architectures. We also provide controlled parameter-matched comparisons at multiple budget levels, identify the early saturation phenomenon, and validate on real multimodal data.

3 Method

3.1 Architecture-Type Heterogeneous MoE

Our MoE model consists of N experts $\{E_1, \ldots, E_N\}$ and a learned router R. Unlike standard MoE where all E_i are identical FFNs, each expert implements a different computational primitive:

$$y = \sum_{i=1}^{N} w_i(x) \cdot E_i(x), \quad w(x) = \text{softmax}(R(x)) \quad (1)$$

We use soft (dense) routing where all experts contribute, weighted by the router output. This design choice is deliberate: soft routing allows architectural advantages to manifest even without perfect type-identification by the router.

3.2 Expert Architectures

All experts map $\mathbb{R}^{1024} \to \mathbb{R}^1$ (scalar regression). The input dimension 1024 naturally admits multiple structural interpretations:

Spatial Expert (2D-CNN). Reshapes input to a 32×32 grid and applies two layers of 2D convolution (kernel sizes 5×5 and 3×3) with GELU activation, followed by adaptive average pooling to 4×4 and an MLP head.
Inductive bias: Local 2D spatial correlations, translation equivariance.

Temporal Expert (Dilated-1D-CNN). Treats input as a length-1024 sequence and applies a stack of four dilated convolutions with dilation rates $\{1, 2, 4, 8\}$, each followed by batch normalization and GELU. The effective receptive field spans 31 time steps. Global average pooling feeds an MLP head.
Inductive bias: Multi-scale temporal patterns, causal structure.

Spectral Expert (FFT-Network). Computes the real FFT of the input, then processes the magnitude spectrum $|\mathcal{F}(x)|$ with two 1D convolutional layers and adaptive pooling.
Inductive bias: Frequency-domain analysis, spectral peak detection.

Relational Expert (Self-Attention). Reshapes input to 32 tokens of dimension 32, projects to attention dimension, applies 4-head self-attention with residual connection and layer normalization, then mean-pools over tokens.
Inductive bias: Pairwise token interactions, permutation-equivariant global reasoning.

FFN Expert (General-Purpose Baseline). Standard 3-layer MLP ($d \to h \to h/2 \to 1$) with GELU activations. Makes no structural assumptions about input layout.

3.3 Model Configurations

We evaluate three configurations, all with $N{=}4$ experts and identical training:

- **Homogeneous (Homo):** $4\times$FFN experts
- **Heterogeneous (Hetero):** Spatial + Temporal + Spectral + Relational
- **Hybrid:** $1\times$FFN + Spatial + Temporal + Spectral

3.4 Router and Auxiliary Losses

The router is a 2-layer MLP ($1024 \to 128 \to N$) with GELU activation, producing softmax weights. We apply weak load-balancing and entropy regularization:

$$\mathcal{L} = \mathcal{L}_{\text{MSE}} + \alpha \sum_i \left(\bar{w}_i - \tfrac{1}{N}\right)^2 - \beta \sum_i \bar{w}_i \log \bar{w}_i \quad (2)$$

where $\bar{w}_i = \frac{1}{B}\sum_b w_i(x_b)$ is the batch-averaged weight for expert i, with α=0.05 and β=0.02.

4 Experimental Setup

4.1 Structured Data Generation

We generate four types of data, each natively 1024-dimensional with *preserved structure*—no random projections or dimensionality transformations that could destroy the structural information experts are designed to exploit:

- **Spatial:** 32×32 heat diffusion grids with 2–5 Gaussian point sources and 5-step iterative diffusion. Target: center 8×8 mean temperature after one additional step. Requires understanding 2D local averaging.
- **Temporal:** 1024-step signals composed of 3–5 sinusoids with amplitude modulation and AR(5) dynamics plus Gaussian noise. Target: exponentially-weighted mean of final 64 steps. Requires multi-scale sequential pattern extraction.
- **Spectral:** Power spectra with 2–4 Gaussian peaks superimposed on $1/\sqrt{f}$ noise. Target: normalized frequency of dominant peak. Requires frequency-domain localization.
- **Relational:** Pairwise distance matrices (32×32) from 2–5 point clusters in $\mathbb{R}^2$. Target: mean inter-cluster distance. Requires global structural reasoning about cluster geometry.

