# Where Can We Mix? From Atom to Cosmic

Paper Reading by Zhiying Lu

2024.11.19

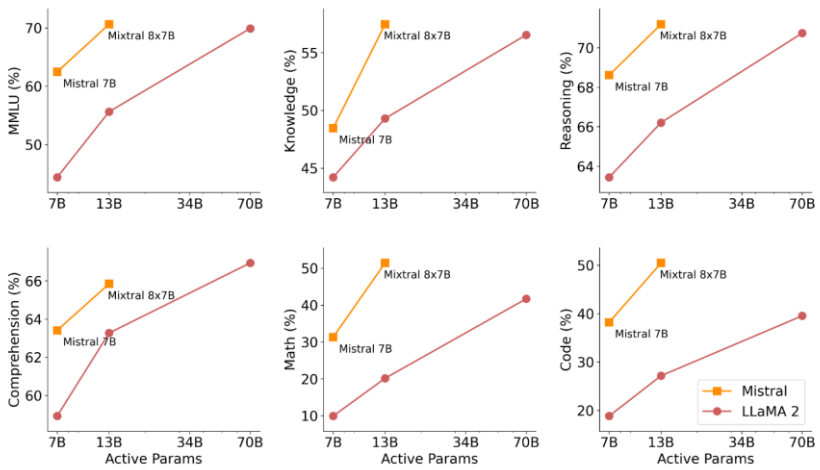
智能多媒体内容计算实验室
Intelligent Multimedia Content Computing Lab

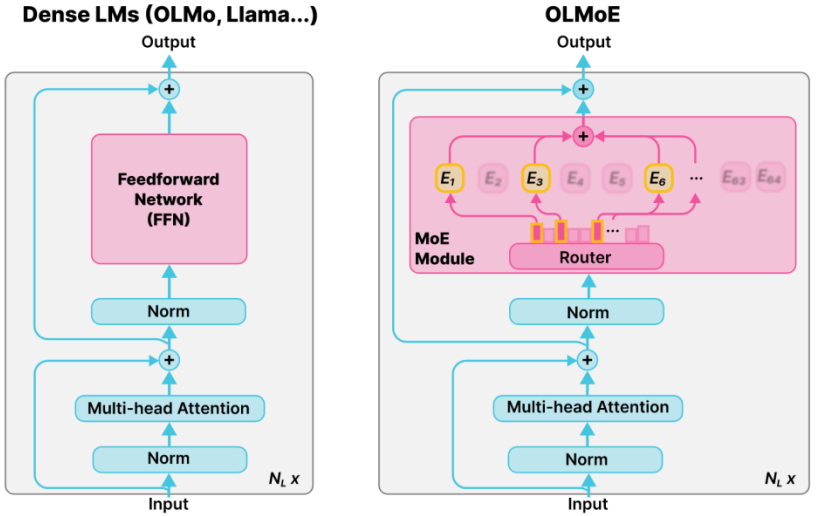
- Standard MoE
- Internal Mixture
- External Mixture
- A New Insight
- Summary

# Mixture-of-Expert (MoE)



- 混合专家架构，一种将多个domain-specific expert融合起来构建统一模型的技术，常用赋予模型multi-task能力
- 在LLM结构中，FFN部分可以被替换为MoE架构，从Dense架构转换为Sparse结构，通过为每个token分配少量的K个expert，实现稀疏激活
- 扩展expert总数，控制激活的总参数量，构建轻量化高精度模型



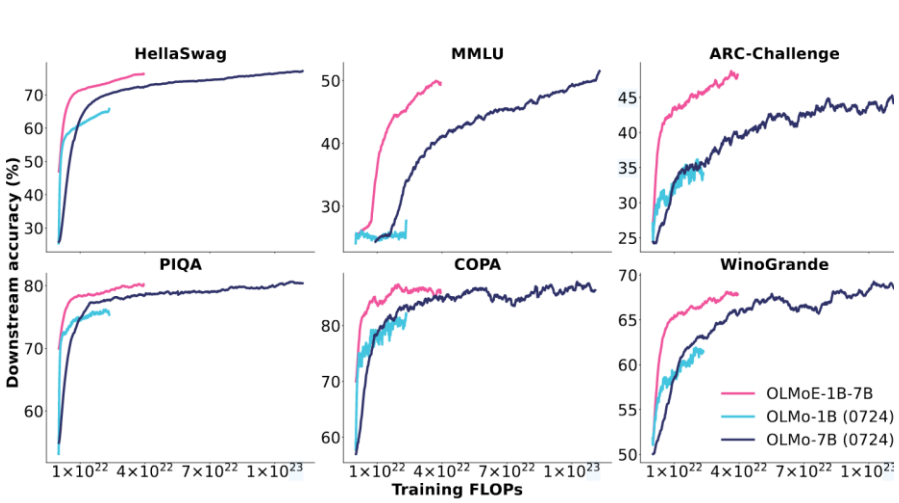

OLMoE: Open Mixture-of-Experts Language Models
Mixtral of Experts

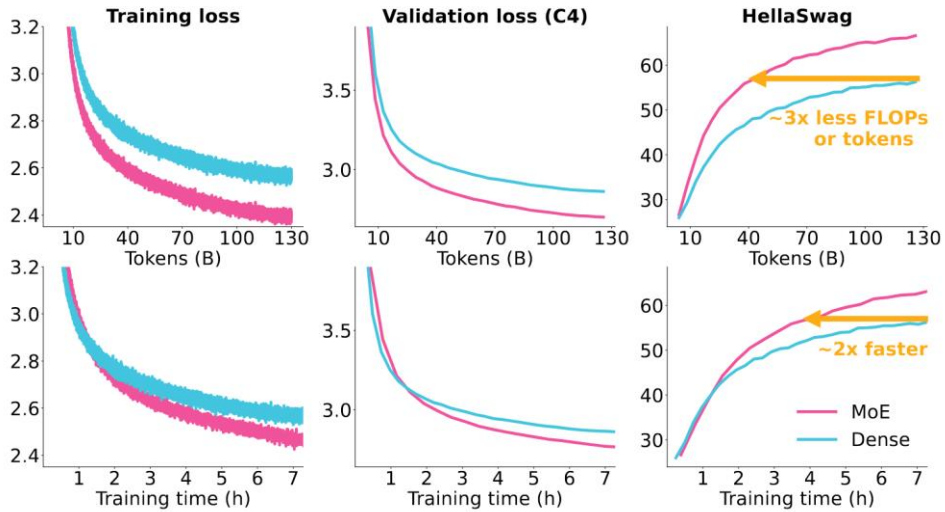
# MoE



- MoE的优势：
  - 更快的收敛速度
  - 更强的模型表征
  - 更丰富的泛化性
  - 更少的推理负担

| | Active params | Open Data | MMLU | Hella-Swag | ARC-Chall. | ARC-Easy | PIQA | Wino-Grande |
|---|---|---|---|---|---|---|---|---|
| LMs with ~7-9B active parameters | | | | | | | | |
| Llama2-7B [181] | 6.7B | ✗ | 46.2 | 78.9 | 54.2 | 84.0 | 77.5 | 71.7 |
| OLMo-7B (0724) [64] | 6.9B | ✓ | 54.9 | 80.5 | 68.0 | 85.7 | 79.3 | 73.2 |
| Mistral-7B [77] | 7.3B | ✗ | 64.0 | 83.0 | 78.6 | 90.8 | 82.8 | 77.9 |
| DCLM-7B [89] | 6.9B | ✓ | 64.4 | 82.3 | 79.8 | 92.3 | 80.1 | 77.3 |
| Llama3.1-8B [50] | 8.0B | ✗ | 66.9 | 81.6 | 79.5 | 91.7 | 81.1 | 76.6 |
| Gemma2-9B [175] | 9.2B | ✗ | **70.6** | **87.3** | **89.5** | **95.5** | **86.1** | **78.8** |
| LMs with ~2-3B active parameters | | | | | | | | |
| OpenMoE-3B-9B [198] | 2.9B | ✓ | 27.4 | 44.4 | 29.3 | 50.6 | 63.3 | 51.9 |
| StableLM-2B [16] | 1.6B | ✗ | 40.4 | 70.3 | 50.6 | 75.3 | 75.6 | 65.8 |
| DeepSeek-3B-16B [39] | 2.9B | ✗ | 45.5 | 80.4 | 53.4 | 82.7 | 80.1 | **73.2** |
| JetMoE-2B-9B [156] | 2.2B | ✗ | 49.1 | **81.7** | 61.4 | 81.9 | 80.3 | 70.7 |
| Gemma2-3B [175] | 2.6B | ✗ | 53.3 | 74.6 | 67.5 | 84.3 | 78.5 | 71.8 |
| Qwen1.5-3B-14B [178] | 2.7B | ✗ | **62.4** | 80.0 | **77.4** | **91.6** | **81.0** | 72.3 |
| LMs with ~1B active parameters | | | | | | | | |
| Pythia-1B [18] | 1.1B | ✓ | 31.1 | 48.0 | 31.4 | 63.4 | 68.9 | 52.7 |
| OLMo-1B (0724) [64] | 1.3B | ✓ | 32.1 | 67.5 | 36.4 | 53.5 | 74.0 | 62.9 |
| TinyLlama-1B [209] | 1.1B | ✓ | 33.6 | 60.8 | 38.1 | 69.5 | 71.7 | 60.1 |
| DCLM-1B [89] | 1.4B | ✓ | 48.5 | 75.1 | 57.6 | 79.5 | 76.6 | 68.1 |
| **OLMoE-1B-7B** | 1.3B | ✓ | **54.1** | **80.0** | **62.1** | **84.2** | **79.8** | **70.2** |



# MoE



- 其关键点在于：
- Expert -- What to be Mixed?
- Router -- How to Mix?

