MoME: Mixture of Matryoshka Experts for Audio-Visual Speech Recognition

NEURAL INFORMATION PROCESSING SYSTEMS

U. **Cappellazzo**[1], M. Kim[2], P. Ma[2], H. Chen[2], X. Liu[2], S. Petridis[1], M. Pantic[1]

1 IMPERIAL

2 Meta

LLM-based Audio-Visual Speech Recognition

- **Audio-Visual Speech Recognition** (AVSR) seeks to enhance the robustness and accuracy of speech recognition systems under noisy conditions by incorporating an additional *visual* modality (i.e., lip movements).

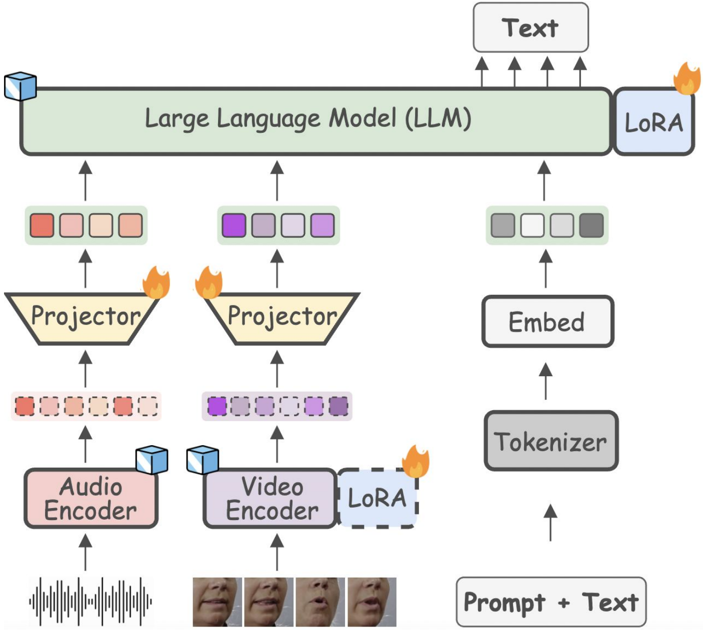
LLM-based Audio-Visual Speech Recognition

- **Audio-Visual Speech Recognition** (AVSR) seeks to enhance the robustness and accuracy of speech recognition systems under noisy conditions by incorporating an additional *visual* modality (i.e., lip movements).

- Recently, multiple works have demonstrated that **Large Language Models** (LLMs) can be harnessed to carry out the task of AVSR, ASR, and VSR, achieving state-of-the-art results on multiple benchmarks.

LLM-based Audio-Visual Speech Recognition

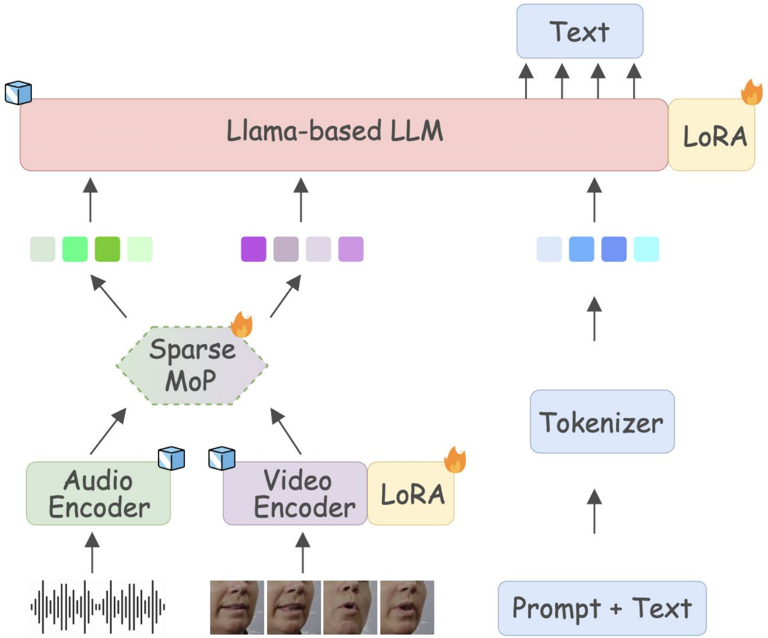
- Examples of LLM-based AVSR systems are: **Llama-AVSR [ICASSP 2025]**, Llama-SMoP [Interspeech 2025], and MMS-Llama [ACL Findings 2025].



[Cappellazzo et al., 2025(a)]