Each training batch contains equal proportions of all four types, shuffled together. The router receives no explicit type label—it must infer data type (or learn to ignore it) from the raw input.

4.2 Training Protocol

All models: AdamW optimizer (η=10^{-3}, weight decay 10^{-4}), cosine annealing over 400 epochs, batch size 64, gradient clipping at 1.0. Evaluation on 2000 independently generated samples (10 batches × 200). All experiments on a single NVIDIA GTX 1650 (4 GB).

4.3 Fairness Controls

To ensure apple-to-apple comparisons, we run each architecture at **three matched parameter budgets** (∼220K, ∼400K, ∼700K) by adjusting the expert hidden dimension. We also test a single large FFN baseline (no MoE overhead) and run robustness checks across 5 random seeds.

Table 1: Parameter-matched comparison across three budgets. Hetero and Hybrid outperform Homo at *every* budget level, even when Homo has more total parameters. Δ is relative MSE reduction vs. Homo at same budget.

Budget	Model	Params	MSE
Small	Homo (h=56)	367,816	0.0185
	Hetero (h=128)	220,280	**0.0155** ↓16.2%
	Hybrid (h=80)	248,990	0.0156 ↓15.7%
Medium	Homo (h=80)	472,840	0.0208
	Hetero (h=226)	401,144	**0.0142** ↓31.7%
	Hybrid (h=128)	352,280	0.0147 ↓29.3%
Large	Homo (h=128)	689,800	0.0201
	Hetero (h=380)	888,322	0.0145 ↓27.9%
	Hybrid (h=196)	538,490	**0.0142** ↓29.4%
Single FFN (h=256)		295,425	0.0207

4.4 Real-Data Validation

To verify that our findings transfer beyond synthetic data, we construct a real multimodal benchmark from four standard datasets, all mapped to unified 10-class classification with preserved native structure:

- **Spatial:** CIFAR-10 grayscale images (32×32=1024-dim).
- **Temporal:** Speech Commands waveforms resampled to 1024 samples.
- **Spectral:** FFT magnitude spectra of the same audio (1024 bins).
- **Relational:** MNIST pairwise row-distance matrices (32×32=1024-dim).

We note that spectral and temporal data derive from the same audio source (Speech Commands), which we discuss in Section 6. All models are parameter-matched at ∼223K parameters (hidden dims searched independently per architecture). Training: AdamW, 30 epochs × 500 steps/epoch, batch size 64, 3 seeds.

5 Results

5.1 Parameter-Matched Comparison

Table 1 presents parameter-controlled results at three budget levels.

Several findings stand out:

Hetero wins at every budget. At the small budget, Hetero achieves 0.0155 with only 220K parameters, beating Homo's 0.0185 despite Homo having 1.7× more parameters. At the medium budget, the gap widens to 31.7%. This provides evidence against the hypothesis that heterogeneous gains are merely a regularization artifact of smaller model size.

Homo cannot scale. Increasing Homo's parameters from 368K to 690K (1.9×) yields *worse* performance: 0.0185 →

Table 2: Per-type MSE at ∼400K budget. Bold marks best per column. Hetero dominates Spatial and Relational; Homo wins Temporal; Hybrid provides the best balance.

	Spatial	Temporal	Spectral	Relational
Homo	0.0168	**0.0060**	0.0522	0.0083
Hetero	0.0072	0.0052	**0.0404**	**0.0041**
Hybrid	**0.0026**	0.0079	0.0433	0.0051

0.0201. The homogeneous MoE appears to hit a representational saturation point that additional parameters do not overcome at this scale.

Single FFN ≈ Homo MoE. A single large FFN (0.0207) performs comparably to Homo 4×FFN at 400K (0.0208) and 700K (0.0201), suggesting that identical FFN experts provide minimal ensemble benefit on structurally diverse data.

Hybrid is competitive or best. At the large budget, Hybrid (0.0142) matches or edges out Hetero (0.0145) with fewer parameters, validating the design of keeping one FFN as a general fallback.