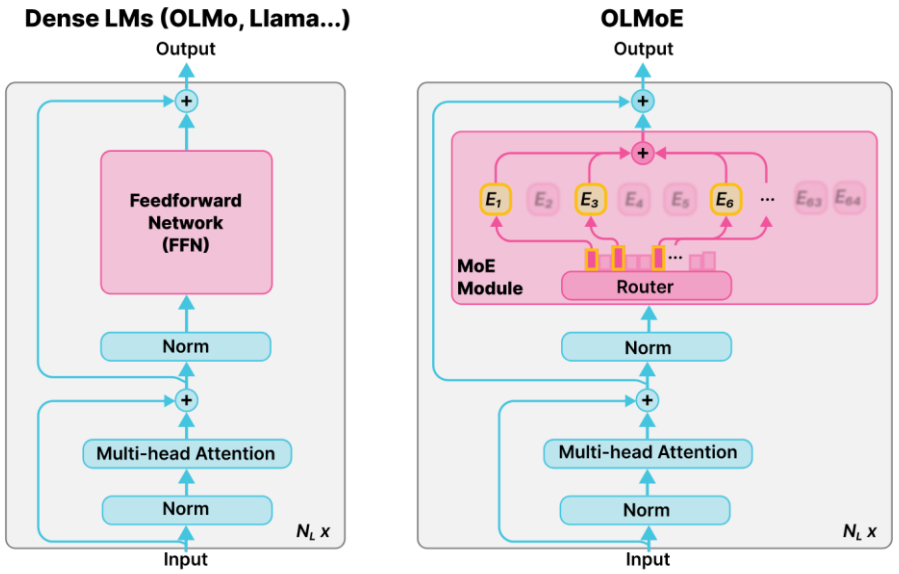
$$\sum_{i=0}^{n-1} G(x)_i \cdot E_i(x).$$

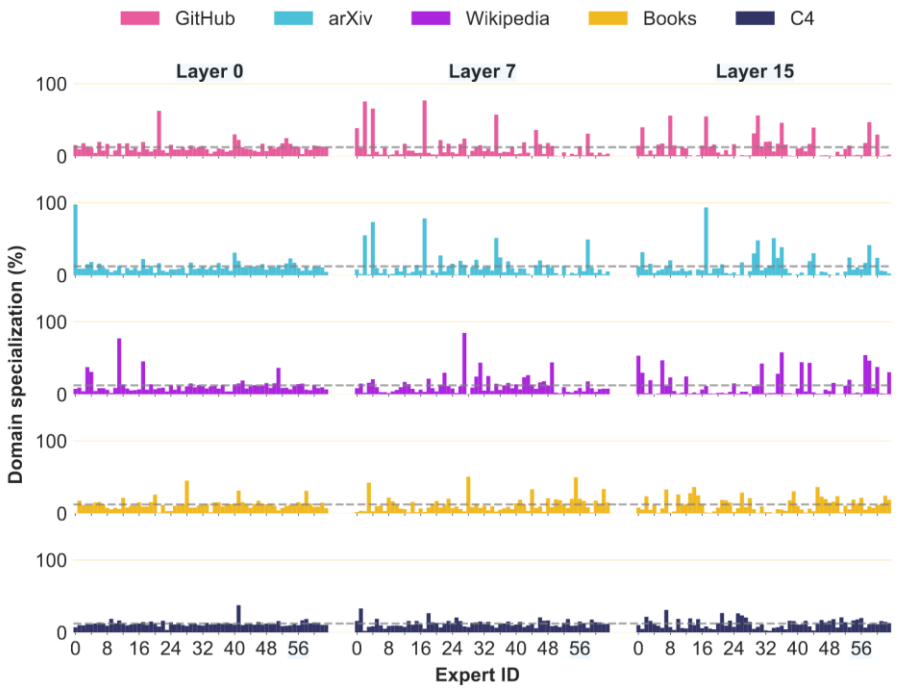
$$G(x) := \text{Softmax}(\text{TopK}(x \cdot W_g)),$$

- 本报告专注于分享第三个问题：

  Where to Mix?

  - 单模态内
  - 单模态->多模态
  - 单任务->多任务



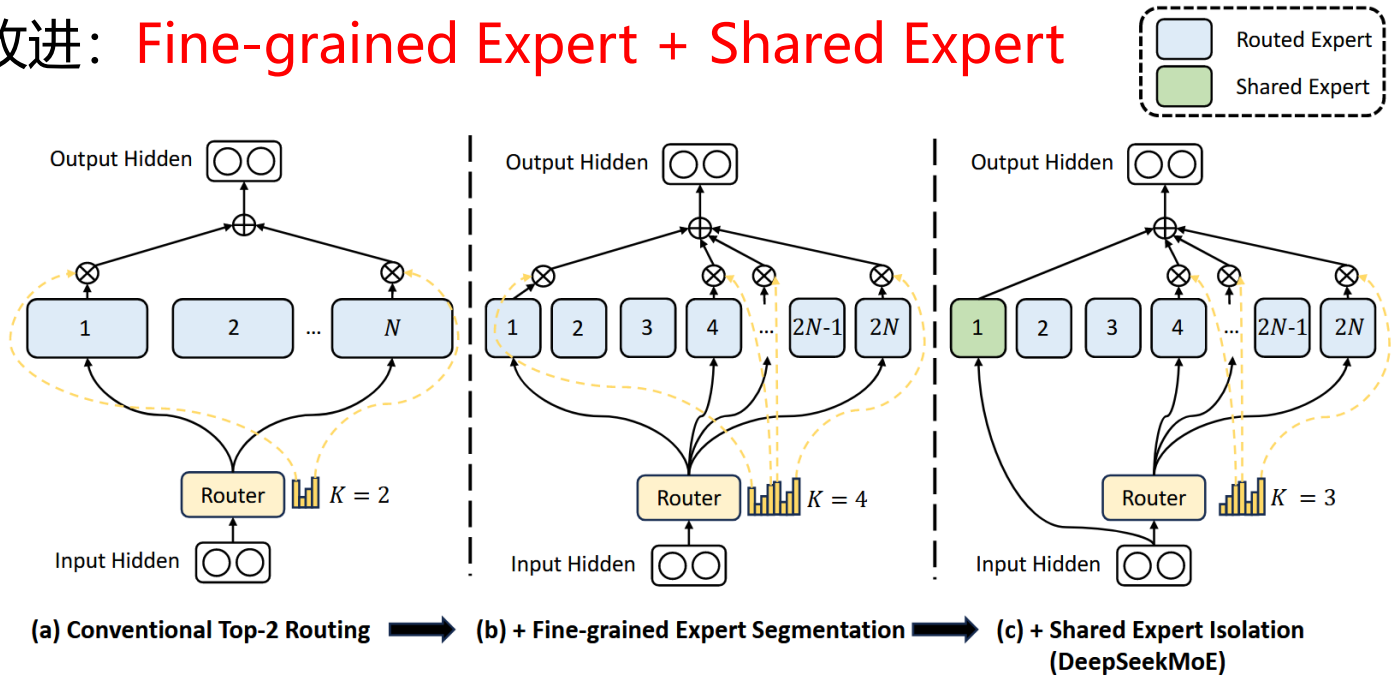



- Standard MoE
- Internal Mixture
- External Mixture
- A New Insight
- Summary

# DeepSeekMoE: Towards Ultimate Expert Specialization in Mixture-of-Experts Language Models



- 两种改进：Fine-grained Expert + Shared Expert




$$\mathbf{h}_t^l = \sum_{i=1}^{K_s} \mathrm{FFN}_i\left(\mathbf{u}_t^l\right) + \sum_{i=K_s+1}^{mN} \left(g_{i,t}\,\mathrm{FFN}_i\left(\mathbf{u}_t^l\right)\right) + \mathbf{u}_t^l,$$

$$g_{i,t} = \begin{cases} s_{i,t}, & s_{i,t} \in \mathrm{Topk}(\{s_{j,t} | K_s + 1 \leqslant j \leqslant mN\}, mK - K_s), \\ 0, & \text{otherwise}, \end{cases}$$

$$s_{i,t} = \mathrm{Softmax}_i\left(\mathbf{u}_t^{l\,T} \mathbf{e}_i^l\right).$$

- 固定activate params，增加组合数

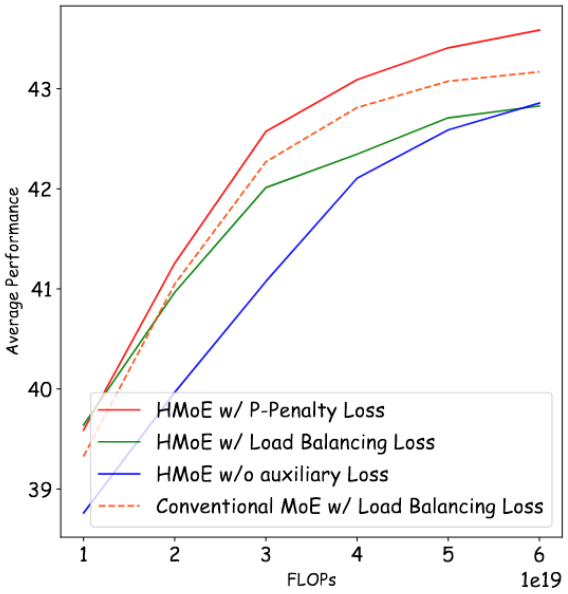
$$\binom{16}{2} = 120$$

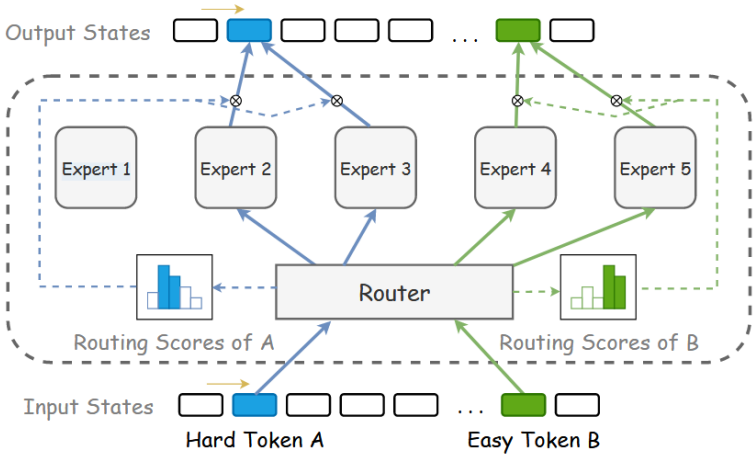
$$\binom{64}{8} = 4,426,165,368$$

# HMoE: Heterogeneous Mixture of Experts for Language Modeling

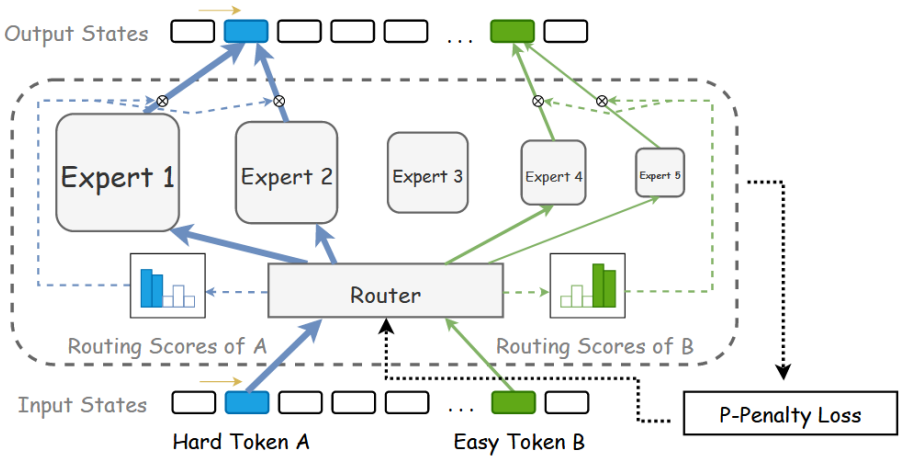


- 传统MoE模型需要控制每个expert的负载均衡，即分配给每个expert的token数量接近，因此需要每个expert的参数量相等
- 该工作设置了参数相关的惩罚性loss，使得expert的大小可以不相等




(a) Conventional homogenerous MoE.




