Unveiling Multimodal Processing: Exploring Activation Patterns in Multimodal LLMs for Interpretability and Efficiency

Chuan Wu[1], Meng Su[3], Youxuan Fang[3], Shaolin Zhu[2]*
[1]School of New Media and Communication, Tianjin University, Tianjin, China
[2]College of Intelligence and Computing, Tianjin University, Tianjin, China
[3] China Mobile Information Technology Co., Ltd., Shenzhen, China
{wuchuan, zhushaolin}@tju.edu.cn
{sumeng, fangyouxuan}@chinamobile.com

Abstract

Recent Multimodal Large Language Models (MLLMs) have achieved remarkable advancements, yet their internal mechanisms for concurrently processing diverse modalities like text, image, and audio remain largely opaque. In this paper, we propose a methodology to convert dense MLLMs into fine-grained Mixture-of-Experts (MoE) architectures. This allows us to visually investigate their multimodal activation patterns through expert activation frequency heatmaps. Conducting comprehensive experiments on representative MLLMs, we analyze the similarities and differences in internal neuron activations when handling distinct modalities. Specifically, we examine the distribution of high-frequency activated experts, the distinct roles of high-frequency (e.g., fundamental logic) and low-frequency (e.g., domain-specific concepts) multimodal shared experts, and the prevalence and localization of modality-specific experts. Furthermore, we explore leveraging these discovered activation discrepancies to guide sparse activation and model pruning. Experimental results demonstrate that our approach substantially outperforms random expert pruning and can achieve comparable or even superior performance to the original unpruned models while utilizing significantly fewer active parameters. Our work not only sheds light on the multimodal processing mechanisms within MLLMs but also provides a practical pathway toward developing more interpretable and efficient multimodal systems.

1 Introduction

Large Language Models (LLMs) have demonstrated extraordinary capabilities in processing and generating human language, catalyzing a new era in natural language processing (Hadi et al., 2023; Zhu et al., 2024b). The frontier has rapidly expanded towards multimodal LLMs (MLLMs), which integrate capabilities for understanding and reasoning across diverse modalities such as text, images, and audio (Li et al., 2024c; Liu et al., 2024a). These models, often architected by coupling powerful visual and/or audio encoders with a pre-trained LLM backbone, have achieved remarkable success on a wide array of multimodal tasks, from visual question answering to image captioning and audio-grounded dialogue (Wang et al., 2024; Wu et al., 2024; Li et al., 2025).

However, despite their impressive performance, a fundamental question remains largely unanswered: how do these MLLMs process and integrate information from disparate modalities within a unified architectural framework? While an LLM's core is trained on text, its adaptation to handle visual or auditory signals introduces new complexities. Previous studies in neuroscience suggest that while the multilingual LLMs exhibit considerable overlap in regions processing different languages, discernible specializations also exist (Zhu et al., 2024a; Tang et al., 2024). Analogously, we hypothesize that MLLMs might develop both modality-agnostic (shared) computational pathways and modality-specific ones. Uncovering these internal mechanisms is crucial not only for advancing our theoretical understanding but also for addressing practical challenges. MLLMs, inheriting the scale of their LLM parents, are computationally intensive, making deployment and inference costly (Caffagni et al., 2024). A deeper insight into their internal workings could pave the way for more efficient model designs, targeted pruning strategies, and improved interpretability.

Currently, we lack an intuitive, fine-grained understanding of the internal neuron-level activity within MLLMs as they process different modalities—the system largely remains a "black box." This opacity hinders our ability to diagnose failures, improve robustness, and optimize performance efficiently. Therefore, we aim to investigate the distinct and shared internal neuron activation patterns

* Corresponding Author.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 9005–9016

of MLLMs when confronted with inputs from various modalities. We term these phenomena "multimodal activation patterns."

To demystify this black box and gain an intuitive understanding, we devise a methodology inspired by recent explorations in multilingual LLMs (Liu et al., 2024c). Our approach involves converting standard, dense MLLMs into fine-grained Mixture-of-Experts (MoE) architectures (Zhang et al., 2021). This transformation allows us to treat groups of neurons as "experts" and subsequently calculate their activation frequencies when processing data from different modalities (e.g., text-only, image-only, audio-only, or combined inputs). These frequencies are then visualized as heatmaps, offering a global view of expert utilization across layers and modalities.

Through comprehensive experiments on leading MLLMs, we analyze: (1) The distribution of high-frequency activated experts across different modalities. (2) The existence and functional roles of multimodal shared experts—those frequently activated regardless of input modality, potentially handling core reasoning or abstract concepts. We further differentiate these by activation frequency, hypothesizing that high-frequency shared experts manage fundamental cross-modal logic, while lower-frequency shared experts might capture more nuanced, less common inter-modal abstractions. (3) The characteristics and distribution of modality-specific experts—those predominantly activated by a single modality (e.g., visual experts, textual experts, auditory experts). Furthermore, we explore their practical application in guiding sparse activation and model pruning. By identifying which experts are critical for specific modalities or general reasoning, we propose methods to selectively activate or retain only the most relevant experts during inference. This aims to significantly reduce computational load while preserving, or even in some cases enhancing, model performance on specific tasks. Our preliminary findings indicate that this informed approach can substantially outperform random pruning baselines. In this work, we make the following contributions:

- We introduce a novel methodology to convert and visualize the internal multimodal activation patterns of MLLMs by transforming them into MoE structures and analyzing expert activation frequencies.
- We demonstrate that these identified activation patterns can be effectively leveraged to guide sparse activation and model pruning, leading to significant computational savings with minimal performance degradation, and sometimes even improvements.
- Our findings offer new insights into the internal processing mechanisms of MLLMs, contributing to a better understanding of how these complex models handle and integrate multimodal information, and paving the way for more interpretable and efficient MLLM architectures.

2 Related Work

Our work builds upon several interconnected lines of research: understanding internal mechanisms of large language models, the development and analysis of MLLMs, and techniques for model compression and efficient inference, particularly through sparsity.

