BrainMoE: A Brain-Inspired Modular Mixture-of-Experts Model for Multimodal Reasoning

Anonymous ACL submission

Abstract

Reliable cross-modal reasoning requires Multimodal Large Language Model (MLLMs) with interpretable and controllable computation. However, existing models rely on data-driven architectures that entangle multimodal processing, obscuring the functional processes underlying their outputs. Inspired by the human brain's specialized and dynamically collaborative networks, we propose **BrainMoE**, a multimodal MoE model with six sensory and cognitive experts: Language, Logic, Social, World Knowledge, Vision, and Auditory. To support flexible sensory–cognitive coordination, we incorporates a layer-wise adaptive router that dynamically determines which experts are activated at each layer based on task-processing demands. Experiments on text, image, and audio reasoning benchmarks demonstrate that BrainMoE outperforms strong baselines by 1.59 and 2.2 points on text-only and multimodal benchmarks, while matching competitive external multimodal models. Routing and ablation analyses demonstrate clear expert specialization and causally important module-level contributions. Experiments on the cognitive benchmark further show that adaptive-routing BrainMoE aligns more closely with human behavioral patterns than baselines and static routing methods.

1 Introduction

Interpretable and controllable multimodal large language models (MLLMs) are important not only for reliable cross-modal reasoning, but also for cognitive science, as they can provide an in silico framework for examining how functionally specialized brain networks coordinate to support complex behavior. However, existing MLLMs typically rely on end-to-end learning over densely shared parameters, entangling visual, auditory, linguistic, and higher-level cognitive processes. Although Mixture-of-Experts (MoE) models partition computation into multiple expert modules (Fedus et al., 2022; Riquelme et al., 2021), their experts often conflate heterogeneous functions, limiting functional interpretability. In contrast, the human cortex relies on functionally specialized yet dynamically coordinated networks for language, reasoning, social cognition, and perception (Yeo et al., 2011; Power et al., 2011; Duncan, 2010), motivating neurocognitive organization as an inductive bias for efficient and interpretable multimodal experts.

Drawing on these neurocognitive principles, recent efforts have sought to embed functional specialization into artificial neural networks. A notable step in this direction is MiCRo (AlKhamissi et al., 2025a), which partitions transformer layers into four specialized modules mirroring distinct cognitive brain networks. While this framework improves interpretability and maintains competitive textual reasoning, it remains restricted to text inputs and lacks dedicated perceptual modules. Furthermore, its reliance on static routing prevents it from modeling the demand-dependent dynamic coordination characteristic of human sensory and cognitive systems (Mesulam, 1998; Capouskova et al., 2023).

To address these limitations, we propose **BrainMoE**, a brain-inspired multimodal mixture-of-experts model built on functionally grounded modularity. BrainMoE organizes reasoning into six sensory and cognitive experts: *Vision*, *Auditory*, *Language*, *Logic*, *Social*, and *World Knowledge* (Fedorenko et al., 2011; Duncan, 2010; Saxe and Kanwisher, 2013; Raichle et al., 2001; Buckner et al., 2008; Yeo et al., 2011; Power et al., 2011). These modules provide cognitively grounded functional abstractions for interpretable multimodal reasoning. To support flexible sensory–cognitive coordination, we incorporate a layer-wise adaptive router that determines which experts to activate at each layer according to task-processing demands. This design allows the model dynamically adjusting expert participation based on the evolving token-level context. In this way, BrainMoE combines the efficiency ad-

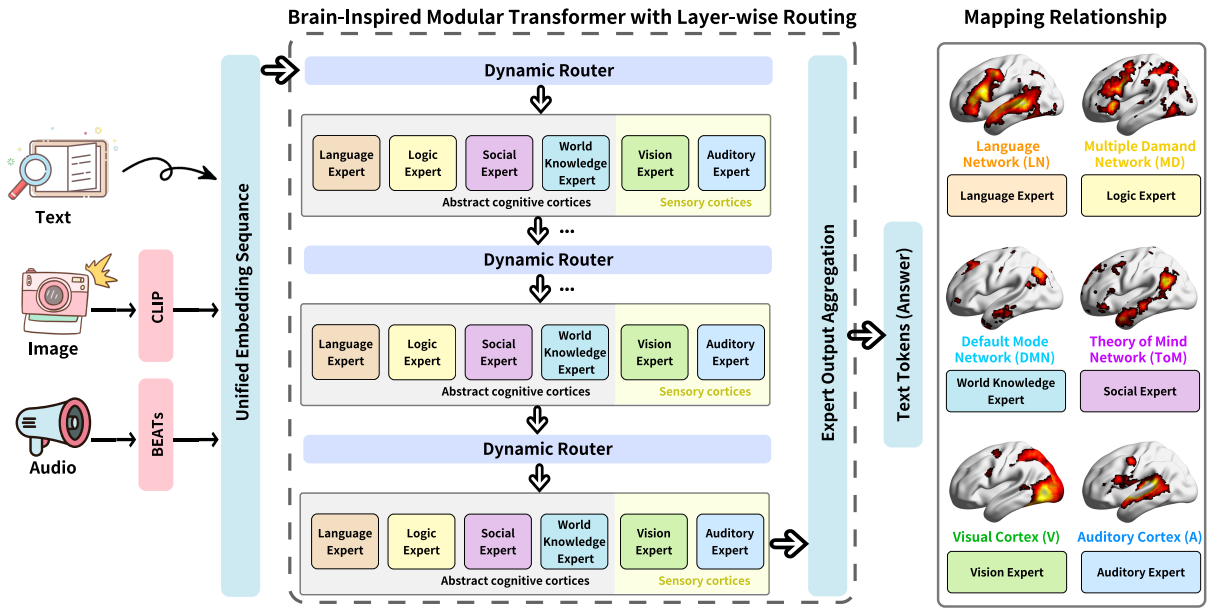

Figure 1: Overview of BrainMoE. BrainMoE maps text, image, and audio inputs into a unified token sequence and performs layer-wise token-level routing through a brain-inspired modular Transformer. At each layer, a dynamic router adaptively selects both the relevant experts and the number of activated experts from six specialized modules: Language, Logic, Social, World Knowledge, Vision, and Auditory. Their outputs are aggregated to generate the final answer, enabling flexible sensory–cognitive coordination. The right panel shows the intended mapping between these experts and representative human brain systems, including language, multiple-demand, default-mode, theory-of-mind, visual, and auditory networks.

vantages of sparse MoE with a more interpretable and flexible routing mechanism.

We evaluate BrainMoE on text, image, and audio reasoning benchmarks. BrainMoE improves over strong baselines by 1.59 points on text-only test sets and 2.2 points on multimodal test sets, while remaining competitive with strong external vision- and audio-language models. Beyond accuracy, routing analysis shows that BrainMoE assigns tokens to task-relevant functional experts, while expert ablation demonstrates that removing the most relevant expert leads to the largest performance degradation. Human-behavior correlation analysis further shows that BrainMoE with adaptive routing aligns more closely with human behavioral patterns than both baseline models and static-routing variants. These results suggest that BrainMoE provides an effective and interpretable framework for brain-inspired multimodal reasoning.

Our contributions are threefold:

- We propose **BrainMoE**, a brain-inspired multimodal MoE model that organizes experts into sensory and cognitive functional modules.
- We incorporate a layer-wise adaptive router that determines which experts to activate at each layer according to task-processing demands, enabling sparse and flexible sensory–cognitive coordination.
- We demonstrate strong performance across text, image, and audio reasoning benchmarks, with routing, ablation, and human-behavior correlation analyses supporting functional specialization and cognitive plausibility.