(b) Our proposed heterogeneous MoE.

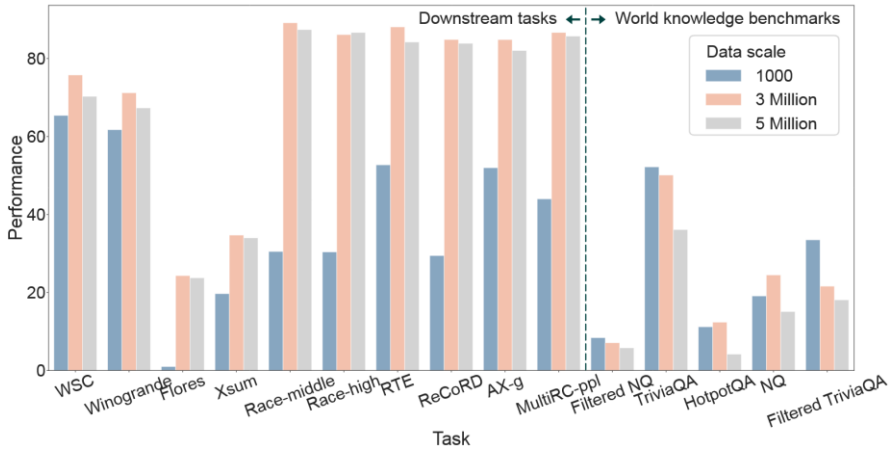
# LoRAMoE



- 在SFT阶段，用MoE方式结合多个Lora，防止通用知识的遗忘

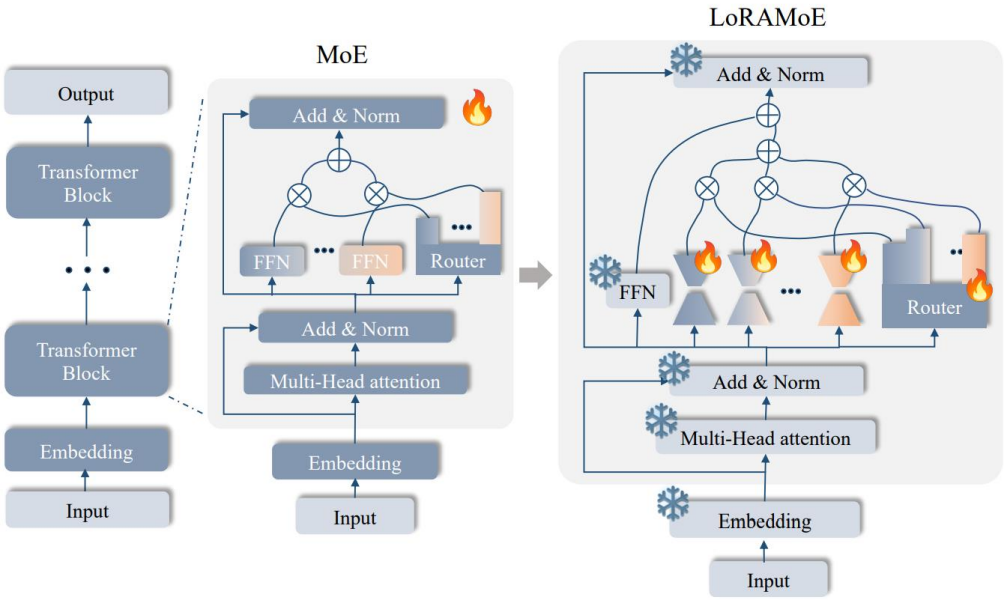
$$o = W_0 x + \gamma \sum_{i=1}^{N} G(x)_i E_i(x)$$

$$= W_0 x + \frac{\alpha}{r} \sum_{i=1}^{N} \omega_i \cdot B_i A_i x$$




**LoRAMoE: Alleviating World Knowledge Forgetting in Large Language Models via MoE-Style Plugin**

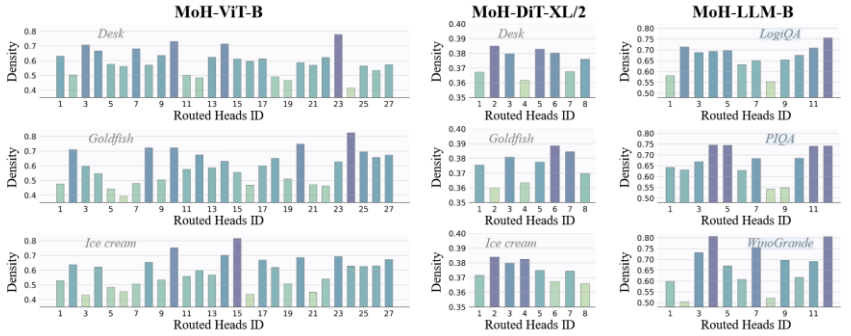
Shihan Dou[1*], Enyu Zhou[1*], Yan Liu[1], Songyang Gao[1], Wei Shen[1], Limao Xiong[1], Yuhao Zhou[1], Xiao Wang[1], Zhiheng Xi[1], Xiaoran Fan[1], Shiliang Pu[5], Jiang Zhu[5], Rui Zheng[1], Tao Gui[2†], Qi Zhang[1,3†], Xuanjing Huang[1,4†]




| Task | baseline | SFT solely on CBQA | SFT | LoRA | LoRAMoE | LoRAMoE (with $\mathcal{L}_{lbc}$) |
|---|---|---|---|---|---|---|
| **WSC** | 65.4 | - | **76.0** | 65.4 | 71.2 | 70.2 |
| **winogrande** | 61.7 | - | **71.2** | 64.3 | 66.3 | 69.6 |
| **Flores** | 0.1 | - | 24.3 | **26.6** | 26.4 | 25.9 |
| **Xsum** | 19.7 | - | 34.7 | 34.5 | **34.8** | 33.2 |
| **Race-middle** | 30.5 | - | 89.1 | 78.8 | 84.5 | **90.0** |
| **Race-high** | 30.4 | - | 86.1 | 75.3 | 80.6 | **86.5** |
| **RTE** | 52.7 | - | **88.1** | 77.3 | 80.9 | 87.4 |
| **ReCoRD** | 29.4 | - | 84.8 | 83.2 | 84.3 | **85.9** |
| **AX-g** | 52.0 | - | 84.8 | 76.1 | 81.7 | **87.1** |
| **multiRC** | 44.0 | - | 86.7 | 81.4 | 87.3 | **87.9** |
| **TriviaQA** | 52.2 | 57.8 | 51.1 | 47.8 | 55.3 | **58.1** |
| **NQ** | 18.5 | 28.6 | 24.5 | 16.2 | 23.8 | **28.0** |
| **Filtered TriviaQA** | 33.5 | 36.2 | 21.6 | 33.4 | **38.5** | 35.4 |
| **Filtered NQ** | 7.8 | 12.8 | 7.3 | 11.6 | **13.4** | 12.0 |
| **hotpot QA** | 11.2 | 16.1 | 13.4 | 10.7 | 14.4 | **16.1** |

# MoH






- MHSA本身也是一个MoE的形式，因此也可以改进成sparse结构
- 激活的比例比FFN MoE要大

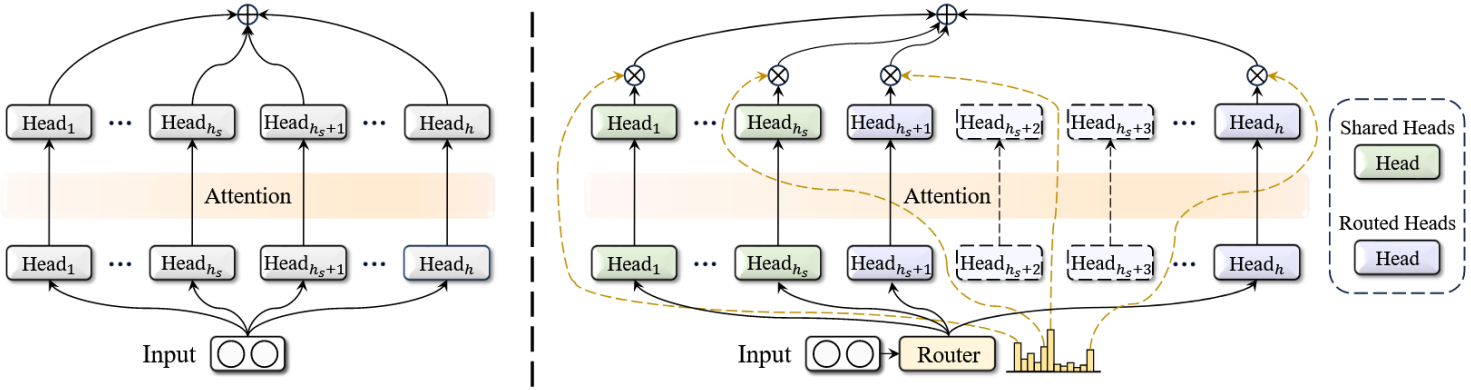
| Methods | #Activated Heads (%) | MMLU (5) | CEVAL (5) | CMMLU (5) | GSM8K(8) | TruthfulQA |
|---|---|---|---|---|---|---|
| LLaMA3-8B (Dubey et al., 2024) | 100 | 65.2 | 52.3 | 50.7 | 49.5 | 35.4 |
| **MoH-LLaMA3-8B** | 75 | 65.8 | 61.5 | 64.4 | 56.9 | 44.0 |