The quest to understand the inner workings of LLMs has gained significant traction. Much of this research has focused on monolingual or multilingual textual LLMs. For instance, studies have investigated how LLMs represent linguistic structures (Hewitt and Manning, 2019; Davies and Khakzar, 2024), store factual knowledge (Meng et al., 2022; Sharma et al., 2024), and perform reasoning (Wei et al., 2022). More recently, inspired by neuroscience, researchers have begun to explore language-specific neurons or expert-like structures within LLMs (Kojima et al., 2024). Liu et al. (2024c) provided a direct inspiration by converting dense LLMs into MoE architectures to study multilingual activation patterns, demonstrating the existence of language-specific and shared experts. For MLLMs, while their architectures combining vision/audio encoders with LLMs are well-documented (Li et al., 2024c; Zhu et al., 2023), fine-grained analysis of how different modalities are processed and integrated at the neuron or expert level is still nascent. Early work has explored cross-modal attention mechanisms (Lu et al., 2019) or attempted to localize concepts across modalities (Goh et al., 2021), but a systematic, layer-wise, and expert-level understanding of activation patterns across multiple modalities, akin to what we propose, remains less explored. Our work extends the MoE-based activation analysis from the multilingual to the multimodal domain, aiming to

uncover similar notions of shared and modality-specific computational units.

MoEs (Fedus et al., 2022) have emerged as a promising approach to scale up model capacity while keeping computational costs manageable by sparsely activating only a subset of "experts" per input. While initially prominent in LLMs, their application and analysis in the multimodal context are growing. Some MLLMs have started to incorporate MoE layers explicitly in their design for efficiency (Lin et al., 2024). However, much of the existing MLLM landscape still relies on dense architectures. Zhang et al. (2021) demonstrated that pre-trained dense Transformer FFN layers can be post-hoc converted into MoE structures without significant performance loss, providing a powerful tool for analysis and potential efficiency gains. Our methodology leverages this MoEfication concept, not primarily for pre-training efficient MLLMs, but as an analytical lens to decompose existing dense MLLMs and study their internal multimodal specialization. Subsequently, the identifying frequently or infrequently activated experts per modality can inform strategies for sparse activation during inference or targeted pruning (Heurtel-Depeiges et al., 2024; Ding et al., 2023). Unlike works that design MoE MLLMs from scratch, we focus on understanding and re-purposing existing dense models.

3 Exploring Multimodal Activation Patterns in MLLMs

To investigate how MLLMs internally process and integrate information from diverse modalities, we propose a methodology centered around converting dense MLLM architectures into fine-grained Mixture-of-Experts (MoE) structures. This allows us to analyze the activation patterns of these "experts" in response to unimodal and multimodal inputs.

3.1 Expert Construction in Multimodal Architectures

Our first step is to transform the feed-forward network (FFN) layers within the MLLM's backbone into distinct experts. MLLMs like Llava-NeXT (Liu et al., 2024b) and Qwen-Omni (Yang et al., 2025) often employ LLM backbones whose FFNs consist of up-projection, gate-projection, and down-projection layers.

Modality-Aware Parameter Clustering. MLLMs process embeddings that can originate from text, projected visual features, or projected auditory features. While the core LLM processes these as sequences of vectors, the *origin* and *nature* of these vectors differ significantly. We hypothesize that neurons within FFNs might specialize not just based on abstract features but also subtly influenced by the statistical properties of embeddings derived from different modalities. Therefore, for expert construction, we adopt a parameter clustering approach (Zhang et al., 2021) but with considerations for multimodal processing. We perform balanced K-Means clustering (Malinen and Fränti, 2014) on the parameters of the up-projection layer of each FFN, dividing it into a predefined number of clusters. The neurons and corresponding parameters in the down-projection layers are then grouped according to the up-projection clustering. This creates fine-grained "experts" within each FFN layer. The key distinction from a purely textual MoEfication is that these experts will subsequently be evaluated based on their activation by inputs derived from distinct modalities.

3.2 Cross-Layer Expert Selection for Multimodal Inputs

Shallow layers might handle low-level unimodal features, while deeper layers integrate information and perform abstract reasoning. To capture this, and acknowledging the direct incomparability of raw activation magnitudes across different FFN layers, we extend a cross-layer expert selection strategy for multimodal contexts.

Modality-Specific Activation Scoring and Normalization. For each input token, we calculate an activation score for every expert in every FFN layer. This score is the sum of the activation values of all neurons within that expert *before* the down-projection layer. Crucially, to enable fair comparison and selection across all layers for a given input token, regardless of its originating modality, we perform a Z-score normalization of these expert scores *within each FFN layer*.

Global Expert Ranking and Selection. After layer-wise normalization, we rank all experts across all FFN layers based on their normalized scores for the current input token. We then select the top-K% of experts as the "activated experts" for that specific token. The activation count for these selected experts is incremented by 1. This cross-layer approach allows us to identify experts that are significantly active relative to their peers