2 Related Work

Functional specialization and adaptive coordination in the brain. Human cognition relies on functionally specialized and dynamically coordinated neural systems. Sensory processing is supported by specialized visual and auditory pathways (Goodale and Milner, 1992; Hickok and Poeppel, 2007), while multisensory cognition depends on interactions between modality-specific cortices and association areas (Ghazanfar and Schroeder, 2006; Driver and Noesselt, 2008). At higher cognitive levels, neuroscience has identified systems related to language, domain-general control, social reasoning, and semantic or episodic situation modeling, including the language network, multiple-demand system, theory-of-mind network, and default-mode network (Fedorenko et al., 2011; Duncan, 2010; Saxe and Kanwisher, 2013; Raichle

et al., 2001; Buckner et al., 2008). These systems are adaptively coordinated under changing task demands, as suggested by flexible-hub and control-network studies (Cole et al., 2013). Inspired by this view, BrainMoE treats multimodal reasoning as adaptive coordination among functionally grounded sensory and cognitive experts.

Sparse expert models and brain-aligned interpretability. Mixture-of-Experts (MoE) models improve efficiency by routing inputs to sparse expert subsets, enabling larger model capacity with reduced per-token computation (Shazeer et al., 2017; Lepikhin et al., 2020; Fedus et al., 2022). Recent work further improves routing stability, load balancing, and multimodal fusion (Zhou et al., 2022; Zoph et al., 2022; Yu et al., 2024; Zong et al., 2024). However, conventional MoE routing is typically optimized through end-to-end objectives, and the resulting activation patterns may entangle task, modality, and cognitive factors, leading to limited interpretability. This makes fine-grained analyses such as module localization and routing inspection less straightforward. Recent neuroscience-inspired studies address related issues by identifying causally task-relevant units in LLMs through brain-style functional localization (AlKhamissi et al., 2025b), or by constructing modular language models aligned with cognitive systems (AlKhamissi et al., 2025a). Building on this line of work, BrainMoE constructs a brain-inspired multimodal MoE with explicit sensory and cognitive experts, enabling efficient multimodal reasoning with improved module-level interpretability.

3 BrainMoE: Brain-Inspired Modular Multimodal Model

3.1 Overview

We propose **BrainMoE**, a modular multimodal mixture-of-experts framework that decomposes multimodal understanding into functionally specialized expert modules, as shown in Figure 1. The model consists of six experts: four cognitive experts for language, reasoning, social inference, and world knowledge, and two sensory experts for visual and auditory inputs. Given a multimodal sequence, BrainMoE first maps heterogeneous modality features into a shared hidden space and then processes them with a Transformer-based MoE backbone.

The key component of BrainMoE is a layer-wise adaptive router. Instead of applying a single static routing policy, the router performs token-level expert selection independently at each layer. This enables the model to dynamically adjust expert usage across both tokens and network depth, allowing lower layers to emphasize modality-specific processing and higher layers to support semantic reasoning (Martín-Fernández et al., 2025).

3.2 Multimodal Representation Alignment

BrainMoE maps text, image, and audio inputs into a shared hidden space before expert routing. Text tokens are encoded by the backbone embedding layer. For non-text modalities, we extract image features using a CLIP visual encoder (Radford et al., 2021) and audio features using a BEATs audio encoder (Chen et al., 2022), and then map them to the language hidden dimension with separate lightweight MLP projectors. The resulting modality tokens are concatenated into a unified sequence, with modality indicators retained for modality-aware routing.

3.3 Modular Expert Architecture

BrainMoE contains $E = 6$ experts initialized from the same pretrained backbone and then specialized toward distinct functional roles. We organize them into cognitive and sensory groups.

Cognitive Experts. The cognitive group consists of Language, Logic, Social, and World Knowledge experts. The *Language Expert* is inspired by the left-lateralized frontotemporal language network (LN), including language-selective regions in the left perisylvian cortex, and targets linguistic composition and semantic abstraction (Fedorenko et al., 2010, 2011). The *Logic Expert* is motivated by the domain-general multiple-demand (MD) network, a bilateral frontoparietal system associated with executive control, working memory, and demanding reasoning, and supports controlled multi-step inference (Duncan, 2010; Fedorenko et al., 2013). The *Social Expert* is inspired by the theory-of-mind (ToM) network, including the temporo-parietal junction (TPJ) and medial prefrontal cortex (mPFC), and captures pragmatic, affective, and mental-state-related cues (Saxe and Powell, 2006). The *World Knowledge Expert* is inspired by the default mode network (DMN), which is linked to semantic memory, episodic retrieval, and situation-model construction, and supports background knowledge integration (Andrews-Hanna et al., 2010).

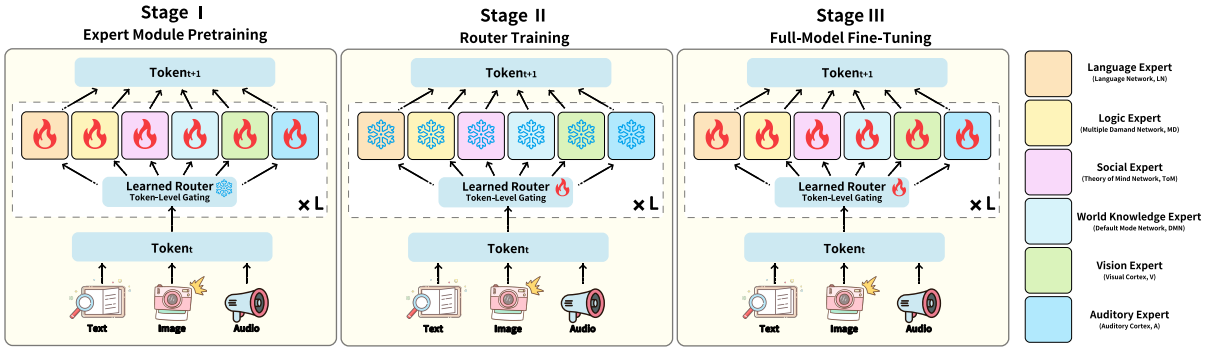

Figure 2: Training pipeline of BrainMoE. BrainMoE is trained in three stages: expert-module pretraining with a frozen token-level router, router training with frozen experts, and end-to-end full-model fine-tuning. ×L indicates repeated layer-wise routing.

Sensory Experts. The sensory group contains Vision and Auditory experts. The *Vision Expert* is inspired by the visual cortex (V), where neural populations encode visual properties such as orientation, spatial frequency, contrast, and color, and extracts perceptual representations from visual tokens (Ng et al., 2006; Grootswagers et al., 2024). The *Auditory Expert* is inspired by the auditory cortex (A), including Heschl's gyrus and adjacent early auditory areas, and captures temporal and structural patterns from audio inputs (Bizley and Walker, 2010; Rupp et al., 2025).

This decomposition provides an architectural inductive bias for functional specialization, with the biological analogy used as design motivation rather than strict neural correspondence (Margulies et al., 2016; Paquola et al., 2019; Martín-Fernández et al., 2025). All experts are optimized end-to-end under standard machine learning objectives.

3.4 Layer-wise Adaptive Router

The routing decision is based on multiple factors. First, a base gate scores each expert according to the current token representation. Second, the router incorporates contextual information by combining local window statistics and global sequence-level statistics. Third, the router uses a long-term expert utilization prior to discourage persistent overuse of the same experts.

Instead of using a fixed top-k value, BrainMoE adopts a threshold-based adaptive activation strategy. For each token, the router activates experts whose probabilities exceed an adaptive threshold. The threshold is higher when the router is confident, leading to selective activation, and lower when the router is uncertain, allowing broader expert collaboration. This mechanism enables sparse, context-sensitive, and layer-specific expert collaboration. More details can be found in Appendix B.1.

3.5 Training Procedure

We train BrainMoE in three stages, as shown in Figure 2. **Stage 1** specializes experts by post-training copies of the same pretrained backbone on function-aligned data. **Stage 2** freezes the experts and trains the router together with multimodal projection modules, with load balancing and steady-state regularization to stabilize token-to-expert assignment (Shazeer et al., 2017; Fedus et al., 2022; Lepikhin et al., 2020). **Stage 3** jointly fine-tunes all parameters on downstream supervised data, while retaining adaptive top-k routing and routing regularizers for sparse and balanced expert coordination. More details can be found in Appendix B.2.