LLM-based Audio-Visual Speech Recognition

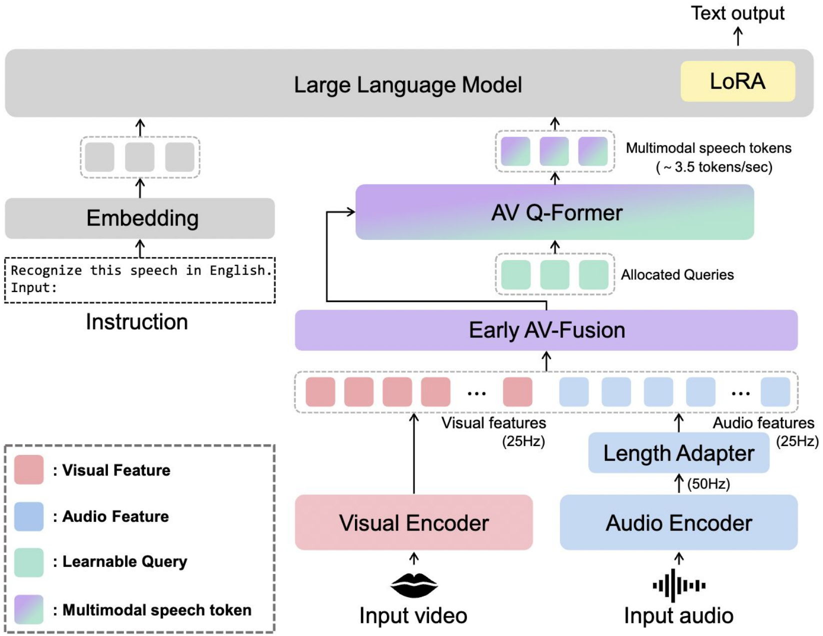
- Examples of LLM-based AVSR systems are: Llama-AVSR [ICASSP 2025], **Llama-SMoP [Interspeech 2025]**, and MMS-Llama [ACL Findings 2025].



[Cappellazzo et al., 2025(b)]

LLM-based Audio-Visual Speech Recognition

- Examples of LLM-based AVSR systems are: Llama-AVSR [ICASSP 2025], Llama-SMoP [Interspeech 2025], and **MMS-Llama [ACL Findings 2025]**.



[Yeo et al., 2025]

Accuray/Efficiency Tradeoff

- Multimodal LLMs are **token-hungry**: they tend to perform better when provided with *fine-grained*, *dense* token representations.

Accuray/Efficiency Tradeoff

- Multimodal LLMs are **token-hungry**: they tend to perform better when provided with *fine-grained*, *dense* token representations.
- Therefore, it is common practice to **reduce** the number of tokens before feeding them to the LLM (e.g., Q-Former/avg pooling).

Accuray/Efficiency Tradeoff

- Multimodal LLMs are **token-hungry**: they tend to perform better when provided with *fine-grained*, *dense* token representations.
- Therefore, it is common practice to **reduce** the number of tokens before feeding them to the LLM (e.g., Q-Former/avg pooling).
- However, since **they require fixing a compression rate in advance, they produce a single fixed-length output**, offering no flexibility to balance information density and efficiency at inference time.

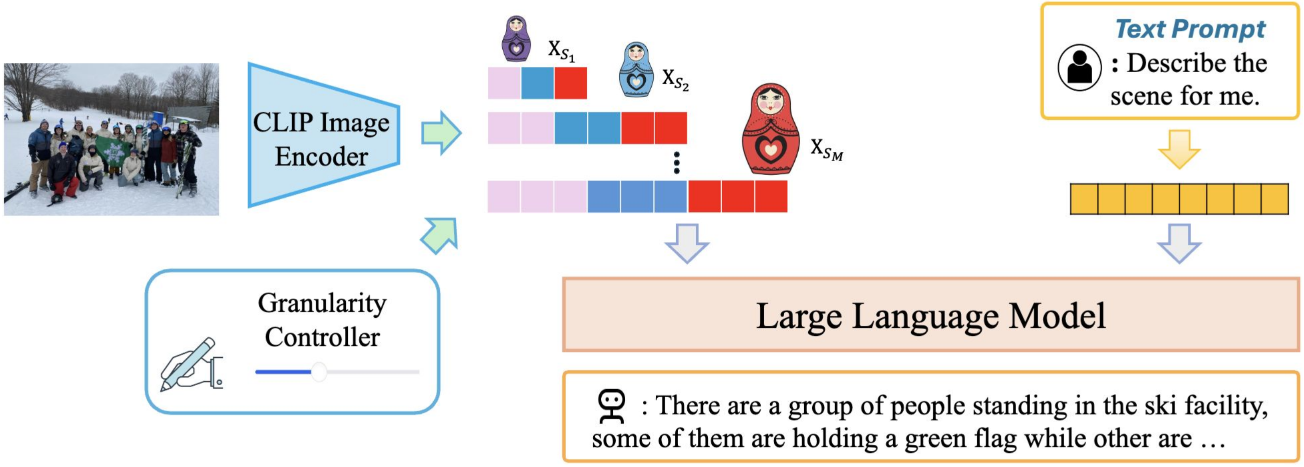
Matryoshka-based MLLMs

- To obviate this issue, Matryoshka-based MLLMs exploit the matryoshka representation learning (**MRL**) principle to train models across multiple token granularities, allowing the number of multimodal tokens to be dynamically adjusted at inference time.

[Kusupati et al., 2022(a)]