| Methods | #Activated Heads (%) | HellaSwag (10) | LogiQA | BoolQ (32) | LAMBADA | SciQ |
|---|---|---|---|---|---|---|
| LLaMA3-8B (Dubey et al., 2024) | 100 | 81.9 | 30.0 | 83.9 | 75.5 | 94.0 |
| **MoH-LLaMA3-8B** | 75 | 80.1 | 30.3 | 84.0 | 76.4 | 92.2 |

| Methods | #Activated Heads (%) | PIQA | WinoGrande | NQ (32) | ARC-C (25) | Average |
|---|---|---|---|---|---|---|
| LLaMA3-8B (Dubey et al., 2024) | 100 | 81.0 | 72.5 | 31.5 | 59.0 | 61.6 |
| **MoH-LLaMA3-8B** | 75 | 78.8 | 72.9 | 28.3 | 60.1 | **64.0** |

$$\text{MultiHead}(\boldsymbol{X}, \boldsymbol{X}') = \text{Concat}(\boldsymbol{H}^1, \boldsymbol{H}^2, ..., \boldsymbol{H}^h)\boldsymbol{W}_O,$$
$$\boldsymbol{H}^i = \text{Attention}(\boldsymbol{X}\boldsymbol{W}_Q^i, \boldsymbol{X}'\boldsymbol{W}_K^i, \boldsymbol{X}'\boldsymbol{W}_V^i),$$

$$\text{MultiHead}(\boldsymbol{X}, \boldsymbol{X}') = \sum_{i=1}^{h} \boldsymbol{H}^i \boldsymbol{W}_O^i. \qquad \text{MoH}(\boldsymbol{X}, \boldsymbol{X}') = \sum_{i=1}^{h} g_i \boldsymbol{H}^i \boldsymbol{W}_O^i,$$



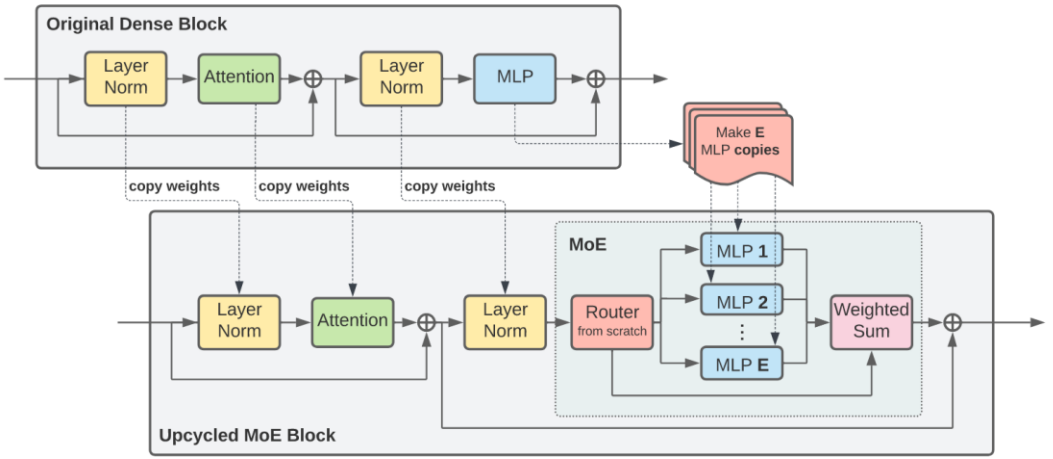

(a) Multi-Head Attention

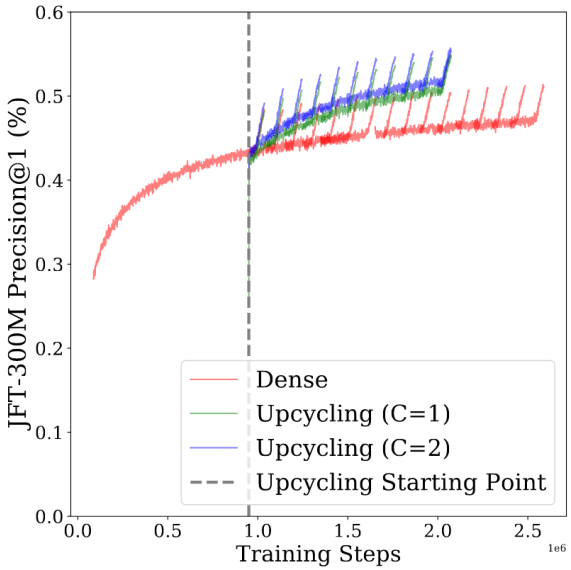
(b) Our proposed Mixture-of-Head Attention