4 Experiments

4.1 Datasets

Expert-labeled supervision (Stages 1–2). For Stages 1–2, we use a small expert-labeled corpus to supervise token-level routing (See Figure **??** in Appendix D.). For samples without reasoning traces, we first augment them with O1-generated reasoning responses. For text tokens, GPT-5 assigns one expert label from {LN, MD, ToM, DMN, V, A} and returns a JSON mapping from token indices to experts. Vision-encoder tokens are labeled as V and audio-encoder tokens as A. We curate and label multiple vision and audio datasets for sensory supervision, as shown in Table 3, while directly adopting the expert-labeled text corpus released by MiCRo (AlKhamissi et al., 2025a). The annotation prompts are provided in Appendix C.

Table 1: Text-only reasoning performance under the LLaMA 3.2-1B backbone. Dense Model is the original non-MoE baseline, MoB Model is a generic mixture-of-blocks baseline, and MiCRo is the brain-inspired text-only baseline. Best results are in bold.

Method	Activated Experts	gsm8k ↑	MMLU ↑	Minerva ↑	PIQA ↑	Average ↑
Text-only / Base Model: LLaMA 3.2-1B						
Dense Model	–	**36.67**	30.23	4.81	72.15	35.97
MoB Model	1/4	30.51	27.21	3.64	70.43	32.95
MiCRo	1/4	34.70	31.20	4.60	73.10	35.90
BrainMoE_text	1/4	33.67	32.48	4.77	72.39	35.83
BrainMoE_text with sensory experts	1/6	33.20	30.58	4.25	70.81	34.71
BrainMoE_text with sensory data	1/4	35.02	32.70	4.85	74.59	36.79
BrainMoE_Multimodal w/o sensory experts	1/4	35.83	**33.56**	**4.91**	**75.67**	**37.49**

Table 2: Multimodal reasoning performance under the LLaMA 3.2-1B backbone. BrainMoE_Multimodal enables six cognitive and sensory experts, fixed settings activate $k \in \{1, 2, 4, 6\}$ experts per token, while the adaptive setting dynamically selects the activation size. Best results are in bold, and second-best results are underlined. ↑ indicates higher is better; WER ↓ indicates lower is better.

Method	Activated Experts	gsm8k ↑	MMLU ↑	Minerva ↑	PIQA ↑	VQAv2 ↑	TextVQA ↑	OK-VQA ↑	LibriSpeech (WER ↓)	CommonVoice (WER ↓)
Multimodal / Base Model: LLaMA 3.2-1B										
Dense Model	–	35.96	33.29	5.07	72.14	46.52	23.25	**37.31**	4.54	8.16
MoB Model	2/6	31.48	29.39	3.90	72.03	44.13	17.17	23.42	5.59	7.05
Shared Trunk Model	2/6	37.52	35.04	5.13	73.49	47.15	23.80	33.47	5.53	7.07
OpenFlamingo-3B	–	20.07	24.58	2.15	51.72	47.61	25.23	30.97	–	–
Qwen3-ASR 1.7B	–	–	–	–	–	–	–	–	**2.35**	**3.92**
MiMo-ASR (∼2.5B)	–	–	–	–	–	–	–	–	5.23	7.51
BrainMoE_Multimodal (Ours)	1/6	35.12	32.87	4.78	73.25	31.61	14.57	22.15	7.42	9.29
	2/6	**38.07**	36.83	7.04	**77.35**	50.95	26.72	31.21	5.17	6.28
	4/6	37.36	**37.92**	6.19	76.15	50.33	25.16	34.90	4.22	6.92
	6/6	35.53	35.41	5.18	73.86	47.16	22.01	33.98	7.34	10.03
	adaptive	37.55	37.06	**7.47**	76.02	**51.39**	**27.48**	34.01	3.04	6.19

Large-scale SFT (Stage 3). Stage 3 performs end-to-end supervised fine-tuning on the AI2 Tulu-3-SFT-Mixture dataset, which provides chat-formatted supervision in a message-based schema (Lambert et al., 2024). Since expert specialization is induced by the routing supervision in Stages 1–2, Stage 3 uses no routing labels.

4.2 Setup

Backbone. All experts are initialized from LLaMA-3.2-1B[1]. For each input, the router activates either a fixed top-k subset or an adaptive number of experts, enabling sparse expert utilization while preserving shared model capacity.

Evaluation benchmarks and metrics. We evaluate BrainMoE on nine benchmarks covering text reasoning, visual question answering, and speech recognition. Text-only reasoning is evaluated on GSM8K (Cobbe et al., 2021), MMLU (Hendrycks et al., 2020), Minerva (Nagrani et al., 2025), and PIQA (Bisk et al., 2020). Visual reasoning is evaluated on VQAv2 (Goyal et al., 2017), TextVQA (Singh et al., 2019), and OK-VQA (Marino et al., 2019). Speech recognition is evaluated on LibriSpeech (Panayotov et al., 2015) and Common Voice (Ardila et al., 2020). We report accuracy for reasoning and VQA tasks, and word error rate (WER) for ASR tasks, where higher accuracy and lower WER indicate better performance.

Baselines. We compare BrainMoE with both internal and external baselines. Internal baselines include a parameter-matched Dense Model; a generic mixture-of-blocks (MoB) model; and the Shared Trunk Model, which uses a shared multimodal backbone without functionally separated experts. We also compare with MiCRo (AlKhamissi et al., 2025a), a brain-inspired text-only modular baseline. For external multimodal and speech baselines, we include OpenFlamingo-3B (Awadalla et al., 2023), Qwen3-ASR 1.7B (Shi et al., 2026), and MiMo-ASR (Xiaomi, 2026).

Since MiCRo is a text-only modular baseline, we split the evaluation into text-only and multimodal settings to ensure fair comparison and to separately assess BrainMoE's multimodal capability. In the text-only setting, we further compare

[1] https://huggingface.co/meta-llama/Llama-3.2-1B

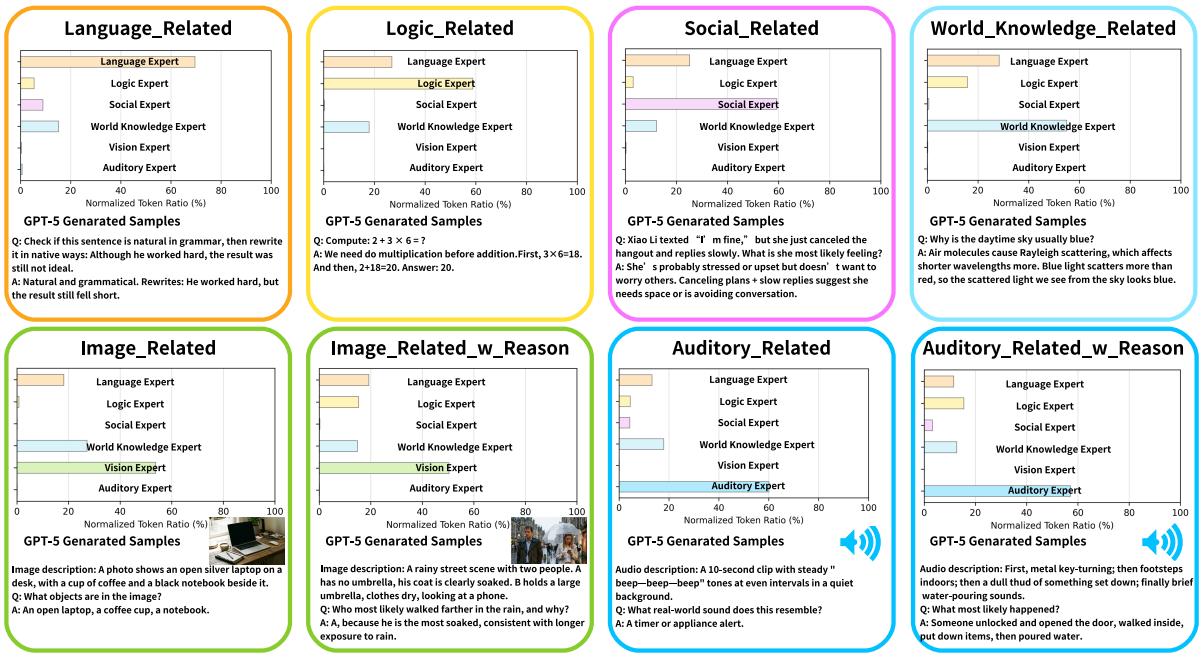

Figure 3: Token-level expert routing in BrainMoE. Routing behavior is evaluated using 80 Q–A pairs: 10 per cognitive expert and 20 per sensory expert (10 easy/10 hard). Bars show the average token proportion (%) routed to each expert within each subset. Experts not shown receive no routing mass in the corresponding subset.