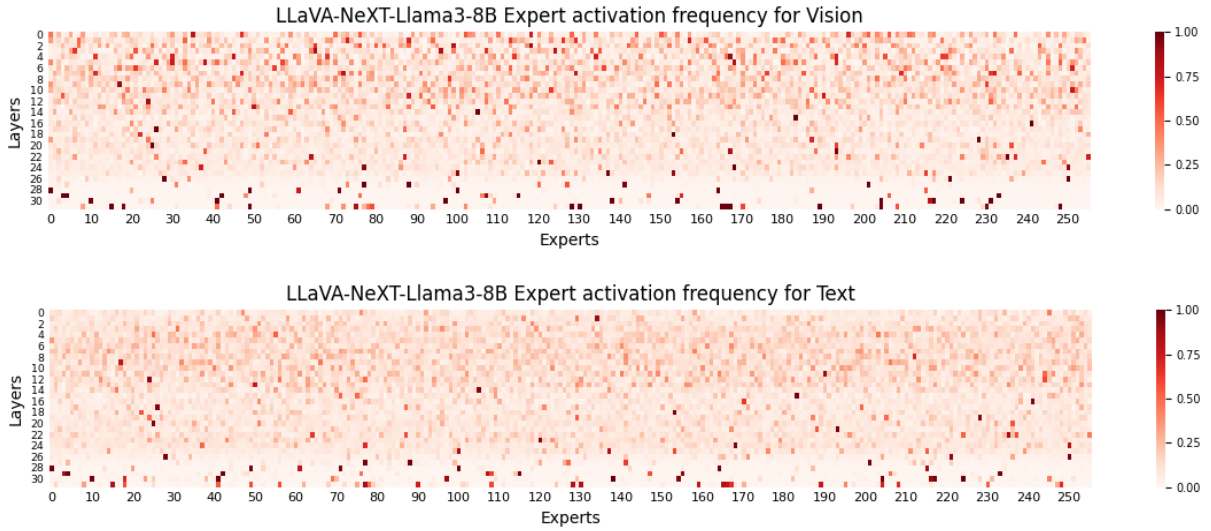

Figure 1: Heatmaps of activation patterns for LLaVA-NeXT-Llama3-8B in Text and Visual modal. Each heatmap is 32*256 (number of layers * number of experts), with darker colors indicating higher activation frequencies.

within their own layer, regardless of the absolute magnitude differences between layers, and whether the token originated from text, vision, or audio.

3.3 Quantifying and Visualizing Multimodal Expert Activation Patterns

After processing a large corpus of data for each modality (and their combinations), we calculate the activation frequency for each expert:

$$\text{ActFreq} = \frac{\text{ActCount}(\text{Expert}_{i,j}, \text{Modality}_m)}{\text{TotalTokens}(\text{Modality}_m)} \quad (1)$$

where $\text{Expert}_{i,j}$ is the j-th expert in the i-th FFN layer, and Modality_m represents a specific input condition (e.g., image-only, text-only, audio-only).

This results in an $L \times N_e$ activation frequency matrix for each modality (where L is the number of FFN layers and N_e is the number of experts per layer). We then visualize these matrices as heatmaps. These heatmaps are central to our analysis, as they allow us to directly observe:

- **Overall Sparsity and Layer-wise Trends:** How sparsely are experts activated for different modalities? Are there layers that are consistently more or less active for images versus text versus audio?

- **Modality-Specific Hotspots:** Are there specific experts or groups of experts that show significantly higher activation for one modality compared to others?

By comparing these heatmaps across different modalities, we can identify patterns of specialization and sharing, providing a granular view into the MLLM's internal processing landscape. This multimodal-centric approach to expert construction, selection, and visualization is key to differentiating our work from purely language-focused analyses.

4 Experiments and Analysis

4.1 Experimental Settings

Models. We conduct experiments on two SOTA MLLMs, LLaVA and Qwen, to investigate their activation patterns across different modalities. Specifically, LLaVA-NeXT-Llama3-8B (Li et al., 2024a) supports text and image modalities, while Qwen2.5-Omni-7B (Jin Xu, 2025) is an all-modal model. This setup enables a comprehensive analysis of modality-specific activation behaviors across model architectures.

Data. For the text modality, we use the MMLU (Massive Multitask Language Understanding) benchmark (Hendrycks et al., 2021b,a), which spans a wide range of subjects including elementary mathematics, history, computer science, law, ethics, and medicine. For the visual modality, we adopt the LLaVA-OneVision-Data corpus (Li et al., 2024b), encompassing both general images and domain-specific scenarios such as documents, charts, screenshots, mathematical reasoning, language comprehension, and OCR tasks. For the au-

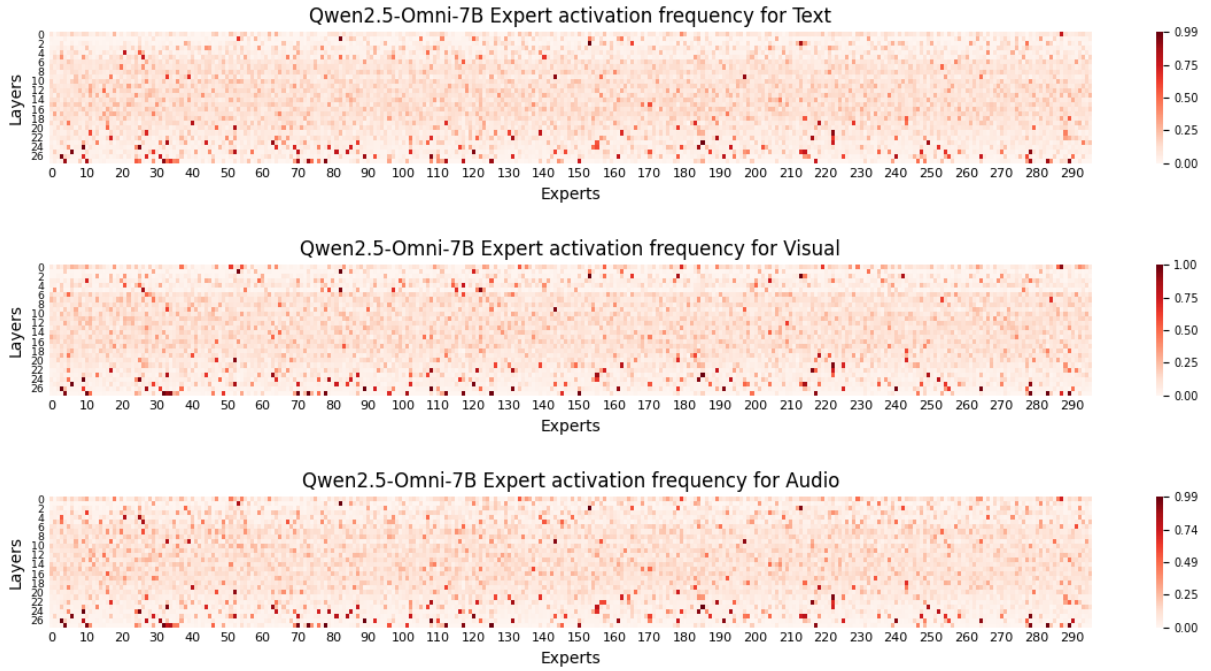

Figure 2: Heatmaps of activation patterns for Qwen2.5-Omni-7B in Text, Visual and Audio modal. Each heatmap is 28*296 (number of layers * number of experts), with darker colors indicating higher activation frequencies.

dio modality, we utilize the LibriSpeech dataset (Panayotov et al., 2015), consisting of approximately 1,000 hours of read English speech. For each modality, we test its activation pattern using 10,000 samples. For image and audio data, we only use the images and audio clips without any text instructions.