Matryoshka-based MLLMs

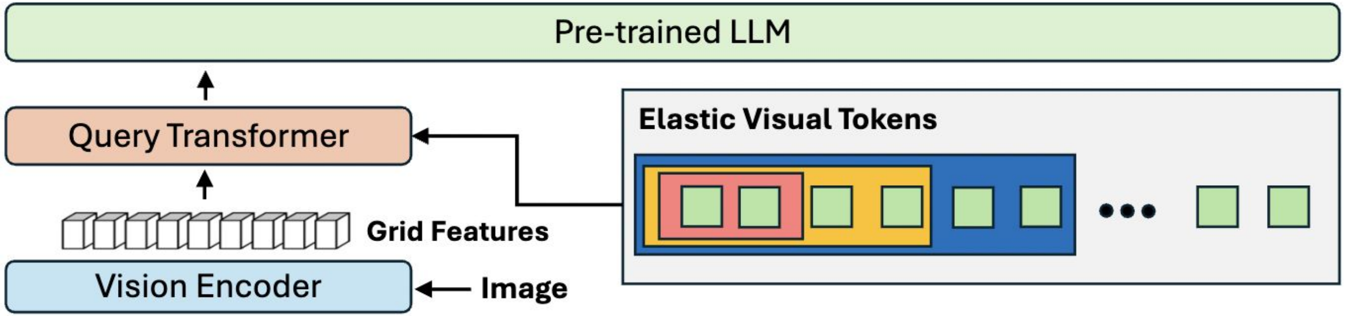
- To obviate this issue, Matryoshka-based MLLMs exploit the matryoshka representation learning (**MRL**) principle to train models across multiple token granularities, allowing the number of multimodal tokens to be dynamically adjusted at inference time.
- Models of this kind include **M^3**, MQT, and Llama-MTSK.



[Cai et al., 2024]

Matryoshka-based MLLMs

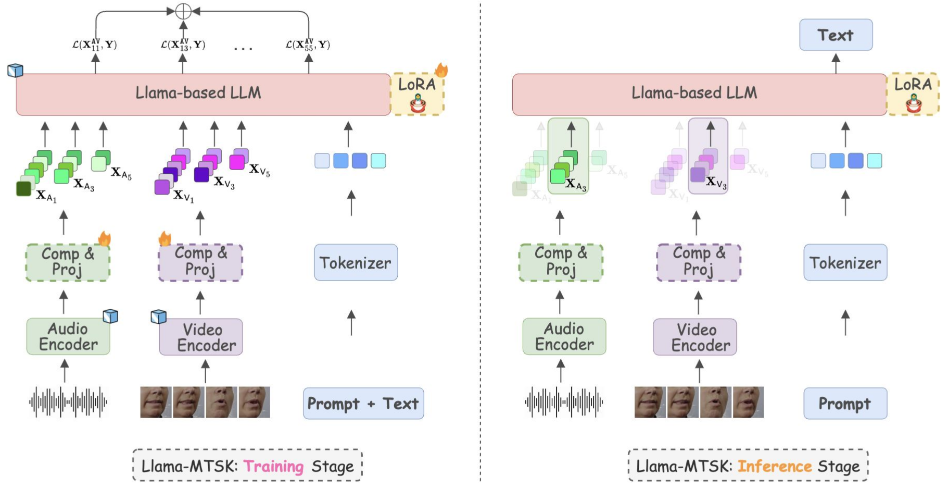
- To obviate this issue, Matryoshka-based MLLMs exploit the matryoshka representation learning (**MRL**) principle to train models across multiple token granularities, allowing the number of multimodal tokens to be dynamically adjusted at inference time.
- Models of this kind include M^3, **MQT**, and Llama-MTSK.



[Hu et al., 2024]

Matryoshka-based MLLMs

- To obviate this issue, Matryoshka-based MLLMs exploit the matryoshka representation learning (**MRL**) principle to train models across multiple token granularities, allowing the number of multimodal tokens to be dynamically adjusted at inference.
- Models of this kind include M^3, MQT, and **Llama-MTSK**.



[Cappellazzo et al., 2025(c)]

Matryoshka-based MLLMs: Current Limitations

- Current Matryoshka models rely on *uniform*, *monolithic* representations at each scale and *treat each resolution independently* during training.

Matryoshka-based MLLMs: Current Limitations

- Current Matryoshka models rely on *uniform*, *monolithic* representations at each scale and *treat each resolution independently* during training.
- This lack of inter-scale interaction forces the model to compromise between generality and specialization, often yielding suboptimal performance at higher rates.

MoME: Mixture of Matryoshka Experts

- **MoME** is a novel module that unifies **MRL** and **MoE** paradigms.

MoME: Mixture of Matryoshka Experts

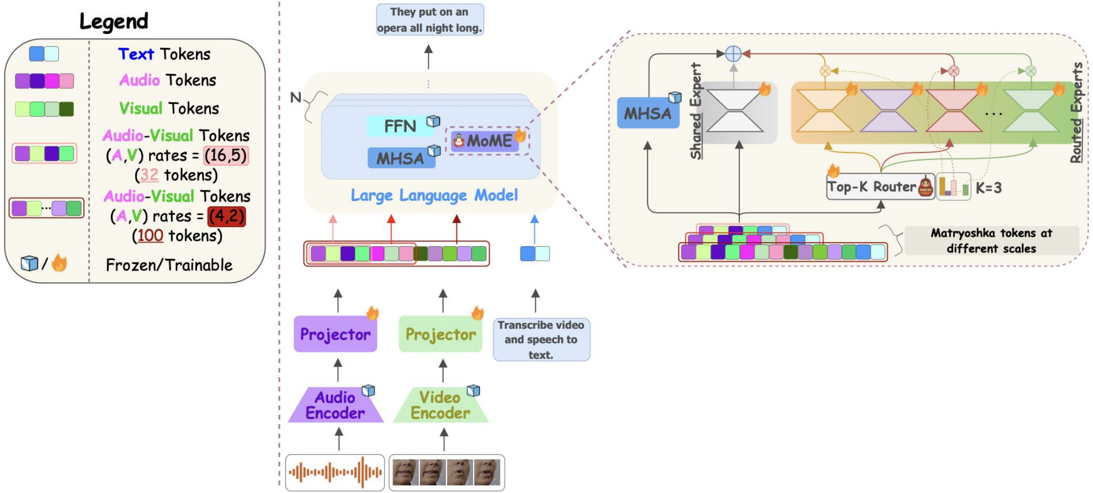
- **MoME** is a novel module that unifies **MRL** and **MoE** paradigms.
- It introduces a set of *routed* and *shared* experts trained jointly across granularities. This design allows experts to learn fine-grained (high-resolution) features that can later be reused when processing compressed (low-resolution) tokens at inference time.