Table 3: Part of the supervision corpus used in Stages 1–2 for sensory-expert routing. For each sample, per-token expert labels from {Language(LN), Logic(MD), World(DMN), Social(ToM), Vision(V), Auditory(A)} are obtained via GPT-5.

Category	Task	Dataset	#Samples
Visual	Color Recognition	Visual Genome	100
Visual	Object Recognition	Visual Genome	100
Visual	Object Localization	Visual Genome	100
Visual	Object Counting	Visual Genome	100
Visual	Image Captioning	MS-COCO	200
Visual	Image Captioning	Flickr30k	100
Visual	Commonsense Reasoning	VCR	300
		Visual Subtotal	**1000**
Audio	Audio Captioning	MACS	300
Audio	Audio Captioning	Clotho	300
Audio	Audio Captioning	WavCaps	200
Audio	Event Classification	FSD50k	200
		Audio Subtotal	**1000**

several BrainMoE variants to analyze the effects of expert composition and training signals. BrainMoE_text uses only cognitive experts trained on cognitive-expert data; BrainMoE_text with sensory experts adds sensory experts while keeping training data unchanged; BrainMoE_text with sensory data introduces sensory-expert data without adding sensory experts; and BrainMoE_Multimodal w/o sensory experts is trained with both cognitive and sensory experts/data, but masks sensory experts during text-only evaluation.

4.3 Results and Discussion

Main Results. Tables 1 and 2 summarize the main results. In the text-only setting, BrainMoE variants outperform the generic MoB baseline and are competitive with or stronger than Dense Model and MiCRo. The best text-only result is achieved by **BrainMoE_Multimodal w/o sensory experts**, which improves the average score over Dense Model, MoB Model, and MiCRo by 1.52, 4.54, and 1.59 points, respectively. This suggests that multimodal training benefits cognitive expert specialization even when sensory experts are masked at inference time.

In the multimodal setting, BrainMoE extends the same backbone to text, vision, and speech tasks. The adaptive router achieves the strongest BrainMoE performance on four benchmarks, including Minerva, VQAv2, TextVQA, and LibriSpeech, and provides a better overall trade-off than fixed routing. Fixed Top-k results show that activating too few experts underuses specialized capacity, while activating all experts can introduce cross-expert interference. Overall, these results indicate that adaptive sparse routing is effective for coordinating sensory and cognitive experts across modalities.

Expert Utilization Patterns. Figure 3 analyzes BrainMoE's routing behavior using controlled diagnostic inputs generated by GPT-5. We construct 80 question–answer pairs covering the six expert domains, including difficulty-controlled subsets for vision and audio, and aggregate token-level router assignments within each subset. The results show clear domain specialization: tokens are preferentially routed to the expert aligned with the input domain, e.g., Logic for reasoning-oriented questions, Vision for visually grounded inputs, and Auditory for audio-centric inputs. We further observe a difficulty-sensitive routing pattern: simple perceptual queries rarely activate the Logic expert, whereas harder visual/audio queries induce stronger Logic-expert activation. These patterns show that BrainMoE routes by domain and reasoning demand, engaging sensory experts for perception and cognitive experts for harder queries.

Single-Expert Ablation. To assess expert functionality, we ablate one expert at a time from the full BrainMoE (Top-6) and measure the performance change across benchmarks (Figure 4). The results show a specialization pattern: ablating an expert causes the largest degradation on tasks aligned with its intended function, while producing smaller effects on unrelated tasks. For example, removing the Logic expert affects math and reasoning benchmarks the most, such as GSM8K, Minerva, and MMLU_logic, but has a limited impact on vision- or audio-centric tasks, which align with the routing patterns in Figure 3, where experts receiving higher routing mass for a domain are also more critical for that domain. Together, the activation and ablation results indicate that BrainMoE learns interpretable and causally relevant expert specialization, consistent with its brain-inspired modular design (Sporns and Betzel, 2016; Binder et al., 2017; Hartwigsen et al., 2021; AlKhamissi et al., 2025b).

Routing Sparsity and Entropy Analysis. We further investigate the impact of the number of activated experts on BrainMoE's performance and routing behavior. As shown in Figure 5, we compare fixed Top-k routing with our adaptive strategy in terms of average accuracy and normalized routing entropy. See Appendix A for more details. Lower entropy indicates stronger expert specificity, as routing decisions are more concentrated on a subset of experts. Top-1 achieves the lowest entropy but suffers from limited capacity, leading to the worst accuracy. Increasing k to 2 or 4 improves

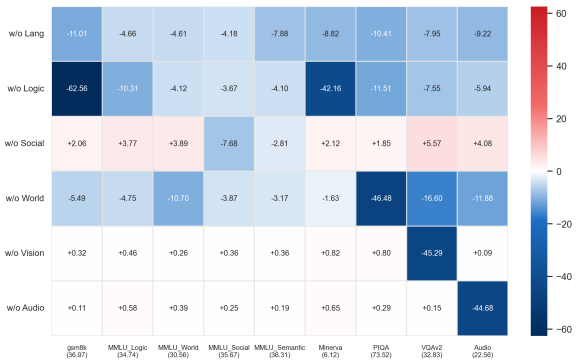

Figure 4: Single-expert ablation heatmap. Performance changes across evaluation benchmarks when ablating each expert module individually from the full model. Rows denote the ablated expert and columns denote benchmarks. Each cell reports the relative change after ablation, computed as $(ablated_score - BrainMoE_Top_6)/(BrainMoE_Top_6) \times 100\%$.

performance by enabling moderate cross-expert collaboration while preserving clear expert roles. In contrast, Top-6 activates all experts, which substantially increases entropy and degrades accuracy, suggesting diluted specialization. Adaptive routing achieves the highest average accuracy with low entropy, demonstrating a favorable trade-off between expert specificity and multimodal collaboration.

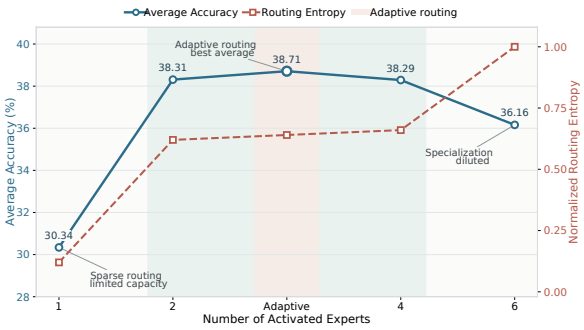

Figure 5: Effect of routing sparsity on accuracy and routing entropy. Fixed Top-k routing shows a trade-off between expert capacity and specialization: Top-1 underuses expert capacity, whereas Top-6 increases entropy and dilutes specialization. Adaptive routing achieves the best average accuracy with moderate entropy.

Human-Behavior Alignment on CogBench. We evaluate BrainMoE on CogBench (Coda-Forno et al., 2024), which contains seven cognitive psychology experiments and ten human-normalized behavioral metrics. Scores are normalized such that 0 corresponds to random behavior and 1 to average human behavior; we report both individual metrics and the aggregate SBRE score.