4.2 Multimodal Activation Patterns of MLLMs

Our analysis of the activation heatmaps reveals distinct patterns both between models (LLaVA-NeXT vs. Qwen-Omni) and across modalities within each model.

As Figure 1, the LLaVA-NeXT model demonstrates a pronounced difference in activation patterns between visual and textual modalities, particularly in the shallower layers. For visual inputs, the initial FFN layers (e.g., layers 0-10) exhibit significantly higher and more broadly distributed expert activation compared to textual inputs. This suggests a substantial allocation of neural resources for processing raw visual features early in the network. As information propagates to middle layers, the activation frequencies become more comparable and relatively stable across both modalities, with fewer discernible hotspots. Interestingly, in the deeper layers (e.g., layers 25-31), while overall sparsity increases, the *pattern* of activated experts for visual and textual inputs shows greater similarity. This convergence indicate that these later layers are primarily involved in abstract reasoning and response generation, integrating information from the processed visual features into the language modeling stream. The sparsity in deeper layers is consistent with findings in LLMs where later layers often show more specialized roles.

The Qwen2.5-Omni-7B model, an all-modal architecture, presents a different activation landscape, as shown in Figure 2. A striking feature in the shallow layers (e.g., layers 0-5) is the appearance of prominent "white stripes" – contiguous regions of uniformly low expert activation. Notably, the expert indices of these low-activation stripes vary depending on the input modality. For instance, visual inputs trigger very limited activation among experts in layers 1, 2, and 5 at specific index ranges, while audio or text inputs might activate these same experts or show different low-activation stripes. This strongly suggests that Qwen-Omni employs highly specialized, modality-specific pathways in its earliest FFN layers, potentially routing different types of unimodal information through distinct subsets of experts before extensive cross-modal fusion. This early specialization could be a mechanism for efficient unimodal feature extraction. Beyond these

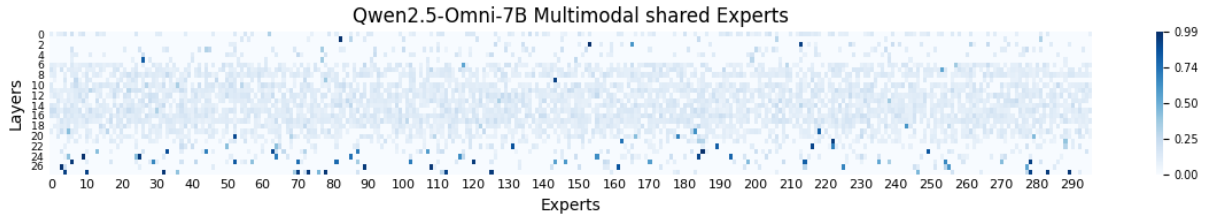

Figure 3: The heatmaps of Multimodal shared experts for Qwen2.5-Omni-7B model. The color shade of each cell indicates the activation frequencies.

initial layers, activation patterns become more uniform and distributed, with fewer stark differences between modalities until the deeper layers, where sparsity again increases, similar to LLaVA-NeXT.

4.3 Multimodal shared experts

We define such units as *multimodal shared experts*, specifically those whose activation frequency surpasses a predefined threshold $\theta_s = 0.05$ for all relevant input modalities (text, vision, and audio for Qwen-Omni; text and vision for LLaVA-NeXT), and where the maximum pairwise difference in activation frequencies between any two modalities does not exceed a small tolerance $\delta_s \leq 0.1$. To further refine our understanding, we categorize these shared experts based on their average cross-modal activation levels into High-Frequency Shared Experts (HF-SEs, average activation frequency $> \theta_{hf} = 0.3$) and Low-Frequency Shared Experts (LF-SEs). As illustrated in Figure 3, it indicates that HF-SEs are few in number and predominantly concentrated in the deeper layers of the network. This strongly suggests they form the core, modality-agnostic reasoning and output generation backbone of the model, playing an indispensable role in final decision-making or generation stages due to their consistent high-frequency activation across diverse inputs.

In contrast, LF-SEs are significantly more numerous, and their distribution patterns unveil model-specific information integration strategies. For instance, in Qwen-Omni, LF-SEs are primarily located in the middle layers, with a near absence of shared experts in the shallowest layers, aligning with this model's strong early-stage modality-specific processing characteristics. This suggests that the middle layers serve as a key zone for integrating unimodal features and representing abstract cross-modal concepts. Broadly, the prevalence of LF-SEs indicates that beyond core reasoning units, MLLMs possess a larger, more diverse set of experts for flexibly handling specific types of cross-modal abstraction and integration tasks as needed. The identification and characterization of these shared experts provide crucial insights into the general-purpose versus specialized computational mechanisms within MLLMs

4.4 Modality Specific Experts

In addition to shared computational units, MLLMs employ *modality-specific experts* dedicated to processing unique unimodal information, defined as experts with significantly higher activation (e.g., by at least 0.1) for one modality over others. Our analysis of Qwen2.5-Omni-7B, visualized through pairwise modality comparisons in Figure 4, reveals a clear hierarchy: vision-specific experts are the most numerous, followed by audio-specific and text-specific experts. This prevalence of vision experts, evident across comparisons in Figure 4, likely stems from the higher token count typically generated from image inputs, demanding greater neural resources. Similar trends in vision expert dominance are noted in LLaVA-NeXT (Appendix Figure 6).