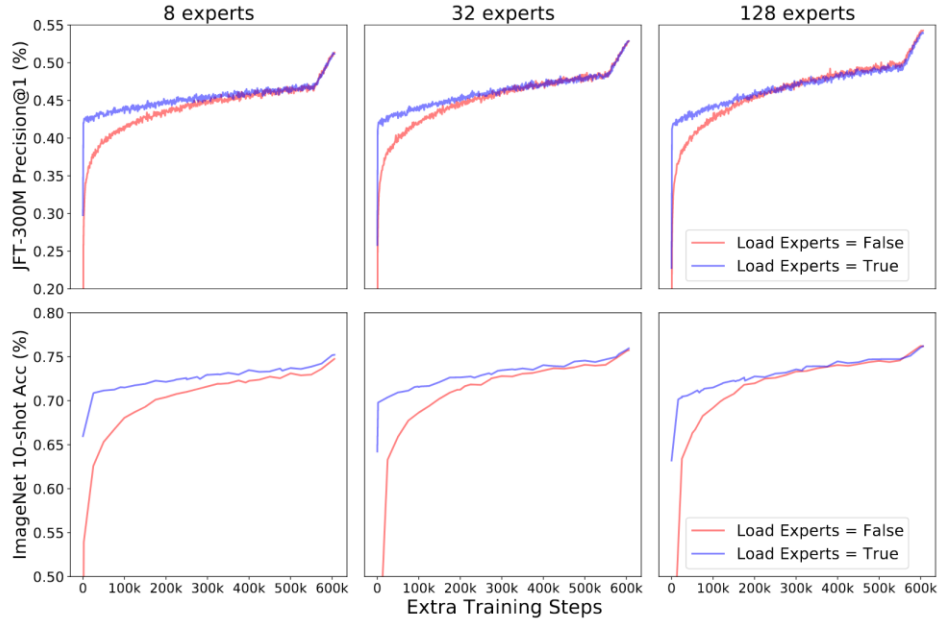
# Sparse Upcycling



- 在初期训练之后将dense模型改造成MoE结构再进行训练
- 直接将MLP参数复制多份，并引入router进行组合
- 效果比直接MoE训练要好

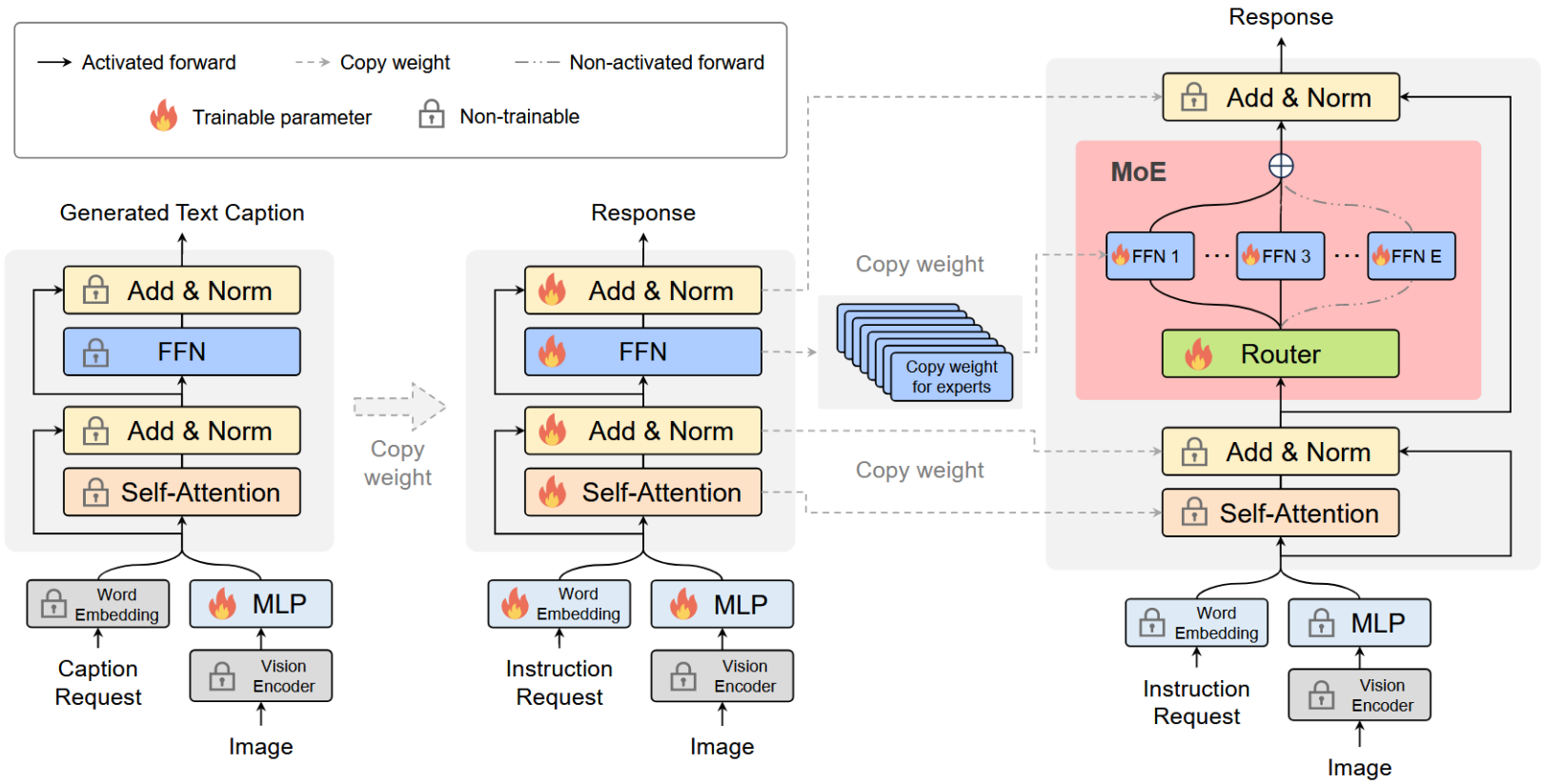


# MoE-LLaVA



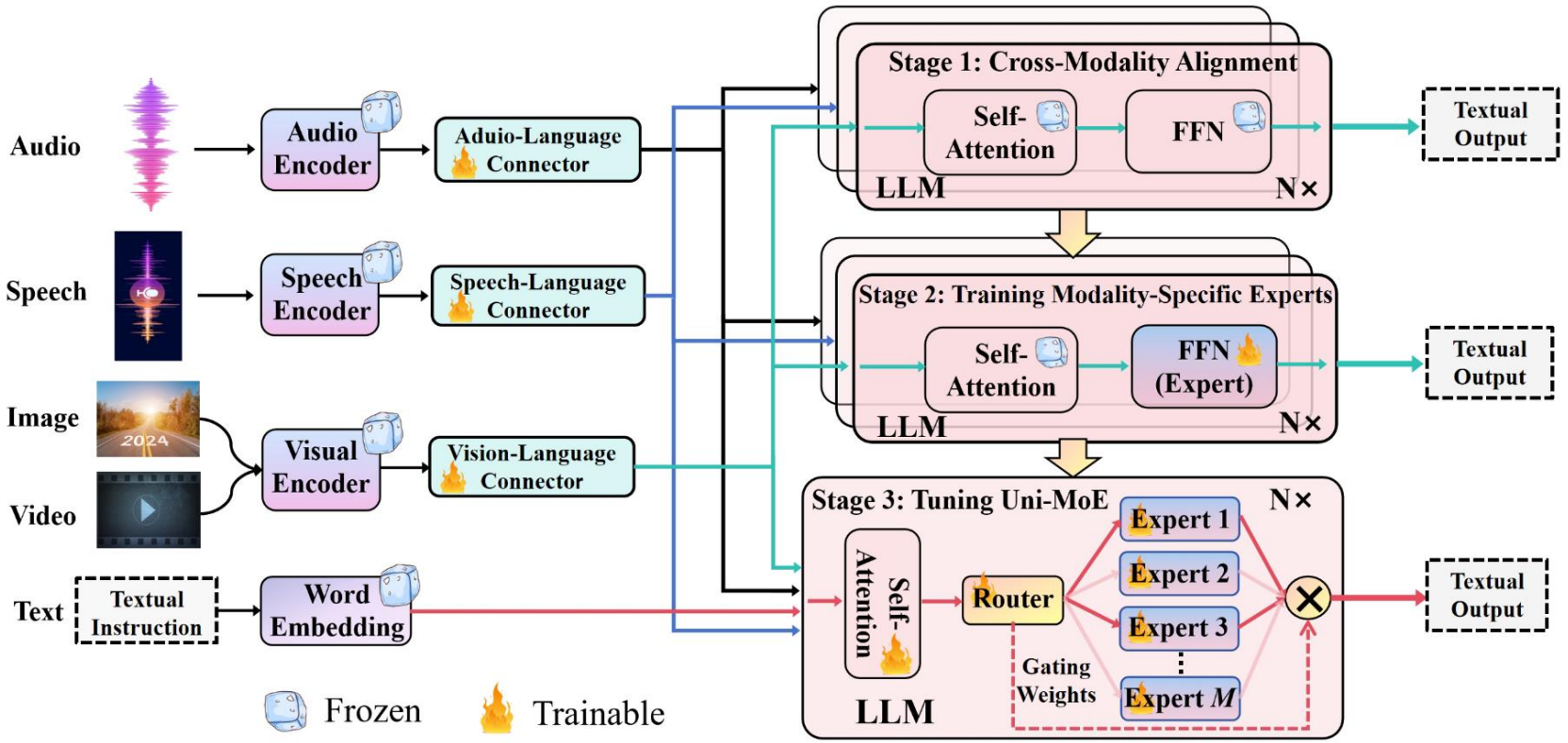



(a) Stage I (b) Stage II (c) Stage III

- Standard MoE
- Internal Mixture
- External Mixture
- A New Insight
- Summary

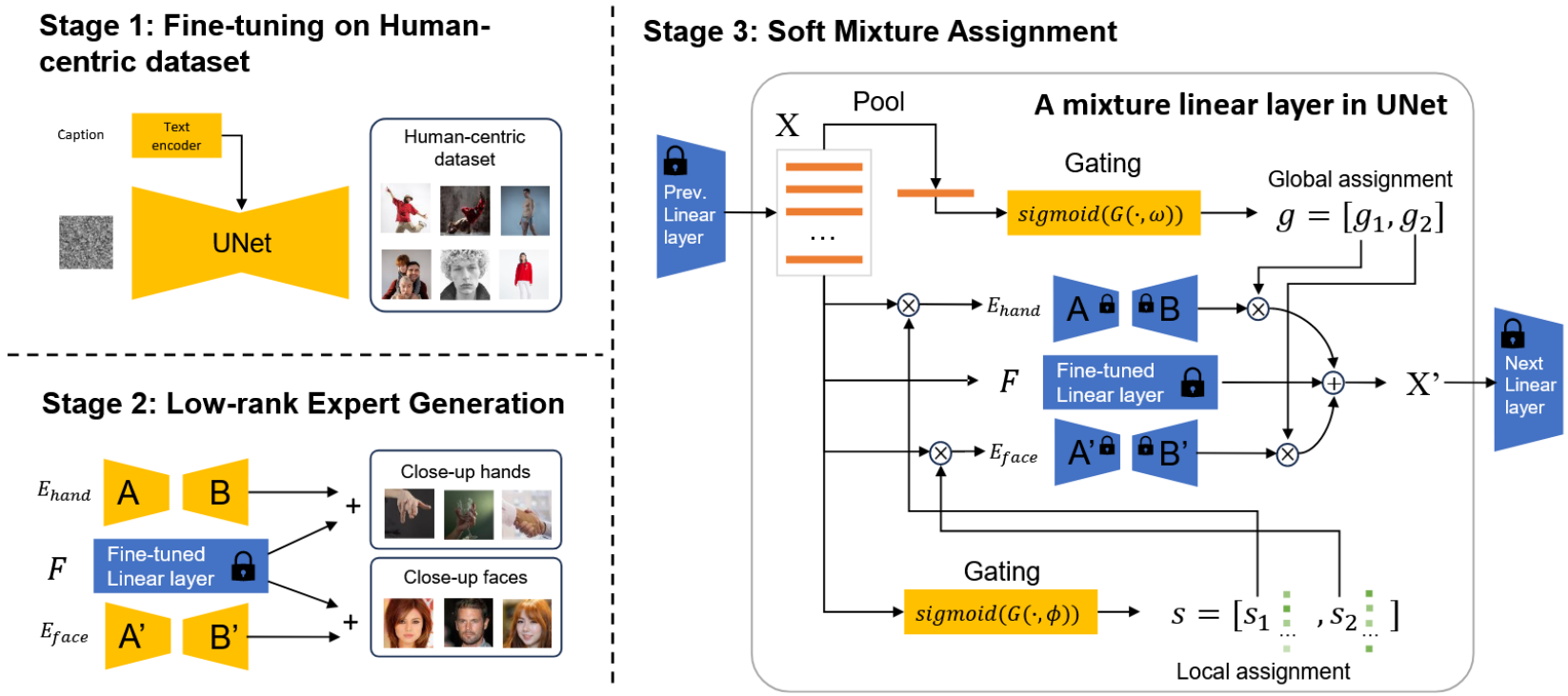
# Uni-MoE

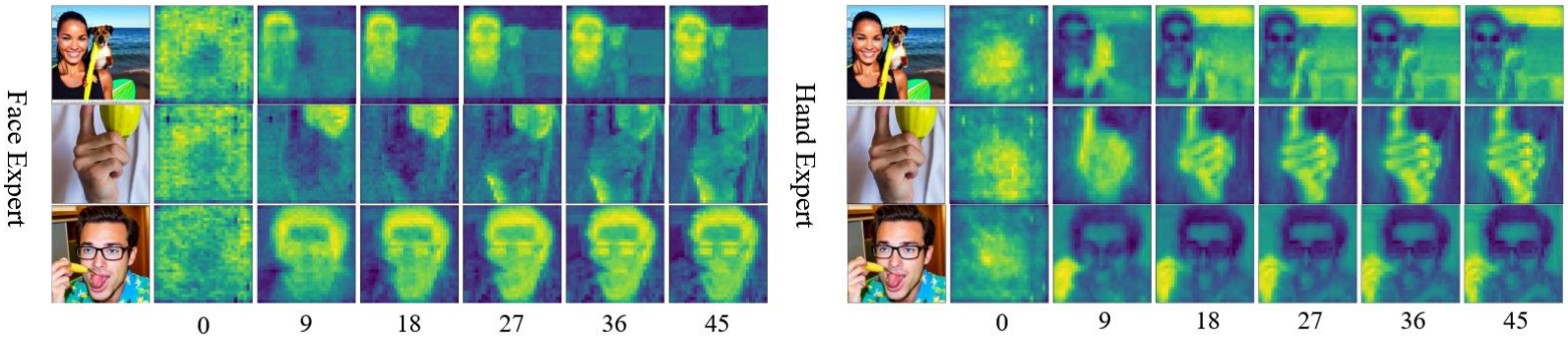



Audio
Audio Encoder
Aduio-Language Connector
Speech
Speech Encoder
Speech-Language Connector
Image
Video
Visual Encoder
Vision-Language Connector
Text
Textual Instruction
Word Embedding
Frozen
Trainable
Stage 1: Cross-Modality Alignment
Self-Attention
FFN
LLM
N×
Textual Output
Stage 2: Training Modality-Specific Experts
Self-Attention
FFN (Expert)
LLM
N×
Textual Output
Stage 3: Tuning Uni-MoE
N×
Self-Attention
Router
Expert 1
Expert 2
Expert 3
Expert M
Gating Weights
LLM
Textual Output