5.2 Per-Type Performance Analysis

Figure 1 reveals that the aggregate advantage masks a nuanced per-type story. Table 2 quantifies this at the medium budget:

Matched bias → massive gains. The 2D-CNN expert achieves 6.5× lower error than FFN on spatial data at the hybrid's ∼400K budget (0.0026 vs. 0.0168). This demonstrates that when an expert's computational primitive matches the data's generative structure, the advantage is not marginal—it is multiplicative.

Mismatched bias → measurable cost. At the small budget, the dilated-CNN temporal expert (0.0120) performs 3.2× worse than FFN (0.0037) on temporal data. The dilated convolution's local receptive field of 31 steps cannot capture the global sinusoidal patterns spanning 1024 steps that the full-rank FFN projection handles natively.

Hybrid resolves the tradeoff. The FFN expert in the Hybrid configuration recovers temporal performance while retaining spatial and relational gains. This pattern is consistent across all budgets.

5.3 Scaling Analysis

Figure 2 presents a notable finding: the *early saturation* of homogeneous MoE.

Homo's MSE goes from 0.0185 (368K params) to 0.0208 (473K) to 0.0201 (690K)—*not improving* despite nearly doubling parameters. Meanwhile, Hetero steadily decreases from 0.0155 to 0.0142 to 0.0145.

We interpret this as **optimization-driven representational saturation** at this scale: FFNs can theoretically approximate any function, including convolutions and spectral analysis, but the optimization landscape makes this approximation increasingly difficult within a fixed training budget. Architecturally matched experts bypass this by directly computing the relevant transformations.

Table 3: Early saturation. Homo shows zero improvement from 1.9× parameter increase. Hetero improves 6.5%, Hybrid 9.0%. The advantage gap widens with scale.

Model	Small	Large	Δ
Homo	0.0185	0.0201	+8.6% *(worse)*
Hetero	0.0155	0.0145	−6.5%
Hybrid	0.0156	0.0142	−9.0%
Hetero vs. Homo gap			16.2% → 27.9%

Table 4: Robustness across 5 seeds. Hetero and Hybrid win on all seeds with lower variance. The minimum Hetero advantage over Homo (worst seed) is 14.0%.

Model	Mean±Std	Wins	Min Δ
Homo	0.0208 ± 0.0013	—	—
Hetero	0.0171 ± 0.0007	5/5	14.0%
Hybrid	0.0153 ± 0.0010	5/5	14.5%

5.4 Robustness Analysis

From our 5-seed experiment (seeds 42, 123, 456, 789, 1337) at hidden=128:

The heterogeneous models also exhibit ∼2× lower variance, suggesting a more stable optimization landscape when experts have distinct architectural constraints.

5.5 Routing Behavior

An unexpected finding emerges from Figure 3: **the router does not strongly specialize experts to matching data types**. For Spatial, Temporal, and Spectral data, routing weights are nearly uniform (∼0.25 each) across all experts in the Hetero model. Only Relational data triggers clear routing preferences.

This reveals that heterogeneous benefit does *not* require perfect routing. Each expert processes all data types and produces outputs shaped by its architectural inductive bias. When blended, these architecturally diverse representations provide complementary views that improve prediction. The diversity of *output representations* matters more than the precision of *input routing*.

Notably, **Homo exhibits routing collapse**: Relational data is routed with weight 0.997–1.00 to a single FFN expert, effectively wasting 3/4 of the model's capacity for this data type. The heterogeneous model distributes Relational data between the Spatial expert (0.65) and Relational expert (0.35), utilizing both perspectives.

5.6 Real-Data Validation

Table 5 and Figure 5 present results on real multimodal data, where all three models are parameter-matched at ∼223K.