The layer-wise distribution of these specific experts, as inferred from Figure 4 for Qwen-Omni, indicates that vision- and audio-specific experts are notably concentrated in shallow and middle layers, forming dedicated pathways for initial unimodal feature extraction. Audio-specific experts in Qwen-Omni also maintain a significant presence into deeper layers, reflecting its all-modal architecture and audio generation capabilities. This observed division of labor underscores the role of modality-specific experts in performing fine-grained unimodal processing, particularly in early network stages, thereby preparing representations for subsequent integration and reasoning by shared experts.

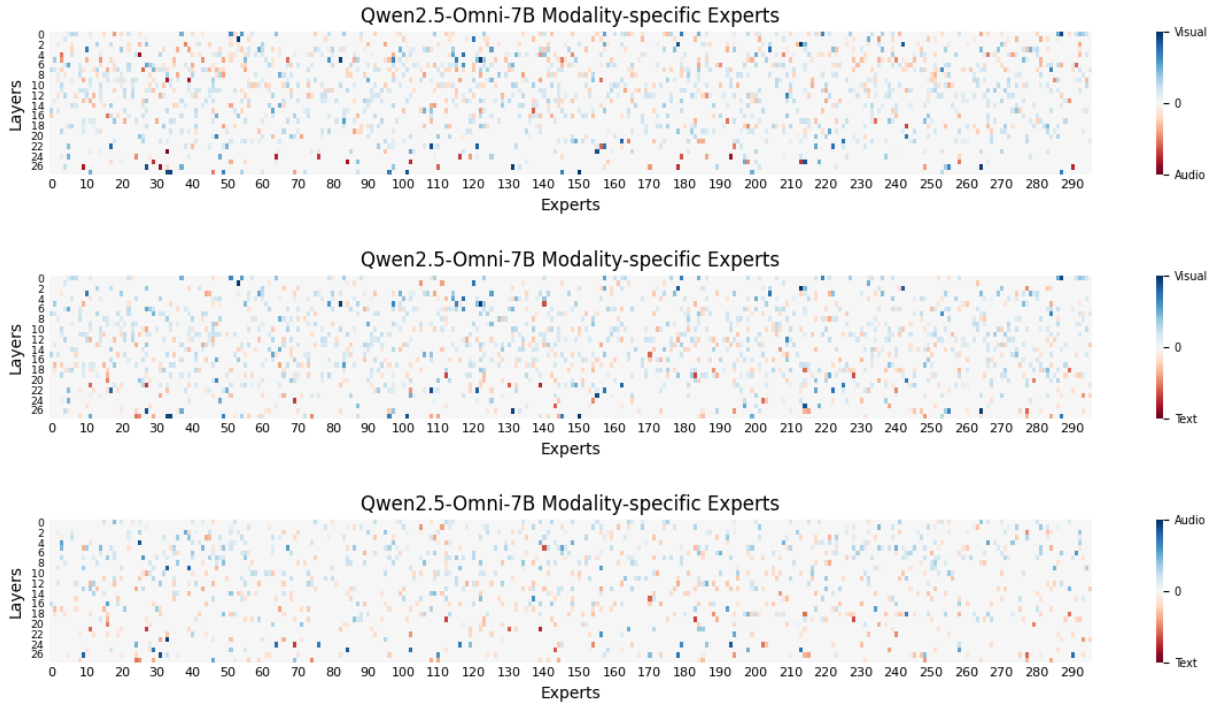

Figure 4: The heatmaps of Modality specific experts for Qwen2.5-Omni-7B model. The color shade of each cell indicates the activation frequencies.

4.5 Can Expert Activation Frequencies Guide Sparse Activation and Model Pruning?

The distinct multimodal activation patterns and functional specializations of experts, identified in previous sections, present opportunities for enhancing MLLM efficiency. We investigate whether these insights can guide sparse activation and targeted model pruning to reduce computational overhead while maintaining robust performance.

Evaluation. We conduct experiments on tasks across different modalities. For text-based tasks, we use CSQA (Talmor et al., 2019) (commonsense question answering) and GSM-8K (Cobbe et al., 2021) (grade-school math problems). For vision tasks, we adopt MMStar (Chen et al., 2024) (vision-indispensable problems) and POPE (Li et al., 2023) (object hallucination). For audio tasks, we employ VocalSound (Gong et al., 2022) (vocal sound classification) and MELD (Poria et al., 2019) (speech emotion recognition). We use accuracy and F1 score as evaluation metrics.

4.5.1 Pruning Based on Expert Functions

This strategy involves selectively activating only those experts deemed functionally relevant—combinations of multimodal shared (reasoning) experts and modality-specific experts—for a given task within the Qwen2.5-Omni-7B model. This targeted activation aims to reduce FFN parameter usage (by 10-20%, translating to ∼9-17% FLOPs reduction) without significant performance loss.

Experimental results for Qwen2.5-Omni-7B across text, visual (Table 1) and audio tasks (Table 2) demonstrate the efficacy of this approach. On text and visual tasks (Table 1), configurations such as "Reasoning+text+visual" (activating shared, text-specific, and vision-specific experts) achieve performance that is remarkably close to, or even surpasses, the original unpruned model. For instance, on CSQA, this configuration yields an accuracy of 67.9, matching the 'Origin' model, and on GSM8K, it achieves 86.2, close to the original 87.6, while utilizing only 92.2% of FFN parameters. Similarly, for visual tasks like MMstar and POPE, this comprehensive expert set maintains strong performance (e.g., 64.0 vs. 63.7 on MMstar; 74.5 vs. 73.8 on POPE). For Qwen-Omni's audio tasks (Table 2), activating "Reasoning+text+audio" experts also yields compelling results, notably on VocalSound (94.2 vs. 93.4 for Origin) and MELD (56.7 vs. 56.3 for Origin). A critical observation across all task types is the performance degradation when high-frequency shared ("Reasoning (w/o