MoME: Mixture of Matryoshka Experts

- **MoME** is a novel module that unifies **MRL** and **MoE** paradigms.
- It introduces a set of *routed* and *shared* experts trained jointly across granularities. This design allows experts to learn fine-grained (high-resolution) features that can later be reused when processing compressed (low-resolution) tokens at inference time.
- By aligning expert training across scales, MoME promotes cross-granularity knowledge transfer and improves the robustness of the model under aggressive compression.

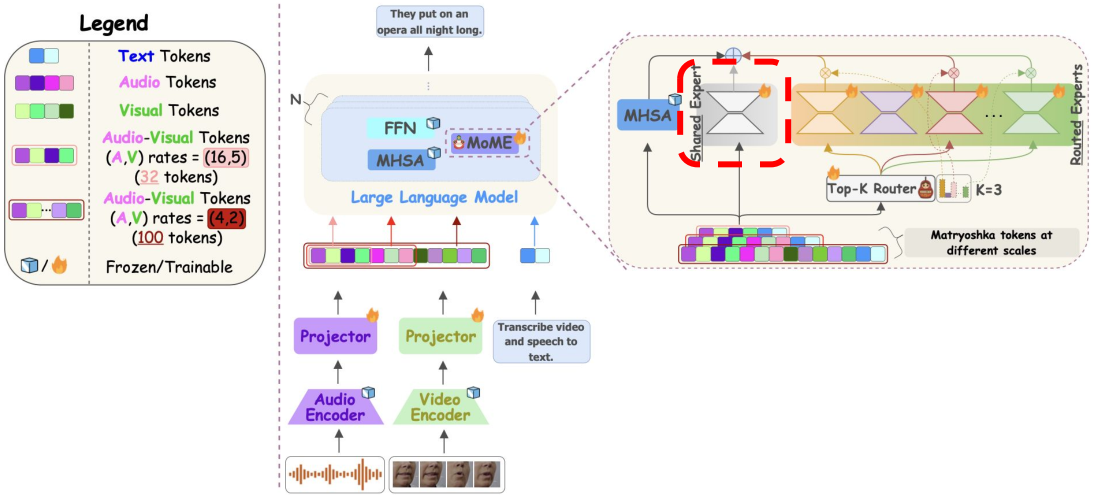
MoME: Mixture of Matryoshka Experts

- **MoME** is inserted in <u>parallel</u> with existing LLM layers (MHSA, FFN, entire layer), allowing efficient fine-tuning of a frozen pre-trained module.



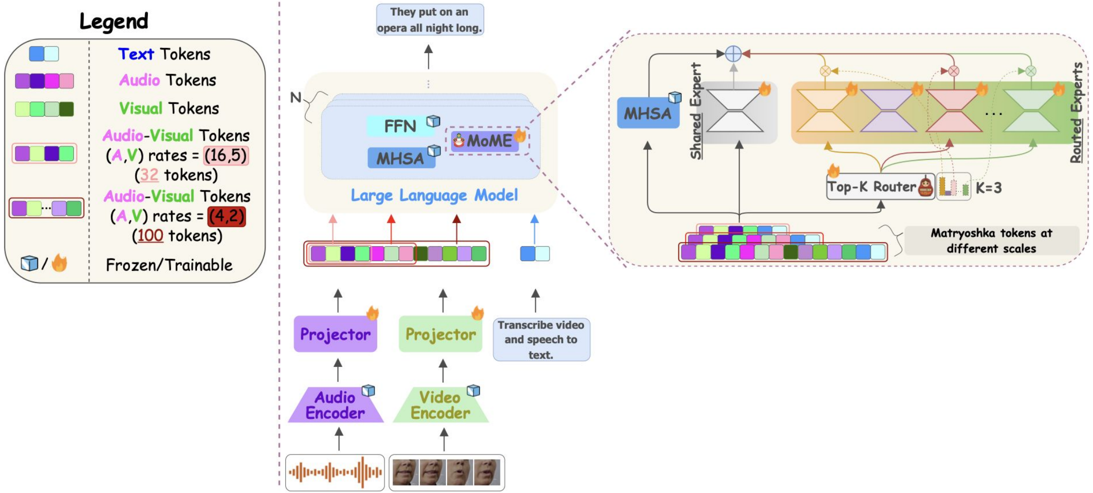
MoME: Mixture of Matryoshka Experts

- **MoME** is inserted in parallel with existing LLM layers (MHSA, FFN, entire layer), allowing efficient fine-tuning of a frozen pre-trained module.
- In addition to routed experts, **MoME** uses one *shared* expert to capture global and scale-invariant knowledge.



MoME: Mixture of Matryoshka Experts

- **MoME** is inserted in parallel with existing LLM layers (MHSA, FFN, entire layer), allowing efficient fine-tuning of a frozen pre-trained module.
- In addition to routed experts, **MoME** uses one shared expert to capture global and scale-invariant knowledge.
- Crucially, both the experts and router in each **MoME** module are shared across all Matryoshka sequences. This design encourages the router to activate similar subsets of routed experts across different granularities, creating implicit alignment.