# MoLE



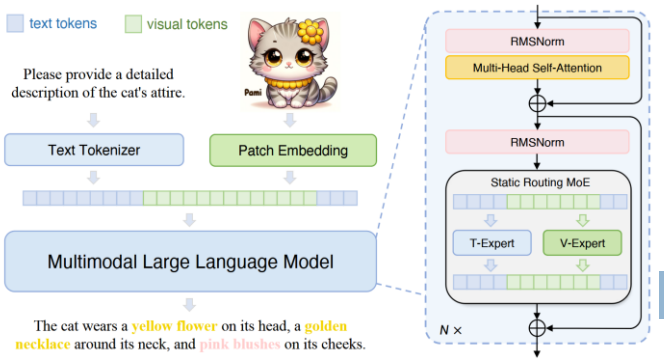


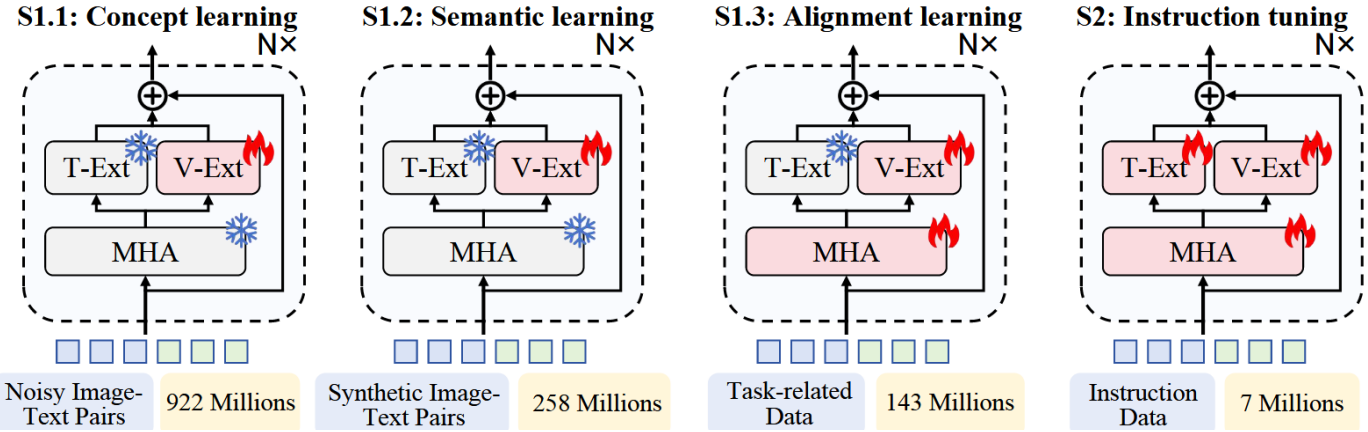




(d) Visualization of score map from face expert (left) and hand expert (right) in different inference step.

# Mono-InternVL





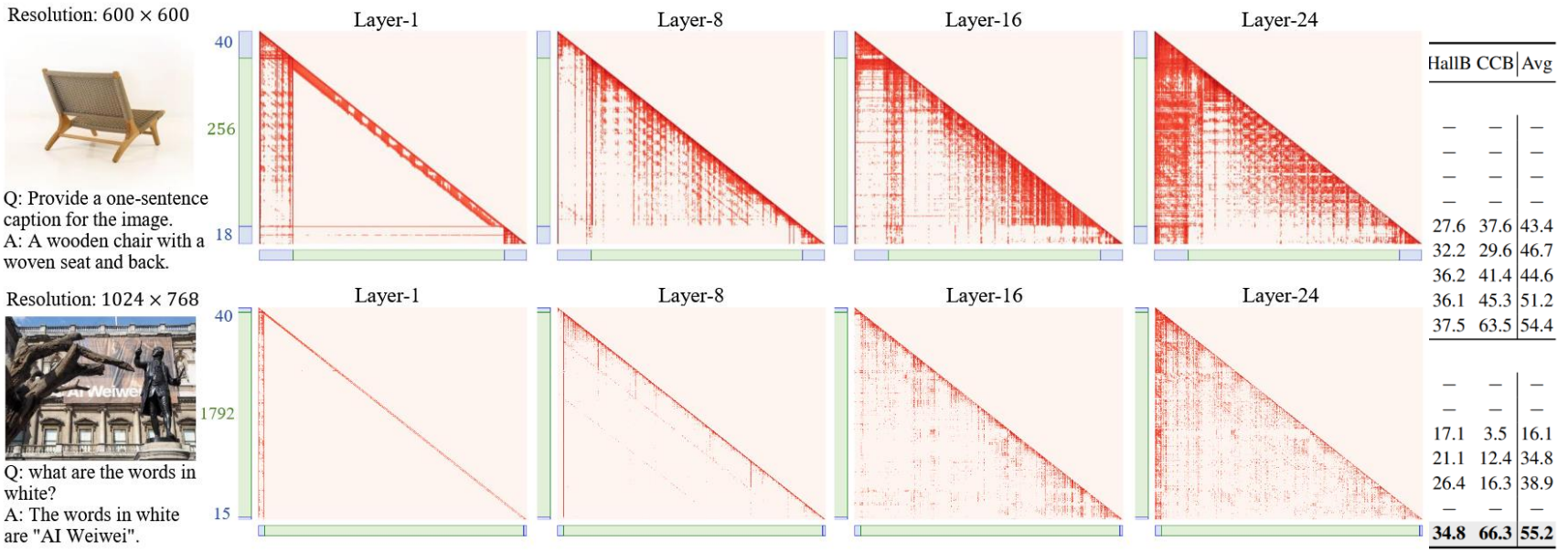




- 不
- 构

$x_m^{l'}$

$x_m^{l}$

MM




| HallB | CCB | Avg |
|---|---|---|
| – | – | – |
| – | – | – |
| – | – | – |
| 27.6 | 37.6 | 43.4 |
| 32.2 | 29.6 | 46.7 |
| 36.2 | 41.4 | 44.6 |
| 36.1 | 45.3 | 51.2 |
| 37.5 | 63.5 | 54.4 |

| | | |
|---|---|---|
| – | – | – |
| – | – | – |
| 17.1 | 3.5 | 16.1 |
| 21.1 | 12.4 | 34.8 |
| 26.4 | 16.3 | 38.9 |
| – | – | – |
| **34.8** | **66.3** | **55.2** |

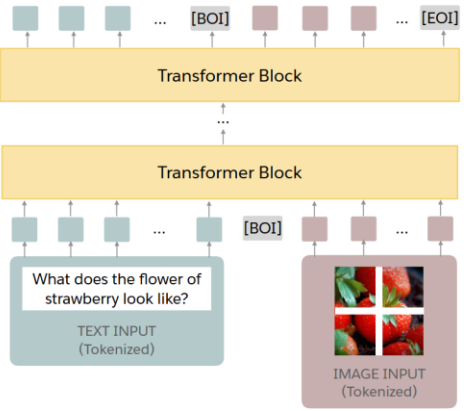
# MoMa: Efficient Early-Fusion Pre-training with Mixture of Modality-Aware Experts

Xi Victoria Lin[1,*], Akshat Shrivastava[1,*], Liang Luo[1], Srinivasan Iyer[1], Mike Lewis[1], Gargi Ghosh[1], Luke Zettlemoyer[1], Armen Aghajanyan[1,*]

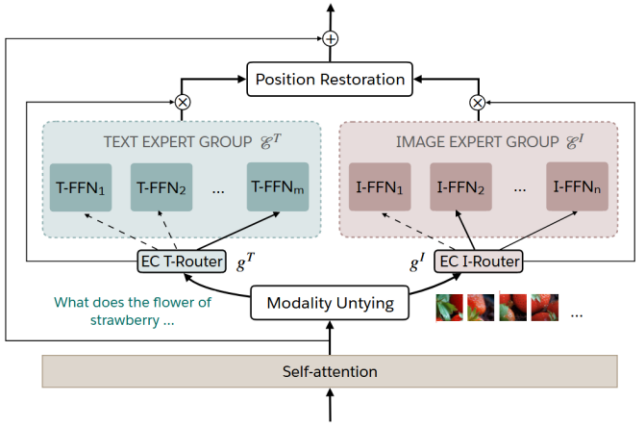
[1]Meta FAIR



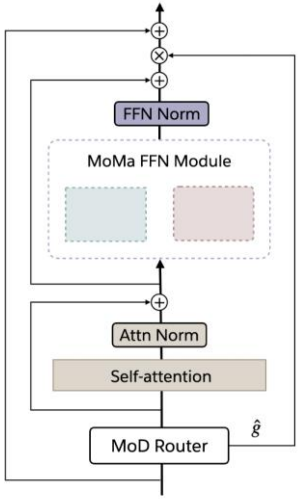
- 扩展FFN到多模态，每个模态内还有MoE，形成多级知识库
- 对整个模型也进行mixture of depth，即选择是否跳过一些编码器层

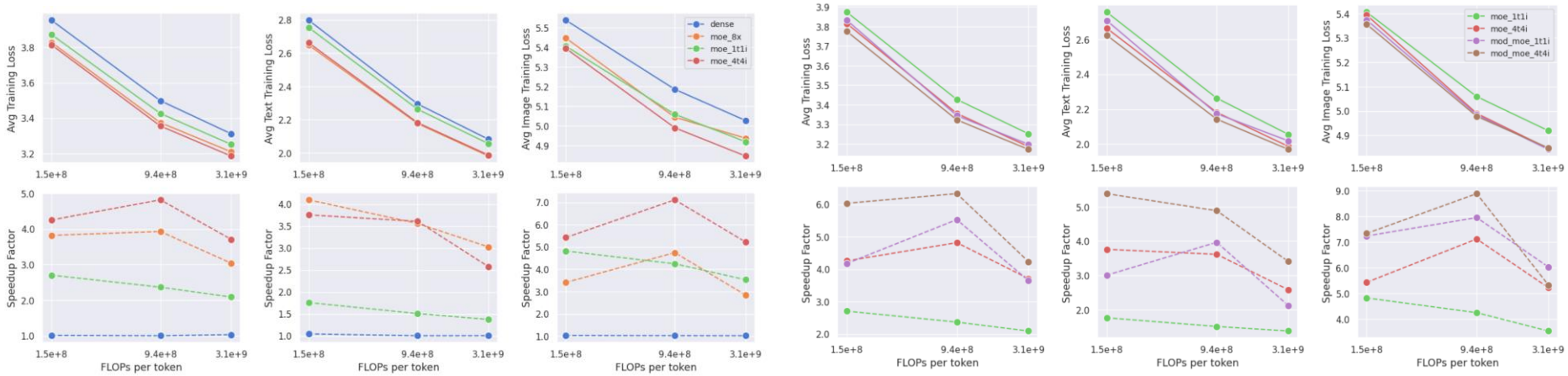



(a) Early-fusion mixed-modal LLM architecture.