Activate Patterns	Parameters	CSQA		GSM8K		MMstar		POPE	
		Random	Experts	Random	Experts	Random	Experts	Random	Experts
Reasoning(w/o high)	80.9%	39.8±6.8	54.6	61.7±2.8	67.3	34.9±6.3	46.4	50.4±1.5	55.4
Reasoning	82.0%	42.5±5.3	61.5	63.8±2.6	78.5	37.4±4.5	55.2	52.9±1.6	61.5
Reasoning+text	85.9%	46.4±4.7	67.7	70.4±1.7	84.3	40.7±2.8	59.3	53.5±1.3	65.9
Reasoning+audio	86.3%	49.7±4.2	65.3	73.4±1.6	81.8	43.2±1.1	58.5	56.4±0.7	68.3
Reasoning+text+audio	90.3%	58.2±3.8	67.5	76.5±0.7	84.5	52.2±0.9	61.5	62.4±0.3	70.2
Reasoning+visual	88.3%	53.5±4.6	66.2	74.7±1.4	83.7	48.5±2.3	63.2	59.6±0.9	**74.5**
Reasoning+text+visual	92.2%	61.7±2.4	**67.9**	78.2±0.8	86.2	56.5±1.2	**64.0**	65.7±0.8	74.2
Origin	100%	67.9		87.6		63.7		73.8	

Table 1: Pruning performance of Qwen2.5-Omni-7B on Text tasks (CSQA, GSM8K) and Visual tasks (MMstar, POPE) . The POPE dataset uses F1-score as the metric. Other datasets use Accuracy (%).

Activate Patterns	Parameters	VocalSound		Meld	
		Random	Experts	Random	Experts
Reasoning(w/o high)	80.9%	79.2±1.5	82.9	30.4±2.3	41.5
Reasoning	82.0%	82.3±1.8	87.5	33.3±1.9	46.8
Reasoning+text	85.9%	86.5±1.7	89.7	38.8±1.5	51.3
Reasoning+visual	88.3%	88.2±0.9	91.8	45.4±0.7	53.6
Reasoning+audio	86.3%	87.7±1.5	93.1	42.7±0.8	56.7
Reasoning+text+visual	92.2%	90.4±0.5	92.6	48.4±0.4	55.1
Reasoning+text+audio	90.3%	90.2±0.4	**94.2**	47.6±0.5	**56.7**
Origin	100%	93.4		56.3	

Table 2: Accuracy (%) of Qwen2.5-Omni-7B on Audio task.

Pruning rate	CSQA		GSM8K		MMstar		POPE		Meld		VocalSound	
	global	equal	global	equal	global	equal	global	equal	global	equal	global	equal
70%	46.3	29.7	66.5	53.7	41.8	24.2	51.6	50.8	38.3	30.4	75.9	69.4
80%	59.7	42.7	78.3	60.8	54.6	32.5	64.1	53.7	46.7	37.9	87.6	76.2
90%	66.2	54.2	84.9	76.2	61.3	53.6	70.3	60.3	53.4	46.8	92.3	88.6
Origin	67.9		87.6		63.7		73.8		56.3		93.4	

Table 3: Pruning performance of Qwen2.5-Omni-7B based on frequency sorting. The "Pruning rate" shows the overall pruning ratio. The "global" column shows unequal pruning for each layer. The "equal" column shows equal pruning for each layer.

high)") experts are excluded, underscoring their foundational role. These function-guided strategies consistently and significantly outperform random expert selection at comparable parameter counts. Similar positive trends for LLaVA-NeXT are detailed in Appendix Tables 4.

4.5.2 Pruning Based on Frequency Sorting

We also explore pruning Qwen2.5-Omni-7B by retaining only the top $n\%$ of experts sorted by their activation frequency, either applied "equally" per layer or "globally" across the model. The results are presented in Table 3.

We can find that the superiority of "global" (unequal per-layer) pruning over "equal" (uniform per-layer) pruning for Qwen2.5-Omni-7B. For example, on the CSQA task, retaining 70% of experts via "global" pruning yields an accuracy of 46.3, substantially better than the 29.7 achieved by "equal" pruning. This pattern holds across different tasks and pruning rates (e.g., on MMstar: 54.6 with "global" vs. 32.5 with "equal" at 80% retention; on MELD for audio: 53.4 with "global" vs. 46.8 with "equal" at 90% retention.) This disparity validates our earlier finding that expert activation sparsity varies significantly across layers in Qwen-Omni. "Global" pruning respects these intrinsic layer-wise differences, leading to more effective compression. This strongly suggests that layer-differentiated pruning rates, informed by activation characteristics, are more effective than uniform approaches for Qwen2.5-Omni-7B, a finding also echoed by experiments on LLaVA-NeXT (see Appendix Table 5).

5 Conclusion

In this paper, we systematically investigated the internal multimodal activation patterns of MLLMs by transforming dense architectures into fine-grained MoE structures and visualizing expert activation frequencies. Our analysis revealed that a small core of high-frequency shared experts, concentrated in deeper layers, appears crucial for fundamental cross-modal reasoning Large contingent of low-frequency shared experts, often in middle layers, likely handles more nuanced multimodal integration and abstract concept representation. We also characterized modality-specific experts, predominantly vision-focused and active in shallower layers, underscoring early-stage unimodal processing. Crucially, we demonstrated that these identified activation patterns can effectively guide sparse activation and model pruning strategies. Our function-based and frequency-sorted pruning methods significantly outperformed random baselines and often matched or even surpassed the original model performance with substantially reduced computational costs. These findings not only offer novel insights into the "black box" of MLLM multimodal processing but also provide a practical pathway towards developing more interpretable and efficient MLLMs.

Limitations

Despite achieving some meaningful conclusions in our research, there are still some limitations.

The limitations of Experimental Data. Although we conduct comprehensive experiments on two representative MLLMs, these models may not fully capture the diversity of all multimodal tasks and scenarios. For instance, certain domain-specific tasks—such as medical image-text integration or multilingual speech-text interaction—may require more specialized model architectures and training strategies. Additionally, the diversity and scale of the experimental datasets may limit our ability to fully uncover the internal mechanisms of the models.