AVSR Results

- Results on LRS2 & LRS3 datasets.

Method	Active Params	LRS2 Dataset (4,2)	LRS2 Dataset (4,5)	LRS2 Dataset (16,2)	LRS2 Dataset (16,5)	Active Params	LRS3 Dataset (4,2)	LRS3 Dataset (4,5)	LRS3 Dataset (16,2)	LRS3 Dataset (16,5)
Llama-AVSR [14]‡	27.5M	4.1	4.5	5.3	8.1	6.8M	2.4	2.8	3.3	4.1
Llama-MTSK MS [23]	27.5M	4.8	5.9	6.4	8.9	8.1M	2.6	2.7	3.7	4.1
Llama-MTSK SS [23]	27.5M	3.4	4.7	4.8	6.4	8.1M	2.3	2.2	3.3	3.6
Llama-MTSK MSS [23]	55.0M	3.6	4.8	6.1	9.0	13.6M	2.4	2.4	3.2	3.5
MoME-23/4-MHSA-I‡	12.7M	3.2	3.1	4.9	5.3	3.5M	2.1	1.9	3.3	3.7
MoME-23/4-FFN	12.7M	3.2	3.1	4.5	4.6	3.5M	2.1	2.2	4.0	4.0
MoME-23/4-MHSA	12.7M	2.9	3.0	**4.2**	4.3	3.5M	1.8	**1.7**	**2.9**	**2.9**
MoME-23/4-LAYER	12.7M	**2.7**	**2.7**	**4.2**	**4.2**	3.5M	**1.5**	1.8	3.1	3.2
MoME-23/4-LAYER	**2.3M**	3.0	3.2	4.3	4.7	**0.9M**	2.0	2.2	3.2	3.7

AVSR Results

Method	Active Params	LRS2 Dataset				Active Params	LRS3 Dataset			
		(4,2)	(4,5)	(16,2)	(16,5)		(4,2)	(4,5)	(16,2)	(16,5)
Llama-AVSR [14]‡	27.5M	4.1	4.5	5.3	8.1	6.8M	2.4	2.8	3.3	4.1
Llama-MTSK MS [23]	27.5M	4.8	5.9	6.4	8.9	8.1M	2.6	2.7	3.7	4.1
Llama-MTSK SS [23]	27.5M	3.4	4.7	4.8	6.4	8.1M	2.3	2.2	3.3	3.6
Llama-MTSK MSS [23]	55.0M	3.6	4.8	6.1	9.0	13.6M	2.4	2.4	3.2	3.5
MoME-23/4-MHSA-I‡	12.7M	3.2	3.1	4.9	5.3	3.5M	2.1	1.9	3.3	3.7
MoME-23/4-FFN	12.7M	3.2	3.1	4.5	4.6	3.5M	2.1	2.2	4.0	4.0
MoME-23/4-MHSA	12.7M	2.9	3.0	**4.2**	4.3	3.5M	1.8	**1.7**	**2.9**	**2.9**
MoME-23/4-LAYER	12.7M	**2.7**	**2.7**	**4.2**	**4.2**	3.5M	**1.5**	1.8	3.1	3.2
MoME-23/4-LAYER	**2.3M**	3.0	3.2	4.3	4.7	**0.9M**	2.0	2.2	3.2	3.7

- Results on LRS2 & LRS3 datasets.
- Results across multiple compression rates.

AVSR Results

Method	Active Params	LRS2 Dataset (4,2)	LRS2 (4,5)	LRS2 (16,2)	LRS2 (16,5)	Active Params	LRS3 Dataset (4,2)	LRS3 (4,5)	LRS3 (16,2)	LRS3 (16,5)
Llama-AVSR [14]‡	27.5M	4.1	4.5	5.3	8.1	6.8M	2.4	2.8	3.3	4.1
Llama-MTSK MS [23]	27.5M	4.8	5.9	6.4	8.9	8.1M	2.6	2.7	3.7	4.1
Llama-MTSK SS [23]	27.5M	3.4	4.7	4.8	6.4	8.1M	2.3	2.2	3.3	3.6
Llama-MTSK MSS [23]	55.0M	3.6	4.8	6.1	9.0	13.6M	2.4	2.4	3.2	3.5
MoME-23/4-MHSA-I‡	12.7M	3.2	3.1	4.9	5.3	3.5M	2.1	1.9	3.3	3.7
MoME-23/4-FFN	12.7M	3.2	3.1	4.5	4.6	3.5M	2.1	2.2	4.0	4.0
MoME-23/4-MHSA	12.7M	2.9	3.0	**4.2**	4.3	3.5M	1.8	**1.7**	**2.9**	**2.9**
MoME-23/4-LAYER	12.7M	**2.7**	**2.7**	**4.2**	**4.2**	3.5M	**1.5**	1.8	3.1	3.2
MoME-23/4-LAYER	**2.3M**	3.0	3.2	4.3	4.7	**0.9M**	2.0	2.2	3.2	3.7

- Results on LRS2 & LRS3 datasets.
- Results across multiple compression rates.
- MoME consistently outperforms the other baselines across all rates.