(b) Mixture of modality-aware experts (MoMa) transformer block.



Step1: High-resolution Adaptation and Training Recipe

Step2: Fusion Paradigm Exploration

| Vision Encoder(s) | Fusion | Token (V) | #Tokens/s | GQA | MME | MMMU | OCR | SQA | POPE | TextVQA | Avg |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *CLIP* | – | 1024 | **47.9** | 64.6 | **1531** | 35.7 | 496 | **72.8** | 87.9 | 64.0 | 644.4 |
| *ConvNeXt* | – | 1024 | 46.8 | 63.7 | 1433 | **37.0** | 527 | 72.0 | 87.9 | 71.7 | 652.3 |
| *CLIP + ConvNeXt* | *SA* | 2048 | 46.1 | 64.6 | 1482 | 35.3 | 536 | 72.5 | 88.3 | 71.8 | 657.2 |
| | *CC* | 1024 | 47.3 | 64.0 | 1486 | 36.0 | 533 | 72.7 | 88.6 | 71.9 | **658.2** |
| | *LH* | 1024 | 47.0 | 62.9 | 1488 | 36.8 | 521 | 72.6 | **88.9** | 68.7 | 652.1 |
| | *MG* | 1024 | 45.2 | **64.9** | 1481 | 36.7 | 474 | **72.8** | 88.2 | 68.6 | 646.7 |
| | *DA* | 1024 | 47.3 | 63.6 | 1497 | 35.3 | 459 | 71.9 | 88.0 | 69.1 | 640.9 |
| *CLIP + ConvNeXt + SAM* | *SA* | 3072 | 40.3 | **64.9** | 1439 | 35.0 | 504 | 72.7 | 88.2 | 71.3 | 649.2 |
| | *CC* | 1024 | 46.3 | 64.6 | 1497 | 35.2 | **558** | 71.7 | 87.9 | **72.2** | **660.4** |

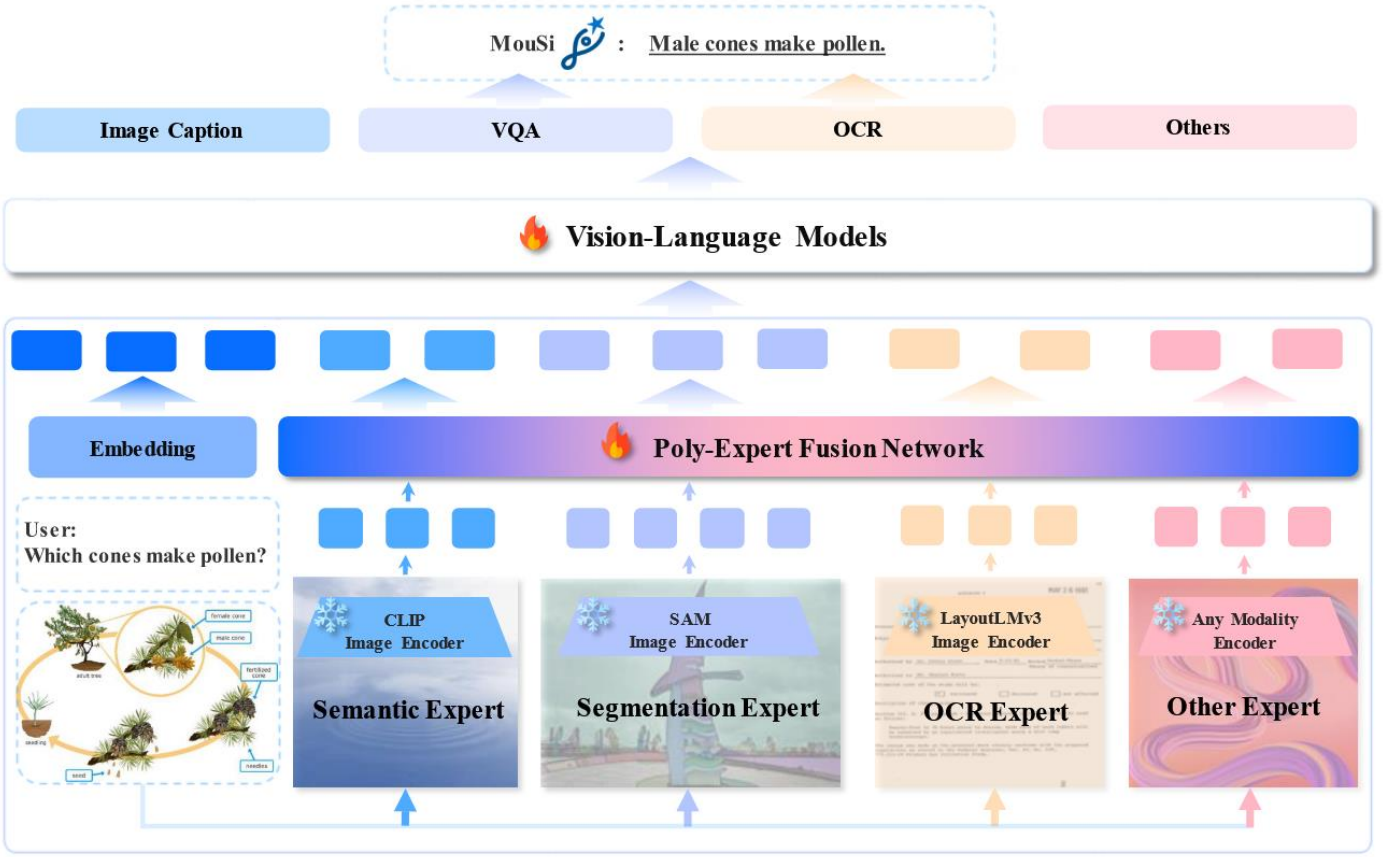
High Resolution

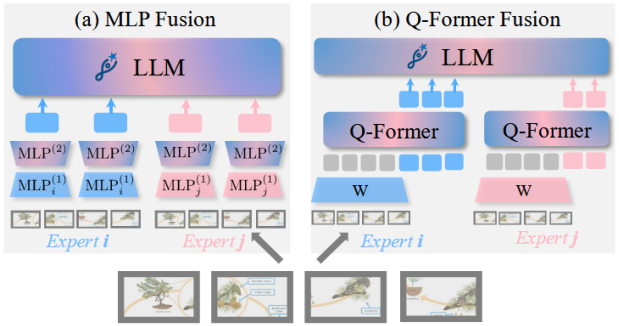
| Model | Knowledge | | | | | General | | | | | OCR and Chart | | | | | Vision-Centric | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Avg | SQA$^I$ | MMMU | MathVista | AI2D | Avg | MME | MMBench | SEED | GQA | Avg | ChartQA | OCRBench | TextVQA | DocVQA | Avg | MMVP | RWQA |
| *Llama3-8B* | | | | | | | | | | | | | | | | | | |
| *MGM-HD* | 55.7 | 75.1 | 37.3 | 37.0 | 73.5 | 72.7 | **1606** | 72.7 | 73.2 | 64.5 | 62.9 | 59.1 | 47.7 | 70.2 | 74.6 | 40.4 | 18.7 | 62.1 |
| *Cambrian-1* | 61.3 | 80.4 | 42.7 | 49.0 | 73.0 | 73.1 | 1547 | 75.9 | 74.7 | 64.6 | 71.3 | 73.3 | 62.4 | 71.7 | 77.8 | 57.6 | 51.3 | 64.2 |
| *Eagle* | **64.2** | **84.3** | **43.8** | **52.7** | **76.1** | **73.8** | 1559 | **75.9** | **76.3** | **64.9** | **76.6** | **80.1** | **62.6** | **77.1** | **86.6** | **69.1** | **71.6** | **66.5** |
| *Vicuna-13B* | | | | | | | | | | | | | | | | | | |
| *MGM-HD* | 54.1 | 71.9 | 37.3 | 37 | 70.1 | 70.7 | 1597 | 68.6 | 70.6 | 63.7 | 60.8 | 56.6 | 46.6 | 70.2 | 69.8 | 38.4 | 19.3 | 57.5 |
| *Cambrian-1* | 60.2 | 79.3 | 40.0 | 48.0 | 73.6 | 73.7 | 1610 | **75.7** | 74.4 | 64.3 | 71.3 | 73.8 | **61.9** | 72.8 | 76.8 | 52.2 | 41.3 | 63.0 |
| *Eagle* | **63.0** | **82.0** | **41.6** | **54.4** | **74.0** | **74.6** | **1651** | **75.7** | **74.8** | **65.3** | **75.1** | **77.6** | **61.9** | **75.5** | **85.4** | **61.4** | **58.0** | **64.8** |
| *Yi-34B* | | | | | | | | | | | | | | | | | | |
| *MGM-HD* | 62.4 | 77.7 | 48 | 43.4 | 80.5 | 76.2 | 1659 | 80.6 | 75.3 | 65.8 | 68.1 | 67.6 | 51.8 | 74.1 | 78.9 | 52.3 | 37.3 | 67.2 |
| *Cambrian-1* | 67.0 | **85.6** | 49.7 | 53.2 | **79.7** | **76.8** | **1689** | **81.4** | 75.3 | **65.8** | 71.9 | 75.6 | 60.0 | 76.7 | 75.5 | 60.3 | 52.7 | 67.8 |
| *Eagle* | **68.6** | 85.5 | **53.2** | **57.9** | 79.1 | 76.3 | 1677 | 81.0 | **75.6** | 64.9 | **75.4** | **77.2** | **62.4** | **78.8** | **83.0** | **68.3** | **67.0** | **69.5** |