The limitations in Cross-Modal Reasoning. Our study primarily focuses on analyzing the internal multimodal activation patterns and their impact on model performance. However, the exploration of deep cross-modal reasoning mechanisms remains limited. Questions such as how models transfer and integrate information across modalities, and how they perform effective reasoning in complex multimodal scenarios, require further investigation.

In future work, we plan to extend our experiments to a broader range of models and datasets, covering more domains and complex multimodal interaction settings. We also aim to explore cross-modal reasoning mechanisms that better align with human cognitive processes.

Ethics Statement

This work was conducted in strict compliance with the ACL Ethics Policy. All datasets and multimodal large language models (MLLMs) used for analysis are publicly available. Furthermore, our work aims to explore multimodal processing mechanisms inside MLLMs. We do not foresee any negative ethical impacts arising from our work.

Acknowledgements

The present research was supported by the National Key Research and Development Program (Grant No.2023YFE0116400), National Natural Science Foundation of China Youth Fund (Grant No.62306210) and the Tianjin Natural Science Foundation of Youth Fund (Grant No.23JCQNJC01690). We would like to thank the anonymous reviewers for their insightful comments.

References

Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. 2024. The revolution of multimodal large language models: a survey. *arXiv preprint arXiv:2402.12451*.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and 1 others. 2024. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*.

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

Adam Davies and Ashkan Khakzar. 2024. The cognitive revolution in interpretability: From explaining behavior to interpreting representations and algorithms. *arXiv preprint arXiv:2408.05859*.

Tianyu Ding, Tianyi Chen, Haidong Zhu, Jiachen Jiang, Yiqi Zhong, Jinxin Zhou, Guangzhi Wang, Zhihui Zhu, Ilya Zharkov, and Luming Liang. 2023. The efficiency spectrum of large language models: An algorithmic survey. *arXiv preprint arXiv:2312.00678*.

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.

Gabriel Goh, Nick Cammarata, Chelsea Voss, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, and Chris Olah. 2021. Multimodal neurons in artificial neural networks. *Distill*, 6(3):e30.

Yuan Gong, Jin Yu, and James Glass. 2022. Vocalsound: A dataset for improving human vocal sounds recognition. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 151–155.

Muhammad Usman Hadi, Rizwan Qureshi, Abbas Shah, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Jia Wu, Seyedali Mirjalili, and 1 others. 2023. A survey on large language models: Applications, challenges, limitations, and practical usage. *Authorea Preprints*.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021a. Aligning ai with shared human values. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021b. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

David Heurtel-Depeiges, Anian Ruoss, Joel Veness, and Tim Genewein. 2024. Compression via pre-trained transformers: A study on byte-level multimodal data. *arXiv preprint arXiv:2410.05078*.

John Hewitt and Christopher D Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138.

Jinzheng He Hangrui Hu Ting He Shuai Bai Keqin Chen Jialin Wang Yang Fan Kai Dang Bin Zhang Xiong Wang Yunfei Chu Junyang Lin Jin Xu, Zhifang Guo. 2025. Qwen2.5-omni technical report. *arXiv preprint arXiv:2503.20215*.

Takeshi Kojima, Itsuki Okimura, Yusuke Iwasawa, Hitomi Yanaka, and Yutaka Matsuo. 2024. On the multilingual ability of decoder-based pre-trained language models: Finding and controlling language-specific neurons. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6919–6971. Association for Computational Linguistics.

Bo Li, Kaichen Zhang, Hao Zhang, Dong Guo, Renrui Zhang, Feng Li, Yuanhan Zhang, Ziwei Liu, and Chunyuan Li. 2024a. Llava-next: Stronger llms supercharge multimodal capabilities in the wild.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024b. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Chunyuan Li, Zhe Gan, Zhengyuan Yang, Jianwei Yang, Linjie Li, Lijuan Wang, Jianfeng Gao, and 1 others. 2024c. Multimodal foundation models: From specialists to general-purpose assistants. *Foundations and Trends® in Computer Graphics and Vision*, 16(1-2):1–214.

Junchen Li, Qing Yang, Bojian Jiang, Shaolin Zhu, and Qingxuan Sun. 2025. Lrm-llava: overcoming the modality gap of multilingual large language-vision model for low-resource languages. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'25/IAAI'25/EAAI'25. AAAI Press.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305, Singapore. Association for Computational Linguistics.

Bin Lin, Zhenyu Tang, Yang Ye, Jiaxi Cui, Bin Zhu, Peng Jin, Jinfa Huang, Junwu Zhang, Yatian Pang, Munan Ning, and 1 others. 2024. Moe-llava: Mixture of experts for large vision-language models. *arXiv preprint arXiv:2401.15947*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024b. Llavanext: Improved reasoning, ocr, and world knowledge.

Weize Liu, Yinlong Xu, Hongxia Xu, Jintai Chen, Xuming Hu, and Jian Wu. 2024c. Unraveling babel: Exploring multilingual activation patterns of llms and their applications. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11855–11881.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

Mikko I Malinen and Pasi Fränti. 2014. Balanced k-means for clustering. In *Structural, Syntactic, and Statistical Pattern Recognition: Joint IAPR International Workshop, S+ SSPR 2014, Joensuu, Finland, August 20-22, 2014. Proceedings*, pages 32–41. Springer.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 5206–5210. IEEE.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy. Association for Computational Linguistics.

Arnab Sen Sharma, David Atkinson, and David Bau. 2024. Locating and editing factual associations in mamba. *arXiv preprint arXiv:2404.03646*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Wayne Xin Zhao, Furu Wei, and Ji-Rong Wen. 2024. Language-specific neurons: The key to multilingual capabilities in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5701–5715.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Qinzhuo Wu, Weikai Xu, Wei Liu, Tao Tan, Jianfeng Liu, Ang Li, Jian Luan, Bin Wang, and Shuo Shang. 2024. Mobilevlm: A vision-language model for better intra-and inter-ui understanding. *arXiv preprint arXiv:2409.14818*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Zhengyan Zhang, Yankai Lin, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. 2021. Moefication: Transformer feed-forward layers are mixtures of experts. *arXiv preprint arXiv:2110.01786*.