AVSR Results

Method	Active Params	LRS2 Dataset (4,2)	(4,5)	(16,2)	(16,5)	Active Params	LRS3 Dataset (4,2)	(4,5)	(16,2)	(16,5)
Llama-AVSR [14]‡	27.5M	4.1	4.5	5.3	8.1	6.8M	2.4	2.8	3.3	4.1
Llama-MTSK MS [23]	27.5M	4.8	5.9	6.4	8.9	8.1M	2.6	2.7	3.7	4.1
Llama-MTSK SS [23]	27.5M	3.4	4.7	4.8	6.4	8.1M	2.3	2.2	3.3	3.6
Llama-MTSK MSS [23]	55.0M	3.6	4.8	6.1	9.0	13.6M	2.4	2.4	3.2	3.5
MoME-23/4-MHSA-I‡	12.7M	3.2	3.1	4.9	5.3	3.5M	2.1	1.9	3.3	3.7
MoME-23/4-FFN	12.7M	3.2	3.1	4.5	4.6	3.5M	2.1	2.2	4.0	4.0
MoME-23/4-MHSA	12.7M	2.9	3.0	**4.2**	4.3	3.5M	1.8	**1.7**	**2.9**	**2.9**
MoME-23/4-LAYER	12.7M	**2.7**	**2.7**	**4.2**	**4.2**	3.5M	**1.5**	1.8	3.1	3.2
MoME-23/4-LAYER	**2.3M**	3.0	3.2	4.3	4.7	**0.9M**	2.0	2.2	3.2	3.7

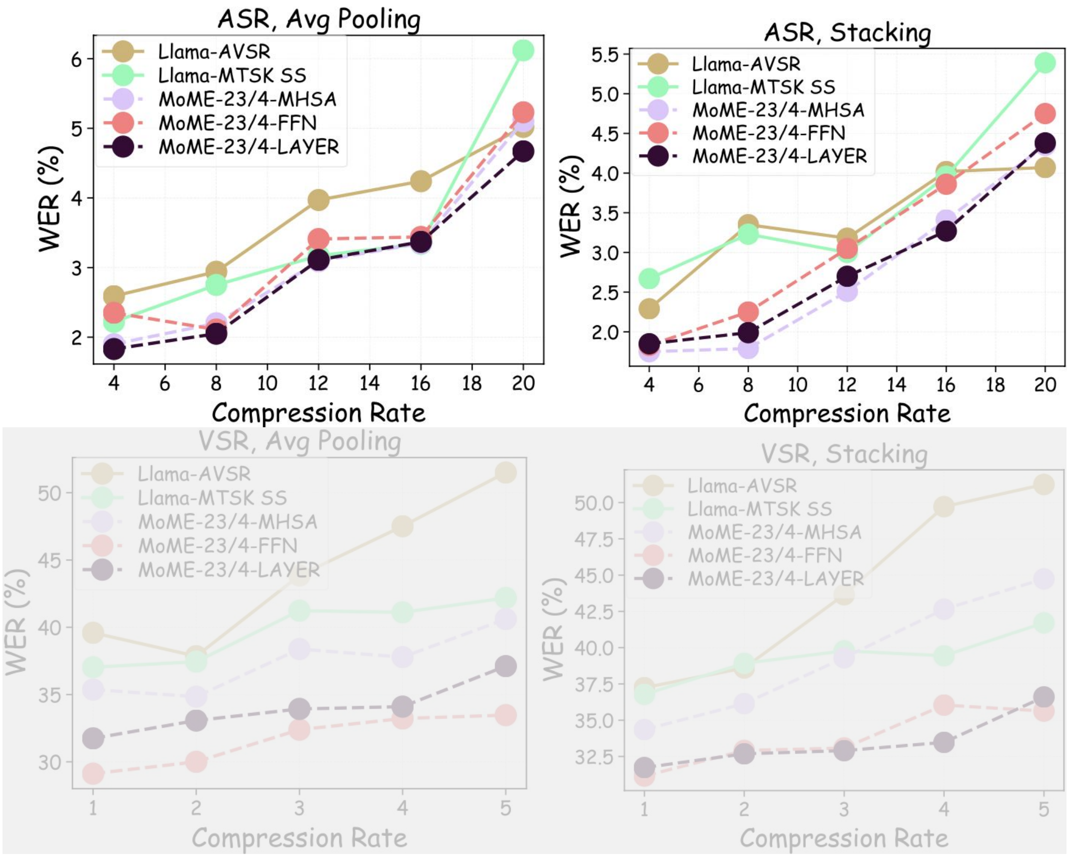
- Results on LRS2 & LRS3 datasets.
- Results across multiple compression rates.
- MoME consistently outperforms the other baselines across all rates.
- MHSA/LAYER configurations bring the best results.