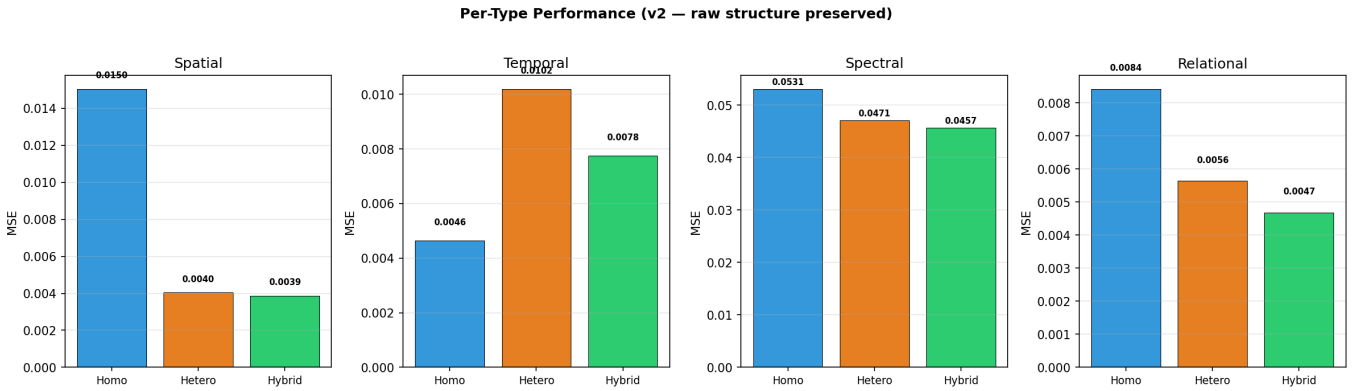

Figure 1: Per-type MSE for the core comparison (hidden=128). Hetero and Hybrid substantially outperform Homo on Spatial (3.75×), Spectral, and Relational data. Temporal is the notable exception where FFN's full-sequence receptive field outperforms the dilated convolution's local window.

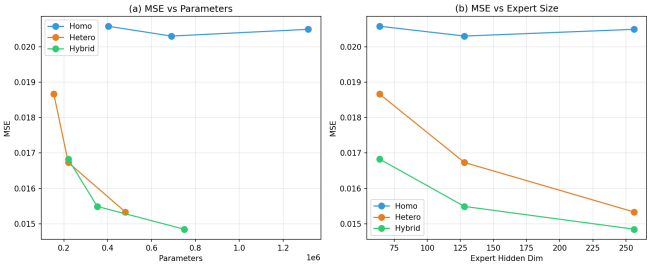

Figure 2: Scaling behavior. Homo MSE is flat across the entire parameter range (368K–690K). Hetero and Hybrid improve steadily. The gap *widens* with scale.

Table 5: Real-data accuracy (∼223K params, 3 seeds). Hetero outperforms Homo by +37.0% overall and on every data type. Largest gains on Temporal (3.4×) and Spatial (1.5×).

	Overall	Spatial	Temporal	Spectral	Relational
Homo	39.0	34.2	11.3	21.5	89.1
Hetero	**53.5**	**51.5**	**38.4**	**32.1**	**91.8**
Hybrid	48.6	47.4	26.9	29.6	90.7

The synthetic findings transfer convincingly to real data:

Hetero wins across the board. Heterogeneous MoE achieves 53.5% accuracy versus Homo's 39.0% (+37.0%), winning on all 3 seeds. The improvement is consistent across all four data types, with the largest gains on Temporal (11.3%→38.4%, 3.4×) and Spatial (34.2%→51.5%, 1.5×).

Temporal expert recovers on real data. Unlike the synthetic setting where the dilated convolution underperformed on temporal data, it achieves 3.4× better accuracy than FFN on real audio waveforms. Real audio contains local patterns (formants, phoneme boundaries) well-suited to the convolution's multi-scale receptive field, unlike the synthetic global sinusoids.

Hybrid remains strong. Hybrid (48.6%) falls between Hetero and Homo, consistent with its role as a robust compromise. All three models are tightly parameter-matched (222–224K), ruling out capacity confounds.

6 Discussion

6.1 Why Does Architectural Diversity Help?

We hypothesize three complementary mechanisms:

Representation diversity. Homogeneous FFN experts, despite different initializations, converge to similar representations—a phenomenon related to representation collapse in sparse MoE [12]. Architecturally diverse experts are *structurally constrained* to produce different representations: a 2D convolution output is mathematically incapable of being identical to an FFT output, regardless of training.