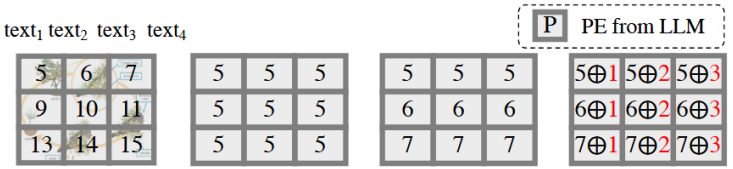
Large Language Model

# Poly-Visual-Expert Vision-Language Models





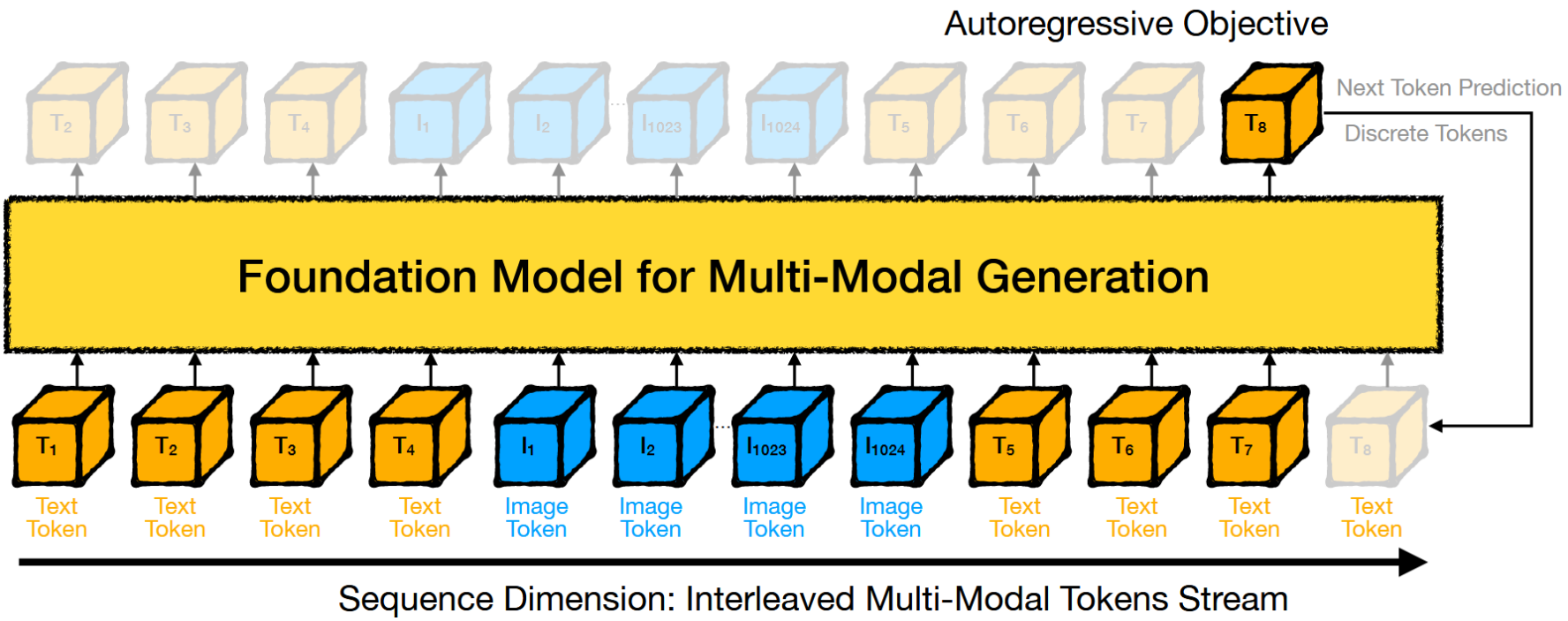



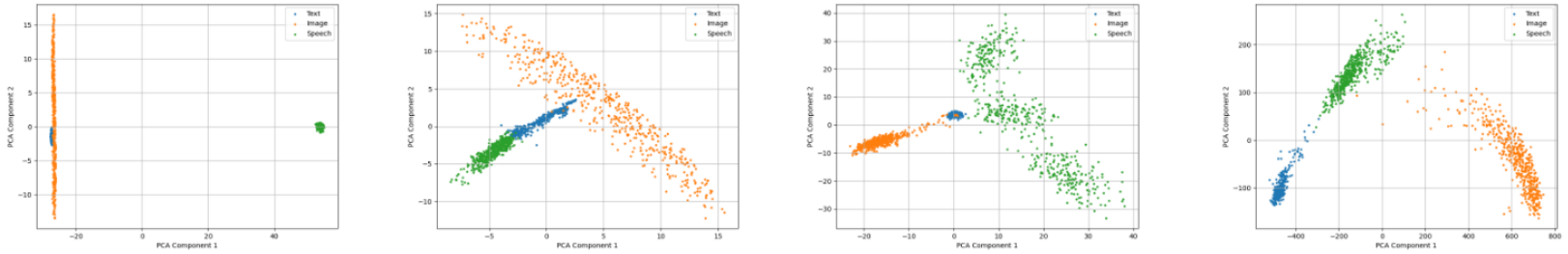



# Mixture-of-Transformers: A Sparse and Scalable Architecture for Multi-Modal Foundation Models

Weixin Liang[2,*], Lili Yu[1†], Liang Luo[1†], Srinivasan Iyer[1], Ning Dong[1], Chunting Zhou[1], Gargi Ghosh[1], Mike Lewis[1], Wen-tau Yih[1], Luke Zettlemoyer[1], Xi Victoria Lin[1]





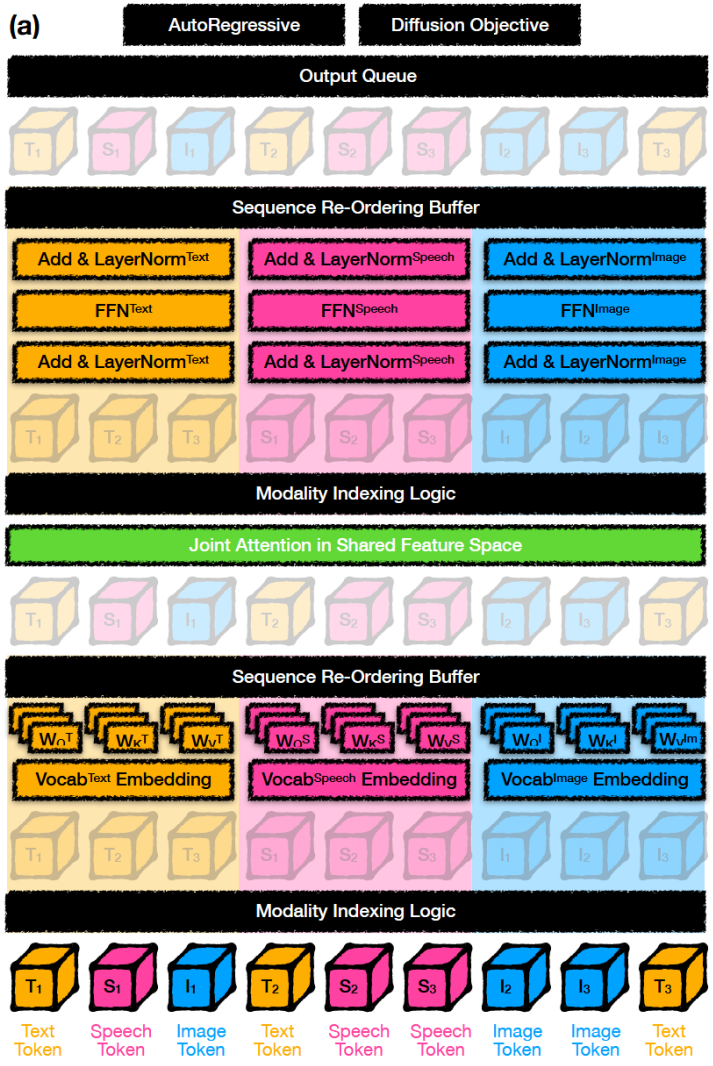
(b) Layer 1 (c) Layer 5 (d) Layer 17 (e) Layer 32

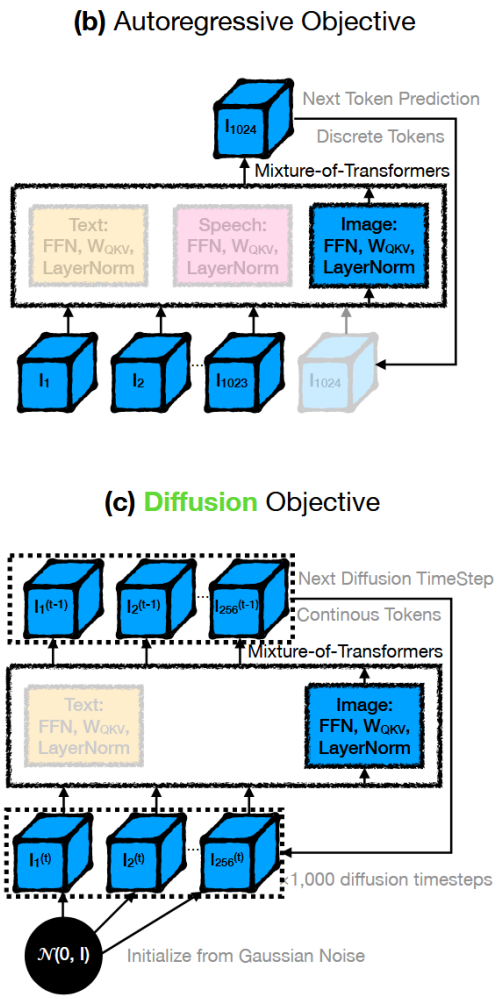
# Mixture-of-Transformers: A Sparse and Scalable Architecture for Multi-Modal Foundation Models

Weixin Liang[2,*], Lili Yu[1†], Liang Luo[1†], Srinivasan Iyer[1], Ning Dong[1], Chunting Zhou[1], Gargi Ghosh[1], Mike Lewis[1], Wen-tau Yih[1], Luke Zettlemoyer[1], Xi Victoria Lin[1]



- 各模态在attn中共同计算注意力，但采用各自的QKV进行映射
- 各模态采用各自的FFN和Norm，在FFN部分单独进行
- 多种方式进行生成训练，自回归或者扩散均可



- Standard MoE
- Internal Mixture
- External Mixture
- A New Insight
- Summary

- Standard MoE
- Internal Mixture
- External Mixture
- A New Insight
- Summary

# 总结反思



- 可以从各个角度构建MoE，只要符合“比较-激活-加权”的形式，均可以视为MoE
- MoE的用途：
  - 多任务：Eagle, Poly,
  - 多模态：Uni-MoE, Mono-InternVL, MoMa, MoT
  - 稀疏高效：Mixtral, DeepseekMoE, MoH, Sparse Upcycling, LoRAMoE, OLMoE, MoE-LLaVA
  - 微调：LoRAMoE, MoLE
  - ……

# 谢谢！