AVSR Results

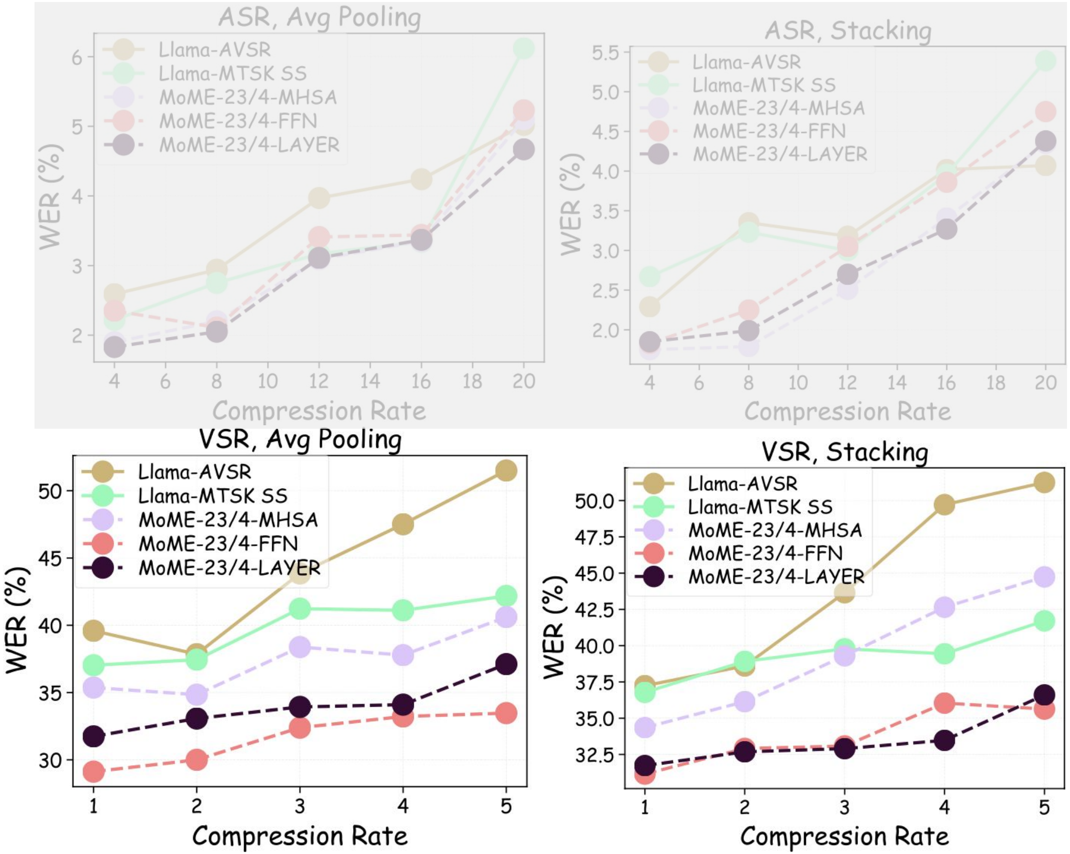
Method	Active Params	LRS2 Dataset (4,2)	(4,5)	(16,2)	(16,5)	Active Params	LRS3 Dataset (4,2)	(4,5)	(16,2)	(16,5)
Llama-AVSR [14]‡	27.5M	4.1	4.5	5.3	8.1	6.8M	2.4	2.8	3.3	4.1
Llama-MTSK MS [23]	27.5M	4.8	5.9	6.4	8.9	8.1M	2.6	2.7	3.7	4.1
Llama-MTSK SS [23]	27.5M	3.4	4.7	4.8	6.4	8.1M	2.3	2.2	3.3	3.6
Llama-MTSK MSS [23]	55.0M	3.6	4.8	6.1	9.0	13.6M	2.4	2.4	3.2	3.5
MoME-23/4-MHSA-I‡	12.7M	3.2	3.1	4.9	5.3	3.5M	2.1	1.9	3.3	3.7
MoME-23/4-FFN	12.7M	3.2	3.1	4.5	4.6	3.5M	2.1	2.2	4.0	4.0
MoME-23/4-MHSA	12.7M	2.9	3.0	**4.2**	4.3	3.5M	1.8	**1.7**	**2.9**	**2.9**
MoME-23/4-LAYER	12.7M	**2.7**	**2.7**	**4.2**	**4.2**	3.5M	**1.5**	1.8	3.1	3.2
MoME-23/4-LAYER	2.3M	3.0	3.2	4.3	4.7	0.9M	2.0	2.2	3.2	3.7

- Results on LRS2 & LRS3 datasets.
- Results across multiple compression rates.
- MoME consistently outperforms the other baselines across all rates.
- MHSA/LAYER configurations bring the best results.
- **We can push the active params down to 2.3/0.9M with minimal performance degradation, ensuring extreme parameter-efficient fine-tuning.**

ASR Results

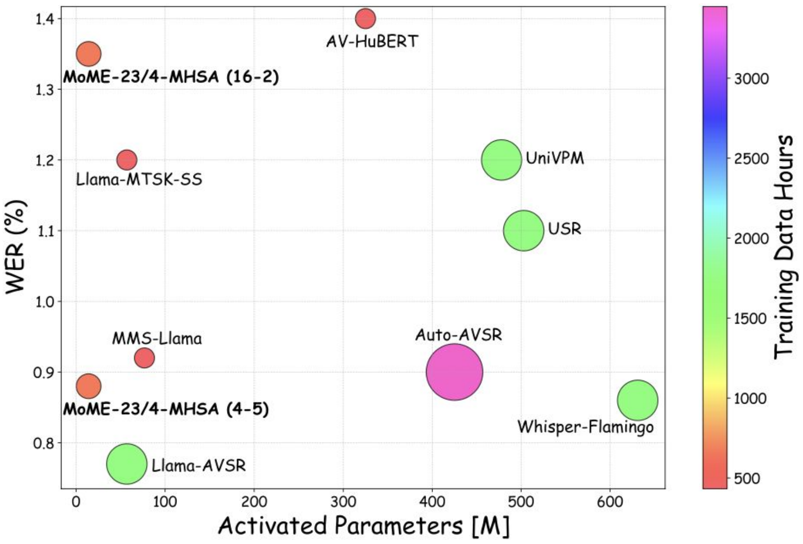

ASR, Avg Pooling
ASR, Stacking
VSR, Avg Pooling
VSR, Stacking
Llama-AVSR
Llama-MTSK SS
MoME-23/4-MHSA
MoME-23/4-FFN
MoME-23/4-LAYER
WER (%)
Compression Rate