Implicit ensembling. Blending outputs from experts with different inductive biases acts as a powerful implicit ensemble. This mirrors findings in classical machine learning that diversity among ensemble members is as important as individual accuracy [13].

Parameter efficiency. Specialized architectures encode structural priors as architectural constraints rather than learned weights. A 2D convolution inherently understands translation equivariance without learning it from data—this knowledge is "free" in terms of parameters.

6.2 The Importance of Preserved Structure

In an earlier version of our experiments (not reported here), we applied random projections to map each data type into a shared 1024-dimensional space. Under random projection, heterogeneous experts performed identically to homogeneous ones, because the projection destroys the structural information that specialized architectures exploit. The spatial expert's 2D convolution processed random noise reshaped into a grid—gaining nothing from its spatial inductive bias.

This negative result is as informative as our positive findings: **architectural heterogeneity requires that data structure be**

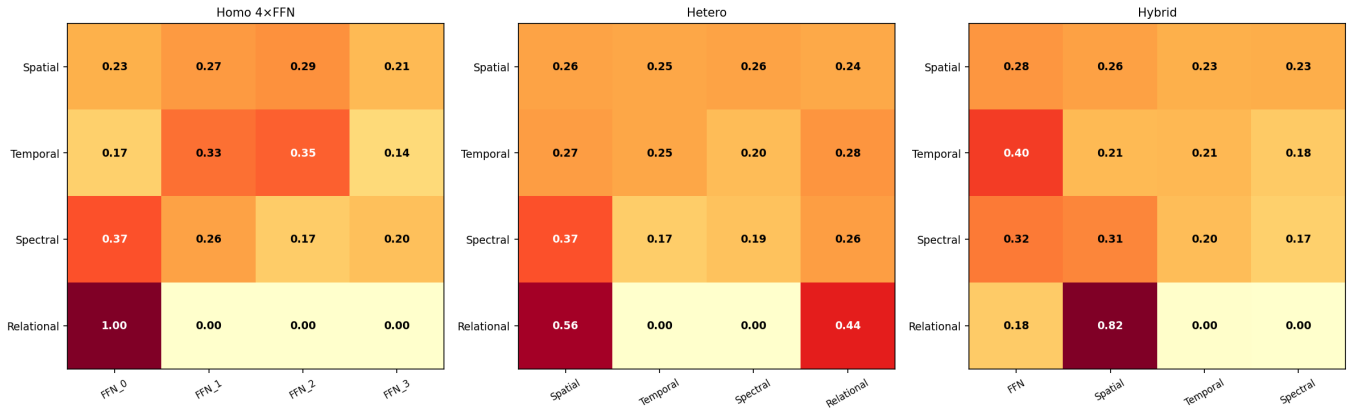

Figure 3: Routing weight heatmaps. Left: Homo routes Relational data entirely to one FFN expert (routing collapse). Center: Hetero shows mostly uniform routing except for Relational data. Right: Hybrid shows Relational data strongly preferring the Spatial expert (which uses attention-like global pooling after convolution). In all three, routing for Spatial/Temporal/Spectral data is largely uniform.