As shown in Table 4, adaptive BrainMoE

Table 4: Human-behavior alignment on CogBench. Scores are human-normalized, where 0 denotes random behavior and 1 denotes average human behavior. SBRE aggregates the overall behavioral similarity, with higher values indicating stronger alignment. Best results are in bold.

Model	Activate Experts	Directed Exp.	Random Exp.	Prior Weight	Likelihood Weight	Learning Rate	Optimism Bias	Model-basedness	Temporal Discount	Risk Taking	Meta-cognition	SBRE ↑
Dense Model	–	0.11	0.08	0.12	0.20	0.30	0.75	0.33	2.05	1.80	0.61	0.58
MoB Model	2/6	0.11	0.07	0.14	0.12	0.55	0.54	0.30	2.00	1.75	0.82	0.59
Shared Trunk Model	2/6	0.12	0.09	0.12	0.15	0.61	0.58	0.39	2.03	1.84	**0.83**	0.60
BrainMoE_Multimodal	1/6	0.13	0.20	0.19	0.18	0.97	0.76	0.51	1.98	1.82	0.57	0.63
	2/6	0.15	0.22	0.20	0.21	1.02	0.78	0.54	1.97	1.83	0.60	0.64
	4/6	0.10	0.15	0.13	0.13	0.83	0.72	0.41	2.01	1.85	0.58	0.61
	6/6	0.09	0.11	0.10	0.12	0.46	0.52	0.34	2.01	1.88	0.51	0.57
	adaptive	**0.16**	**0.24**	**0.20**	**0.24**	**1.04**	**0.79**	**0.57**	**1.96**	**1.80**	0.64	**0.65**

achieves the highest SBRE among all baselines and routing variants, indicating the strongest overall alignment with human behavior. Among fixed-routing settings, Top-2 performs best, while Top-4 and Top-6 degrade alignment. This suggests that text-based cognitive tasks usually require a compact set of relevant cognitive experts, whereas activating more experts may introduce interference. The adaptive router further improves over fixed routing by adjusting the activation size to task demand across probabilistic reasoning, exploration, risk preference, metacognition, and planning tasks.

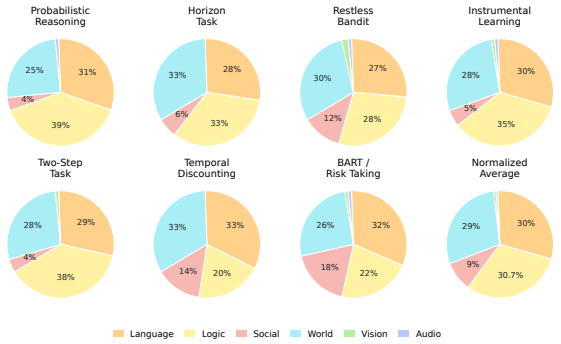

Figure 6: Task-wise routing distribution of adaptive BrainMoE on CogBench. Pie chart reports the cross-layer token routing proportion assigned to each expert.

Figure 6 shows that adaptive routing rarely selects Vision or Audio experts on text-only CogBench, whereas Language, Logic, World Knowledge, and Social experts are recruited in task-dependent proportions, indicating effective suppression of irrelevant experts. Figure 7 further shows that BrainMoE's human-like behavior arises from functionally selective expert contributions rather than uniform gains across modules. Since CogBench is text-based, removing the Language expert causes the largest drop by disrupting instruction parsing. Beyond this broad effect, the ablation reveals module-behavior correspondence: the Logic expert is more important for evidence integration,

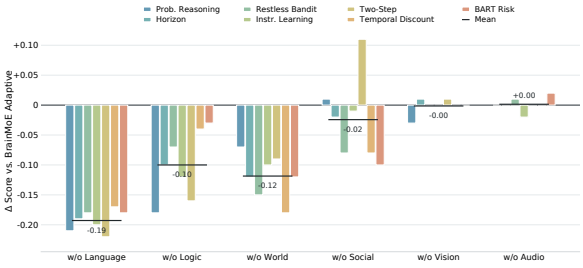

Figure 7: Single-expert ablation on CogBench. Bars show the change in human-normalized score after removing one expert from BrainMoE_Multimodal, using the full Top-6 model as the reference. Negative values indicate performance degradation, and black horizontal lines denote the average change across tasks.

exploration, learning, and model-based planning, whereas the World Knowledge expert contributes more to situation modeling, temporal evaluation, and risk-sensitive decision making. The negligible effects of Vision and Auditory ablations serve as negative controls for this text-only benchmark. Overall, BrainMoE exhibits both global human-behavior alignment and local correspondence between expert roles and cognitive behavior.

5 Conclusion

We introduced **BrainMoE**, a brain-inspired multimodal MoE framework that decomposes multimodal reasoning into six sensory and cognitive experts. Through a layer-wise adaptive router, BrainMoE dynamically selects task-relevant experts for each token, enabling flexible sensory–cognitive coordination. Experiments across text, image, and audio benchmarks show that BrainMoE outperforms strong baselines by 1.59 and 2.2 points on text-only and multimodal benchmarks, respectively. Routing, ablation, and cognitive-benchmark analyses demonstrate BrainMoE's expert specialization, causal module-level contributions, and stronger human-behavior alignment under adaptive routing than baselines and static-routing variants.

Limitations

This work has several limitations. First, BrainMoE is motivated by neurocognitive functional specialization, but the proposed expert design should be viewed as a high-level functional abstraction rather than a precise model of human brain organization. The six experts provide an interpretable way to organize multimodal reasoning, but real cognitive systems involve more fine-grained, overlapping, and dynamically interacting neural processes.

Second, our experiments focus on representative text, vision, and audio benchmarks. Although these tasks cover a range of reasoning and multimodal understanding scenarios, they do not exhaust all possible forms of multimodal reasoning. Therefore, the conclusions should be understood within the scope of the evaluated benchmarks.

Finally, our interpretability analysis is based on routing patterns, expert ablations, and behavioral alignment. These analyses provide useful evidence that the experts acquire task-relevant roles, but they do not fully characterize all internal mechanisms of each expert.

References

Badr AlKhamissi, C Nicolò De Sabbata, Greta Tuckute, Zeming Chen, Martin Schrimpf, and Antoine Bosselut. 2025a. Mixture of cognitive reasoners: Modular reasoning with brain-like specialization. *arXiv preprint arXiv:2506.13331*.

Badr AlKhamissi, Greta Tuckute, Antoine Bosselut, and Martin Schrimpf. 2025b. The llm language network: A neuroscientific approach for identifying causally task-relevant units. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10887–10911.

Jessica R Andrews-Hanna, Jay S Reidler, Jorge Sepulcre, Renee Poulin, and Randy L Buckner. 2010. Functional-anatomic fractionation of the brain's default network. *Neuron*, 65(4):550–562.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the twelfth language resources and evaluation conference*, pages 4218–4222.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, and 1 others. 2023. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*.

Julia C Binder, Ladina Bezzola, Aurea IS Haueter, Carina Klein, Jürg Kühnis, Hansruedi Baetschmann, and Lutz Jäncke. 2017. Expertise-related functional brain network efficiency in healthy older adults. *BMC neuroscience*, 18(1):2.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, and 1 others. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7432–7439.

Jennifer K Bizley and Kerry MM Walker. 2010. Sensitivity and selectivity of neurons in auditory cortex to the pitch, timbre, and location of sounds. *The Neuroscientist*, 16(4):453–469.

Randy L Buckner, Jessica R Andrews-Hanna, and Daniel L Schacter. 2008. The brain's default network: anatomy, function, and relevance to disease. *Annals of the new York Academy of Sciences*, 1124(1):1–38.

Katerina Capouskova, Gorka Zamora-López, Morten L Kringelbach, and Gustavo Deco. 2023. Integration and segregation manifolds in the brain ensure cognitive flexibility during tasks and rest. *Human brain mapping*, 44(18):6349–6363.

Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, and Furu Wei. 2022. Beats: Audio pre-training with acoustic tokenizers. *arXiv preprint arXiv:2212.09058*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Julian Coda-Forno, Marcel Binz, Jane X Wang, and Eric Schulz. 2024. Cogbench: a large language model walks into a psychology lab. *arXiv preprint arXiv:2402.18225*.

Michael W Cole, Jeremy R Reynolds, Jonathan D Power, Grega Repovs, Alan Anticevic, and Todd S Braver. 2013. Multi-task connectivity reveals flexible hubs for adaptive task control. *Nature neuroscience*, 16(9):1348–1355.

Jon Driver and Toemme Noesselt. 2008. Multisensory interplay reveals crossmodal influences on 'sensory-specific'brain regions, neural responses, and judgments. *Neuron*, 57(1):11–23.

John Duncan. 2010. The multiple-demand (md) system of the primate brain: mental programs for intelligent behaviour. *Trends in cognitive sciences*, 14(4):172–179.

Evelina Fedorenko, Michael K Behr, and Nancy Kanwisher. 2011. Functional specificity for high-level linguistic processing in the human brain. *Proceedings of the National Academy of Sciences*, 108(39):16428–16433.

Evelina Fedorenko, John Duncan, and Nancy Kanwisher. 2013. Broad domain generality in focal regions of frontal and parietal cortex. *Proceedings of the National Academy of Sciences*, 110(41):16616–16621.

Evelina Fedorenko, Po-Jang Hsieh, Alfonso Nieto-Castañón, Susan Whitfield-Gabrieli, and Nancy Kanwisher. 2010. New method for fmri investigations of language: defining rois functionally in individual subjects. *Journal of neurophysiology*, 104(2):1177–1194.

William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39.

Asif A Ghazanfar and Charles E Schroeder. 2006. Is neocortex essentially multisensory? *Trends in cognitive sciences*, 10(6):278–285.

Melvyn A Goodale and A David Milner. 1992. Separate visual pathways for perception and action. *Trends in neurosciences*, 15(1):20–25.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Tijl Grootswagers, Amanda K Robinson, Sophia M Shatek, and Thomas A Carlson. 2024. Mapping the dynamics of visual feature coding: Insights into perception and integration. *PLOS Computational Biology*, 20(1):e1011760.

Gesa Hartwigsen, Yoshua Bengio, and Danilo Bzdok. 2021. How does hemispheric specialization contribute to human-defining cognition? *Neuron*, 109(13):2075–2090.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Gregory Hickok and David Poeppel. 2007. The cortical organization of speech processing. *Nature reviews neuroscience*, 8(5):393–402.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, and 1 others. 2024. Tulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. 2020. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*.

Daniel S Margulies, Satrajit S Ghosh, Alexandros Goulas, Marcel Falkiewicz, Julia M Huntenburg, Georg Langs, Gleb Bezgin, Simon B Eickhoff, F Xavier Castellanos, Michael Petrides, and 1 others. 2016. Situating the default-mode network along a principal gradient of macroscale cortical organization. *Proceedings of the National Academy of Sciences*, 113(44):12574–12579.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204.

Jesús Martín-Fernández, Nayra Caballero-Estebaranz, Esteban Félez, Natalia Navarro-Peris, Pedro Pérez Del Rosario, Raúl Hernández Bisshopp, and Jaime Domínguez-Báez. 2025. Where are higher-order cognitive functions? the paradox of non-locality in awake cognitive mapping using a complex dynamic system framework. *Frontiers in Psychology*, 16:1542505.

M-Marsel Mesulam. 1998. From sensation to cognition. *Brain: a journal of neurology*, 121(6):1013–1052.

Arsha Nagrani, Sachit Menon, Ahmet Iscen, Shyamal Buch, Ramin Mehran, Nilpa Jha, Anja Hauth, Yukun Zhu, Carl Vondrick, Mikhail Sirotenko, and 1 others. 2025. Minerva: Evaluating complex video reasoning. *arXiv preprint arXiv:2505.00681*.

Jeffrey Ng, Anil A Bharath, and Li Zhaoping. 2006. A survey of architecture and function of the primary visual cortex (v1). *EURASIP Journal on Advances in Signal Processing*, 2007(1):097961.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

Casey Paquola, Reinder Vos De Wael, Konrad Wagstyl, Richard AI Bethlehem, Seok-Jun Hong, Jakob Seidlitz, Edward T Bullmore, Alan C Evans, Bratislav Misic, Daniel S Margulies, and 1 others. 2019. Microstructural and functional gradients are increasingly dissociated in transmodal cortices. *PLoS biology*, 17(5):e3000284.

Jonathan D Power, Alexander L Cohen, Steven M Nelson, Gagan S Wig, Kelly Anne Barnes, Jessica A Church, Alecia C Vogel, Timothy O Laumann, Fran M Miezin, Bradley L Schlaggar, and 1 others. 2011. Functional network organization of the human brain. *Neuron*, 72(4):665–678.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Marcus E Raichle, Ann Mary MacLeod, Abraham Z Snyder, William J Powers, Debra A Gusnard, and Gordon L Shulman. 2001. A default mode of brain function. *Proceedings of the national academy of sciences*, 98(2):676–682.

Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. 2021. Scaling vision with sparse mixture of experts. *Advances in Neural Information Processing Systems*, 34:8583–8595.

Kyle M Rupp, Jasmine L Hect, Emily E Harford, Lori L Holt, Avniel Singh Ghuman, and Taylor J Abel. 2025. A hierarchy of processing complexity and timescales for natural sounds in the human auditory cortex. *Proceedings of the National Academy of Sciences*, 122(18):e2412243122.

Rebecca Saxe and Nancy Kanwisher. 2013. People thinking about thinking people: the role of the temporo-parietal junction in "theory of mind". In *Social neuroscience*, pages 171–182. Psychology Press.

Rebecca Saxe and Lindsey J Powell. 2006. It's the thought that counts: specific brain regions for one component of theory of mind. *Psychological science*, 17(8):692–699.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.

Xian Shi, Xiong Wang, Zhifang Guo, Yongqi Wang, Pei Zhang, Xinyu Zhang, Zishan Guo, Hongkun Hao, Yu Xi, Baosong Yang, and 1 others. 2026. Qwen3-asr technical report. *arXiv preprint arXiv:2601.21337*.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326.

Olaf Sporns and Richard F Betzel. 2016. Modular brain networks. *Annual review of psychology*, 67(1):613–640.

LLM-Core-Team Xiaomi. 2026. Mimo-v2.5-asr: Robust speech recognition across languages, dialects, and complex acoustic scenarios.

BT Thomas Yeo, Fenna M Krienen, Jorge Sepulcre, Mert R Sabuncu, Danial Lashkari, Marisa Hollinshead, Joshua L Roffman, Jordan W Smoller, Lilla Zöllei, Jonathan R Polimeni, and 1 others. 2011. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of neurophysiology*.

Haofei Yu, Zhengyang Qi, Lawrence Keunho Jang, Russ Salakhutdinov, Louis-Philippe Morency, and Paul Pu Liang. 2024. Mmoe: Enhancing multimodal models with mixtures of multimodal interaction experts. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10006–10030.

Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Zhao, Andrew M Dai, Quoc V Le, James Laudon, and 1 others. 2022. Mixture-of-experts with expert choice routing. *Advances in Neural Information Processing Systems*, 35:7103–7114.

Zhuofan Zong, Bingqi Ma, Dazhong Shen, Guanglu Song, Hao Shao, Dongzhi Jiang, Hongsheng Li, and Yu Liu. 2024. Mova: Adapting mixture of vision experts to multimodal context. *Advances in Neural Information Processing Systems*, 37:103305–103333.

Barret Zoph, Irwan Bello, Sameer Kumar, Nan Du, Yanping Huang, Jeff Dean, Noam Shazeer, and William Fedus. 2022. St-moe: Designing stable and transferable sparse expert models. *arXiv preprint arXiv:2202.08906*.