Shaolin Zhu, Shangjie Li, Yikun Lei, and Deyi Xiong. 2023. PEIT: Bridging the modality gap with pre-trained models for end-to-end image translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13433–13447, Toronto, Canada. Association for Computational Linguistics.

Shaolin Zhu, Leiyu Pan, Bo Li, and Deyi Xiong. 2024a. Landermt: Dectecting and routing language-aware neurons for selectively finetuning llms to machine translation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12135–12148.

Shaolin Zhu, Shaoyang Xu, Haoran Sun, Leiyu Pan, Menglong Cui, Jiangcun Du, Renren Jin, António Branco, Deyi Xiong, and 1 others. 2024b. Multilingual large language models: A systematic survey. *arXiv preprint arXiv:2411.11072*.

A Appendix

A.1 Multimodal shared experts

In Figure 5, we present the distribution of multimodal shared experts in the LLaVA-NeXT-Llama3-8B model.

A.2 Modality specific experts

In Figure 6 , we present the distribution of modality specific experts in the LLaVA-NeXT-Llama3-8B model.

A.3 Pruning results on expert functions

Tabel 4 present the pruning performance of LLaVA-NeXT-Llama3-8B on Text tasks (CSQA, GSM8K) and Visual tasks (MMstar, POPE).

A.4 Pruning results on frequency sorting

Tabel 5 present the pruning performance of LLaVA-NeXT-Llama3-8B on Text and Visual Task.

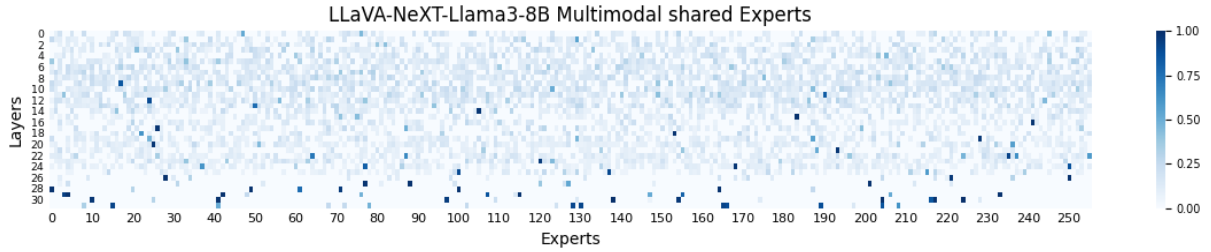

Figure 5: The heatmaps of Multimodal shared experts for LLaVA-NeXT-Llama3-8B. The color shade of each cell indicates the activation frequencies.

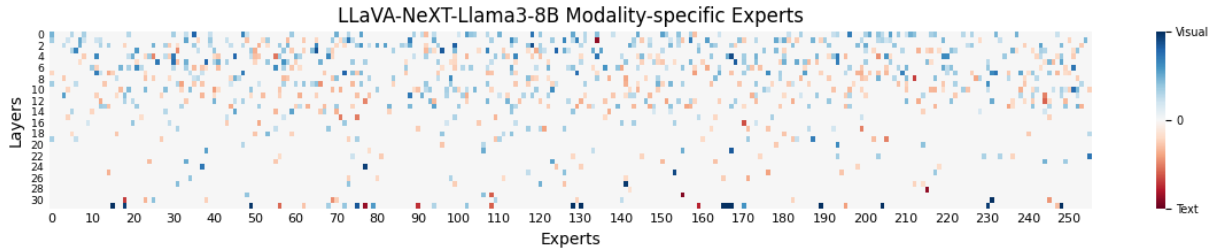

Figure 6: The heatmaps of Modality specific experts for LLaVA-NeXT-Llama3-8B. The color shade of each cell indicates the activation frequencies.

Activate Patterns	Parameters	CSQA		GSM8K		MMstar		POPE	
		Random	Experts	Random	Experts	Random	Experts	Random	Experts
Reasoning(w/o high)	79.8%	43.8±8.3	54.2	53.4±3.1	60.3	27.7±3.9	32.8	56.7±1.6	67.5
Reasoning	82.1%	49.5±6.5	62.8	61.7±2.4	68.6	34.5±2.6	42.1	62.4±0.8	74.7
Reasoning+text	85.7%	57.7±5.3	71.8	65.8±1.7	75.1	37.8±1.7	47.6	71.6±0.5	81.8
Reasoning+visual	87.5%	60.4±2.7	69.4	68.7±1.3	74.7	40.4±0.9	51.5	74.6±0.5	**86.5**
Reasoning+text+visual	91.1%	64.7±0.8	71.9	71.8±0.6	77.4	42.6±0.8	**52.3**	77.4±0.3	86.2
Origin	100%	72.0		77.6		52.1		86.2	

Table 4: Pruning performance of LLaVA-NeXT-Llama3-8B on Text tasks (CSQA, GSM8K) and Visual tasks (MMstar, POPE) . The POPE dataset uses F1-score as the metric. Other datasets use Accuracy (%).

Pruning rate	CSQA		GSM8K		MMstar		POPE	
	global	equal	global	equal	global	equal	global	equal
70%	56.3	42.5	58.7	40.6	30.6	19.3	64.6	50.5
80%	61.7	49.7	67.2	48.7	45.4	28.9	75.2	58.8
90%	69.4	61.2	76.5	62.4	50.8	46.7	84.5	72.3
Origin	72.0		77.6		52.1		86.2	

Table 5: Pruning performance of LLaVA-NeXT-Llama3-8B based on frequency sorting. The “Pruning rate” shows the overall pruning ratio. The “global” column shows unequal pruning for each layer. The “equal” column shows equal pruning for each layer.