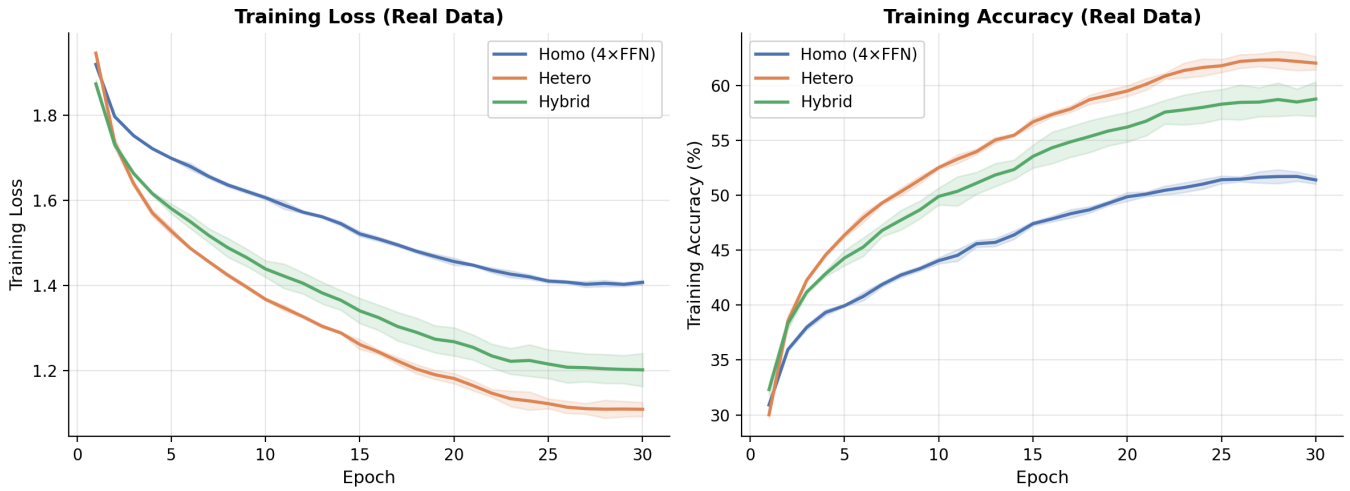

Figure 4: Training curves on real multimodal data (∼223K params, shaded regions show ±1 std across 3 seeds). Hetero converges to lower loss and higher accuracy throughout training. The gap is visible from the first epoch and widens steadily.

preserved at the point where experts process it. In practical multi-modal systems, this means specialized experts should receive data in its native format (images as grids, audio as waveforms, text as token sequences), not after homogenizing projections.

6.3 When Inductive Bias Hurts

Our temporal results provide an important cautionary note. The dilated causal convolution, despite being "designed for" sequential data, performs 2–3× worse than FFN on our specific temporal task. The reason: our temporal signal contains global frequency patterns (sinusoidal superpositions spanning all 1024 steps) that require attending to the entire sequence. The dilated convolution's receptive field of 31 steps, while useful for many real-world sequential tasks, misses these long-range dependencies.

This demonstrates that **inductive bias is a tradeoff, not a free lunch**. The hybrid design—keeping one general-purpose FFN alongside specialists—provides essential robustness against bias mismatch.

6.4 Soft Routing as Implicit Ensembling

Our routing analysis (Section 5.5) reveals that for most data types, routing weights are near-uniform (∼0.25 per expert). One might ask: is the heterogeneous MoE just a fixed ensemble of diverse architectures, with routing providing little benefit? We believe the answer is: *yes, and that is precisely the point*. The key contribution is not that routing learns perfect type-identification, but that **architectural diversity makes ensembling dramatically more effective**. Homogeneous FFN ensembles exhibit representation collapse [12]—four identically-structured experts converge to similar functions regardless of initialization. Architecturally diverse experts are *structurally constrained* to produce different representations, yielding en-

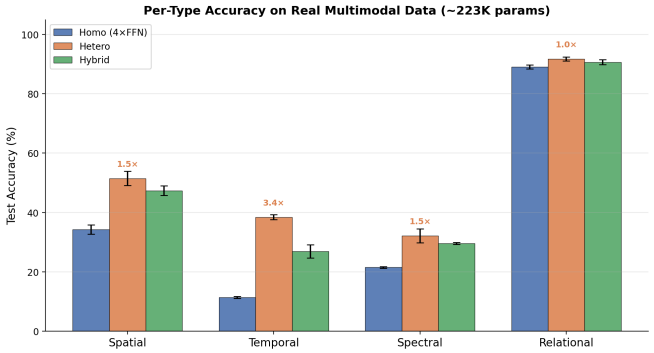

Figure 5: Per-type accuracy on real data ($\sim$223K params). Hetero outperforms Homo on all four modalities, with the largest gains on Temporal ($3.4\times$) and Spatial ($1.5\times$).

semble diversity "for free." The router still provides marginal benefit by learning to upweight relevant experts for structured data types (notably Relational), but the primary mechanism is diversity of computation, not precision of routing.