A Routing Entropy Calculation and Expert Specificity

We quantify expert utilization specificity using the normalized entropy of the empirical routing distribution. Let E denote the number of experts, with $E = 6$ in BrainMoE. For an input sequence $x^{(n)}$, the router at layer ℓ produces a categorical distribution over experts for each non-padding token t:

$$\pi_{\ell,t}^{(n)} = \mathrm{softmax}\Big(g_\ell(h_{\ell,t}^{(n)})\Big) \in \Delta^{E-1},$$

where $h_{\ell,t}^{(n)}$ is the token representation and $g_\ell(\cdot)$ is the layer-specific routing function. After applying the routing policy—either fixed Top-k routing or adaptive threshold routing—we obtain the active expert set $\mathcal{S}_{\ell,t}^{(n)} \subseteq \{1, \ldots, E\}$. For fixed Top-$k$, $\mathcal{S}_{\ell,t}^{(n)}$ contains the k experts with the largest routing probabilities. For adaptive routing, it contains all experts whose routing probabilities exceed the token-specific adaptive threshold, with a top-1 fallback ensuring that at least one expert is activated.

We compute entropy based on the post-selection routing mass, as this accurately reflects the experts

actively utilized for computation. Specifically, the selected routing distribution is defined as:

$$q_{\ell,t,e}^{(n)} = \frac{\pi_{\ell,t,e}^{(n)} \mathbf{1}[e \in \mathcal{S}_{\ell,t}^{(n)}]}{\sum_{j=1}^{E} \pi_{\ell,t,j}^{(n)} \mathbf{1}[j \in \mathcal{S}_{\ell,t}^{(n)}]},$$

where $q_{\ell,t,e}^{(n)} = 0$ for inactive experts. Given an evaluation subset $\mathcal{D}$, we first aggregate the routing mass across all examples, layers, and non-padding tokens:

$$u_e(\mathcal{D}) = \frac{1}{M} \sum_{x^{(n)} \in \mathcal{D}} \sum_{\ell=1}^{L} \sum_{t=1}^{T_n} q_{\ell,t,e}^{(n)},$$

$$\text{where} \quad M = \sum_{x^{(n)} \in \mathcal{D}} L T_n,$$

and T_n denotes the number of non-padding tokens in sequence $x^{(n)}$. By construction, $u(\mathcal{D})$ is a valid empirical distribution over experts, i.e., $\sum_{e=1}^{E} u_e(\mathcal{D}) = 1$. The routing entropy of $\mathcal{D}$ is then computed as:

$$H(\mathcal{D}) = -\sum_{e=1}^{E} u_e(\mathcal{D}) \log u_e(\mathcal{D}),$$

using the standard convention $0 \log 0 = 0$. To make entropy values comparable across different configurations, we report the normalized routing entropy:

$$H_{\text{norm}}(\mathcal{D}) = \frac{H(\mathcal{D})}{\log E}.$$

Thus, $H_{\text{norm}} \in [0, 1]$. Values approaching 0 indicate that the routing mass is concentrated on a small subset of experts, whereas values approaching 1 indicate a near-uniform distribution across all experts.

This entropy statistic captures expert specificity by measuring the concentration of the aggregated routing distribution. If a diagnostic subset consistently activates a small set of functionally relevant experts—such as the Logic expert for reasoning-oriented inputs or the Vision expert for visually grounded inputs—the corresponding utilization distribution $u(\mathcal{D})$ becomes highly peaked, yielding low entropy. In contrast, if the routing mass is spread uniformly, the entropy approaches its maximum value, indicating weak functional selectivity. Therefore, lower normalized entropy reflects stronger expert specificity, signifying that the model selectively assigns computation based on domain characteristics.

Importantly, low entropy alone is insufficient to establish meaningful specialization, as a degenerate router that uniformly collapses to a single expert would also produce low entropy. We therefore interpret routing entropy jointly with domain-conditioned routing patterns and single-expert ablations. A routing pattern is deemed functionally specific only when low entropy is corroborated by two additional pieces of evidence: semantic alignment between the activated experts and the input domain, and a disproportionately large performance degradation when the high-utilization expert is ablated. This rigor prevents conflating genuine expert specialization with trivial expert collapse.

B Implementation Details

B.1 Router Configuration Details

We report the router-specific configuration used in BrainMoE in Table 5. These hyperparameters control routing stability, causal contextual routing, and long-term expert-load regulation. All reported values correspond to the router configuration used in our experiments.

For token position t at layer ℓ, the router produces expert logits

$$z_{\ell,t} \in R^E,$$

where $E = 6$ denotes the number of BrainMoE experts. The expert probabilities are computed using temperature-scaled softmax:

$$p_{\ell,t,e} = \frac{\exp(z_{\ell,t,e}/\tau)}{\sum_{j=1}^{E} \exp(z_{\ell,t,j}/\tau)},$$

where $\tau = \texttt{router-temperature} = 1.0$. This setting leaves the learned router logits neither additionally sharpened nor smoothed.

Because BrainMoE is trained under a causal next-token prediction objective, router-side contextual information is computed using only prefix-available representations. For each position t, the local context window is defined as

$$\mathcal{C}_w(t) = \{t' : \max(1, t - w + 1) \leq t' \leq t\},$$

where $w = \texttt{context-window} = 16$. The context-aware router therefore summarizes only the current and preceding token representations:

$$c_{\ell,t} = \text{Pool}\left(\{h_{\ell,t'} : t' \in \mathcal{C}_w(t)\}\right).$$

This causal formulation prevents the router from accessing future token representations during next-token prediction, while still allowing expert selection to depend on local semantic context.

Table 5: Router-specific hyperparameters used in BrainMoE.

Hyperparameter	Value	Description
`use-layerwise-router`	True	Uses an independent router at each MoE layer.
`use-context-gate`	True	Enables context-aware routing.
`context-window`	16	Size of the causal local context window.
`router-aux-loss-coef`	0.001	Coefficient for the router auxiliary loss.
`router-temperature`	1.0	Temperature for router softmax normalization.
`use-homeostatic-load-prior`	True	Enables the long-term expert-load prior.
`router-load-prior-scale`	0.1	Scaling factor for the load-prior correction.
`router-homeostasis-coef`	0.0005	Coefficient for the homeostatic regularizer.

Table 6: SFT hyperparameters used in Stage 3.

Hyperparameter	Value	Description
`gradient-accumulation-steps`	8	Number of micro-batches accumulated before each optimizer update.
`learning-rate`	2×10^{-5}	Initial learning rate used for SFT optimization.
`warmup-ratio`	0.03	Fraction of training steps used for learning-rate warmup.
`max-length`	4096	Maximum token sequence length during SFT.
`dataset-num-proc`	8	Number of parallel processes for dataset mapping.
`prefilter-num-proc`	8	Number of parallel processes for sample pre-filtering.

The training objective includes weak routing regularization:

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda_{\text{aux}}\mathcal{L}_{\text{aux}} + \lambda_{\text{homeo}}\mathcal{L}_{\text{homeo}},$$

where

$$\lambda_{\text{aux}} = 0.001, \qquad \lambda_{\text{homeo}} = 0.0005.$$

Here, $\mathcal{L}_{\text{aux}}$ denotes the router auxiliary loss, and $\mathcal{L}_{\text{homeo}}$ denotes the homeostatic regularization term. These terms are used to improve routing stability and discourage persistent expert imbalance, while their small coefficients prevent the regularizers from dominating the supervised task objective.

BrainMoE further uses a long-term expert-load prior, controlled by

$$\alpha_{\text{prior}} = \texttt{router-load-prior-scale} = 0.1.$$

This prior provides a weak correction against persistent overuse of the same experts. Importantly, it is used as a global stabilization mechanism rather than a per-example uniformity constraint. Thus, the router can still assign tokens selectively to semantically relevant experts while reducing the risk of expert collapse during training.