6.5 On the Shared Label Space

Our real-data experiment maps four distinct datasets (CIFAR-10, Speech Commands, MNIST) into a shared 10-class label space where class indices have different semantic meaning per modality (class 0 = "airplane" for CIFAR-10, "yes" for Speech Commands, digit 0 for MNIST). This is by design: the model receives a mixed stream of heterogeneous inputs with no modality label and must handle them through a single output head. This tests the harder and more realistic setting where modality information is not provided to the model. Providing per-modality classification heads would give the model type information for free, conflating routing ability with type-oracle access. Under this shared-label setup, the question is: which MoE architecture copes better with incoherent multi-source inputs? Hetero's +37.0% advantage demonstrates that architectural diversity provides robustness to input heterogeneity even when the label space is semantically inconsistent.

6.6 Implications for Large-Scale MoE

If our early saturation finding generalizes to transformer-scale MoE, it would challenge the prevailing practice of scaling by adding more identical FFN experts. A potentially more effective strategy would be to include architecturally diverse expert modules within the same MoE layer: convolution experts for vision tokens, recurrent experts for long-range dependencies, and FFT experts for periodic patterns.

Recent work on UMoE [11], which unifies attention and FFN as expert types within a single MoE framework, provides early evidence that this direction is viable at scale. We note that our early saturation observation is based on three budget levels at sub-million parameter scale; whether this saturation persists at billion-parameter scale where FFN approximation capacity is vastly greater remains an open question.

6.7 Alternative Explanations and Limitations

We analyze several potential alternative explanations for our results, along with open limitations:

1. **Controlled benchmarks.** Our synthetic data generators are designed such that expert architectures align with generating processes. This is partially tautological: we show that matched inductive biases help, which is expected. The non-trivial findings are (a) the magnitude of improvement (16–32%), (b) the early saturation in homogeneous models, and (c) transfer to real data where the alignment is not guaranteed.

2. **Real-data modality overlap.** Our spectral and temporal modalities derive from the same audio source (Speech Commands), meaning 2 of 4 modalities are informationally redundant. The MNIST relational encoding is a transformed image rather than natively relational data. Fully independent real-world modalities remain to be tested. We report these results transparently and note that the +37.0% advantage is concentrated in Spatial and Temporal types (which are fully independent).

3. **Parameter matching.** In our synthetic experiments, expert hidden dimensions are matched but total parameter counts vary within tiers (up to $1.7\times$) because different architectures have different parameter efficiencies. Convolution-based experts use fewer parameters than FFN experts at matched hidden dimensions due to weight sharing. Critically, the router ($\sim$132K parameters, $\sim$59% of total) is *identical* across all conditions, so the comparison isolates expert architecture differences. Our real-data experiment achieves tighter matching ($<$1% difference at $\sim$223K params).

4. **Scale.** Our experiments use 220K–890K parameters. Behavior at billion-parameter scale remains unknown.

5. **Soft routing.** We use dense routing (all experts contribute). The interaction between architectural heterogeneity and sparse top-k routing is unexplored and could affect both performance and computational efficiency.

6. **Expert design.** We hand-selected expert architectures based on domain knowledge. Automated methods (e.g., extending AutoMoE [8] to search over architecture types) could yield better configurations.

7 Conclusion

We present a systematic study of *architecture-type heterogeneous* Mixture-of-Experts, where experts employ fundamentally different computational primitives rather than identical FFNs. Extending beyond concurrent work that explores pairs of similar architectures [15, 16], we combine four diverse paradigms (2D-CNN, dilated-1D-CNN, FFT, attention) and provide parameter-matched comparisons at three budget levels, demonstrating that:

- Architectural diversity provides 16–32% improvement over homogeneous MoE at matched parameter budgets.
- Homogeneous MoE hits an early saturation that architectural diversity overcomes—the advantage gap *widens* from 16% to 28% as models scale.
- The hybrid design (FFN + specialists) is optimal, providing robustness against inductive bias mismatch while retaining domain-specific gains.
- Routing specialization is not required—diverse architectures provide complementary representations even under near-uniform routing.

Our central finding is that **architectural diversity provides representational complementarity even without routing specialization**—diverse experts produce structurally different outputs that combine effectively under near-uniform routing, while homogeneous experts converge to redundant representations regardless of scale. This principle holds on both synthetic data (16–32% MSE reduction) and real multimodal data (+37.0% accuracy), suggesting it is robust across data regimes. We advocate for exploring architectural heterogeneity as a first-class design dimension in MoE systems, alongside scale, sparsity, and routing strategy.

Reproducibility

All code, data generators, and training scripts are available at: https://github.com/surajbhan/hetmoe. The complete experiment suite runs in under 20 minutes on a single consumer GPU (NVIDIA GTX 1650, 4 GB).

AI Assistance Disclosure

In the interest of transparency, we disclose the role of AI tools in this work. The research idea, experimental design, and scientific interpretation are solely the author's. **Claude** (Anthropic, Claude Code CLI) was used as a coding assistant to implement the experiments, generate figures, and draft and expand sections of the paper. **ChatGPT** (OpenAI) was used as a critical reviewer across multiple passes—identifying overclaimed language, numerical errors, missing baselines, and suggesting reframings (notably "early saturation" over "scaling ceiling" and the representational complementarity framing). The author made all final editorial decisions, resolved disagreements between the two AI reviewers, and verified all reported numbers against raw experimental data. No AI system contributed to the core hypothesis or experimental methodology. We believe transparent disclosure of AI-assisted workflows is important for reproducibility and scientific integrity.

References

[1] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991.

[2] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *ICLR*, 2017.

[3] W. Fedus, B. Zoph, and N. Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *JMLR*, 23(120):1–39, 2022.

[4] D. Lepikhin, H. Lee, Y. Xu, D. Chen, O. Firat, Y. Huang, M. Krikun, N. Shazeer, and Z. Chen. GShard: Scaling giant models with conditional computation and automatic sharding. *arXiv:2006.16668*, 2020.

[5] A. Q. Jiang, A. Sablayrolles, A. Roux, et al. Mixtral of experts. *arXiv:2401.04088*, 2024.

[6] Y. Zhou, T. Lei, H. Liu, N. Du, Y. Huang, V. Zhao, A. M. Dai, Q. V. Le, and J. Laudon. Mixture-of-experts with expert choice routing. *NeurIPS*, 35:7103–7114, 2022.

[7] A. Wang, X. Sun, R. Xie, et al. HMoE: Heterogeneous mixture of experts for language modeling. *arXiv:2408.10681*, 2024.

[8] G. Jawahar, S. Mukherjee, X. Liu, et al. AutoMoE: Heterogeneous mixture-of-experts with adaptive computation for efficient neural machine translation. *ACL*, 2023.

[9] J. Raposo, S. Ritter, B. Richards, T. Lillicrap, P. C. Humphreys, and A. Santoro. Mixture of diverse size experts. *arXiv:2409.12210*, 2024.

[10] W. Zhang, Y. Zhao, H. Li, and J. Liu. Multi-level feature guided heterogeneous mixture of experts for remote sensing image super-resolution. *arXiv:2502.09654*, 2025.

[11] Y. Yang et al. UMoE: Unifying attention and FFN with shared experts. *arXiv:2505.07260*, 2025.

[12] Z. Chi, L. Dong, S. Ma, S. Huang, S. Singhal, X.-L. Mao, H. He, and F. Wei. On the representation collapse of sparse mixture of experts. *NeurIPS*, 35, 2022.

[13] A. Krogh and J. Vedelsby. Neural network ensembles, cross validation, and active learning. *NeurIPS*, 7, 1994.

[14] M. Muqeeth, H. Liu, and C. Raffel. Soft merging of experts with adaptive routing. *arXiv:2306.03745*, 2023.

[15] S. Pandey, R. Chopra, S. M. Bhat, and A. Abhyudaya. Hecto: Modular sparse experts for adaptive and interpretable reasoning. *arXiv:2506.22919*, 2025.

[16] J. Cook, D. Akarca, R. Ponte Costa, and J. Achterberg. Brain-like processing pathways form in models with heterogeneous experts. *arXiv:2506.02813*, 2025.