B.2 Supervised Fine-Tuning Configuration

We report the supervised fine-tuning (SFT) configuration used in Stage 3 in Table 6. These hyperparameters control optimization, sequence length, and dataset preprocessing throughput. Only the SFT-specific parameters used in our experiments are reported here.

This SFT configuration is designed to provide stable fine-tuning with moderate memory usage: gradient accumulation improves the effective batch size, the warmup phase stabilizes early optimization, the 4096-token sequence length supports long-form instruction data, and parallel preprocessing reduces data-loading overhead.

C Prompt for Expert-Label Annotation

To construct the expert-labeled supervision data used in the first two training stages, we assign short token sequences to one of the six BrainMoE experts. The goal of this annotation procedure is to provide token-level routing supervision that is aligned with the intended functional role of each expert. For image and audio examples, we first obtain a model-generated scene description and then segment the description into short token sequences. Each sequence is assigned exactly one expert label from $\{\text{LN}, \text{MD}, \text{ToM}, \text{DMN}, \text{V}, \text{A}\}$.

The annotation prompt is designed to separate syntactic, reasoning-related, social, semantic, visual-perceptual, and auditory-perceptual content. A key distinction is made between the Default Mode Network (DMN) and the Visual Perception expert (V). DMN is used for conceptual identity and semantic category information, such as object names, scene labels, named entities, and abstract concepts. In contrast, V is reserved for directly observable low-level visual attributes, such as color, shape, texture, size, illumination, and spatial relations. This distinction prevents visually grounded descriptions

from assigning all object-related tokens to the visual expert and instead encourages the router to distinguish perceptual attributes from semantic object recognition. Similarly, the Auditory Perception expert (A) is used only for sound-related content, including acoustic events, audible properties, and onomatopoeic expressions.

The prompt enforces a single-label decision for each short token sequence and requires the output to be a JSON dictionary. This format makes the resulting labels easy to parse and align with the corresponding token spans during dataset construction.

Prompt for Expert-Label Annotation

```
I am training a mixture-of-experts (MoE) model that
     routes tokens individually (token-level routing)
     to specialized experts. The model includes six
     experts, each reflecting a cognitive domain.
     Below is a concise explanation of each expert's
     responsibility, designed to help assign short
     token sequences to the most appropriate expert.

- Language Network (LN): Handles grammar, syntax, and
     sentence fluency. This expert is best for basic
     function words and structure-related tokens.
  Example tokens: articles ("a", "the"), auxiliary
     verbs ("is", "was"), prepositions ("in", "on"),
     sentence connectors ("and", "but").

- Multiple Demand Network (MD): Specializes in
     structured reasoning, logic, arithmetic, and
     cause-effect relationships.
  Example tokens: logical markers ("because", "
     therefore", "if...then"), comparative reasoning
     ("more than", "less likely"), numbers or steps.

- Theory of Mind Network (ToM): Focuses on emotions,
     intentions, social behavior, and interpersonal
     dynamics.
  Example tokens: emotional states ("happy", "angry",
     "worried"), social roles ("mother", "friend"),
     mental states ("thinks", "wants to").

- Default Mode Network (DMN): Encodes abstract
     concepts and semantic entities (category/identity
      labels) rather than raw perceptual attributes.
     Use DMN for "what it is" at a conceptual level,
     including entity names, object categories, and
     higher-level scene concepts.
  Examples: entities/categories ("apple", "cat", "
     person", "car", "Eiffel Tower"), abstract
     concepts ("restaurant", "airport", "birthday"),
     inferred scene context ("at home", "in a
     classroom"), cultural/historical references.

- Visual Perception Expert (V): Responsible for
     directly observable, low-level visual attributes
     and geometry, i.e., "what it looks like" before
     naming or semantic identification. Use V for
     visual properties that can be directly seen (
     appearance, layout, and spatial relations), not
     for entity/category labels.
  Examples: brightness/illumination ("bright", "dark"),
      colors ("red", "blue"), shapes ("round", "
     rectangular"), textures/patterns ("striped", "
     shiny"), size ("large", "small"), visible parts
     as appearance cues ("eyes", "mouth", "face
     outline"), spatial relations ("left", "under", "
     in front of").

- Auditory Perception Expert (A): Responsible for
     audible features and acoustic content that would
     be perceived by hearing.
  Example tokens: sounds ("barking", "music", "
     applause"), acoustic properties ("loud", "quiet")
```

Prompt for Expert-Label Annotation, continued

```
     , audio events ("footsteps", "thunder"),
     onomatopoeia.

Given a model-generated description of a scene (from
     an image or audio clip), assign each short token
     sequence to exactly one expert (LN, MD, ToM, DMN,
      V, A), according to its content and purpose.

Output your answer as a JSON dictionary in the format:
{
  "sequence_1": "V",
  "sequence_2": "LN",
  "sequence_3": "A"
}

Context Type
This text describes a {modality} scene.

Description
{description}

Token Sequences
Label the following token sequences by index. Use JSON
      keys "1", "2", ... (string indices) mapping to
     one of: {allowed_experts}.
{sequences}
```

D Qualitative Examples of Expert Specialization

Figure 8 provides qualitative examples of BrainMoE's expert specialization across text-only, image-grounded, and audio-grounded scenarios. For each prompt, different experts generate responses that emphasize distinct functional aspects of the same input. In the text-only case, experts focus on linguistic formulation, logical repair, social interpretation, and commonsense plausibility. In the image-grounded example, the Vision expert captures visual evidence such as the traffic sign, while cognitive experts provide rule-based, safety-oriented, and commonsense interpretations. In the audio-grounded example, the Auditory expert identifies prosodic cues such as loudness and sharp intonation, whereas the Social, Logic, and Language experts infer interpersonal intent, response strategy, and appropriate wording. These examples illustrate that BrainMoE does not rely on a single entangled representation for all tasks; instead, its experts provide complementary sensory and cognitive evidence, supporting interpretable and modular multimodal reasoning.

E vLLM server

During testing, we deploy the trained model using vLLM. The model service is exposed through a standardized HTTP/OpenAI-compatible interface, supports long-context inference, and can be efficiently scaled in a multi-GPU environment. For each evaluation round, the client maps the constructed query set into requests and sends them to the vLLM server.

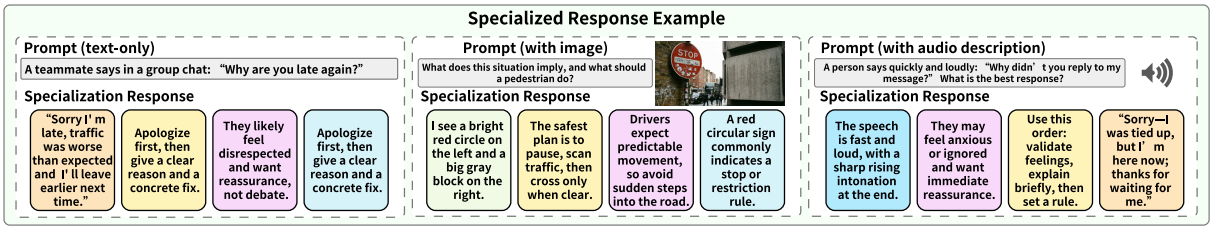

Figure 8: Expert specialization of BrainMoE. Representative text-only, image-grounded, and audio-grounded examples show that different experts produce complementary specialized responses, reflecting linguistic, logical, social, commonsense, visual, and auditory functional specialization.

After receiving the responses, the client writes them into the interaction history, thereby supporting the parallel inference pipeline from information acquisition to final output decision.

vLLM Serving Command

```
vllm serve ${MODEL_PATH} \
    --port ${PORT} \
    --host :: \
    --gpu-memory-utilization 0.8 \
    --max-model-len 32768 \
    --tensor-parallel-size 8 \
    --served-model-name expert \
    --trust-remote-code \
    --disable-log-requests
```