VSR Results



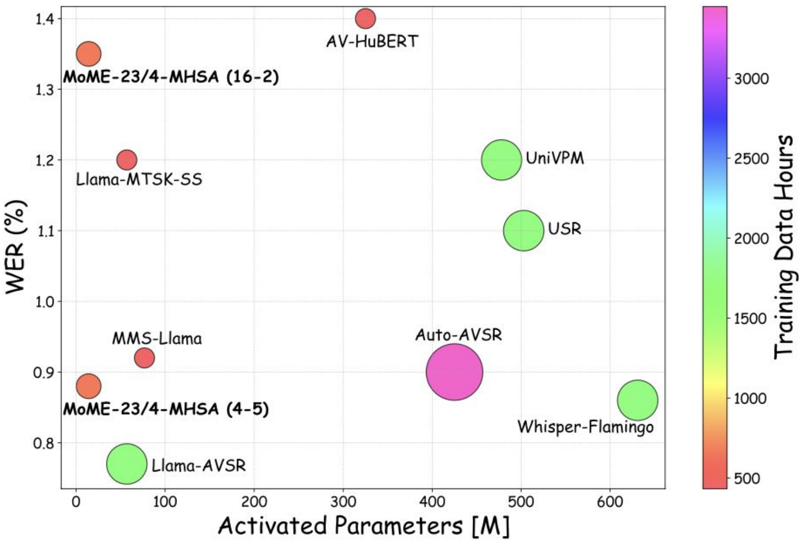
ASR, Avg Pooling
ASR, Stacking
VSR, Avg Pooling
VSR, Stacking
Llama-AVSR
Llama-MTSK SS
MoME-23/4-MHSA
MoME-23/4-FFN
MoME-23/4-LAYER
WER (%)
Compression Rate

AVSR Comparison with SoTA Methods



When comparing MoME with SoTA AVSR methods, it achieves competitive results while activating significantly fewer parameters and requiring fewer training data hours.

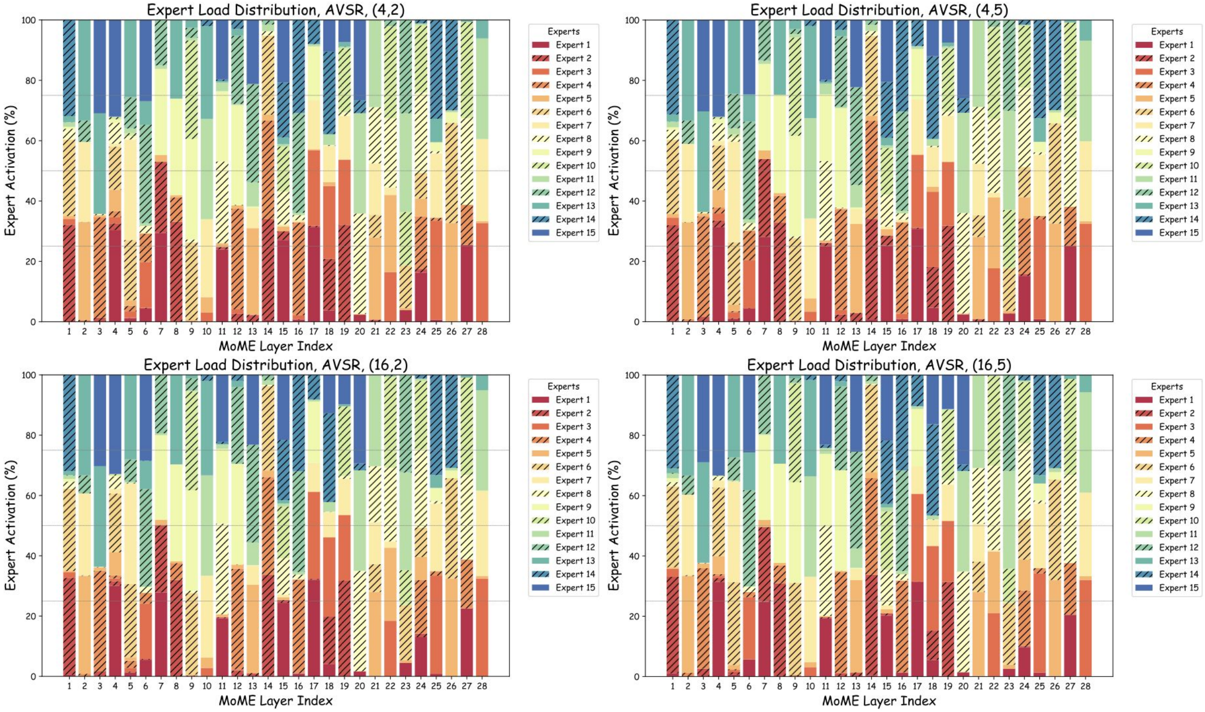
Robustness to Noise



Method	SNR (dB)				
	7.5	5	2.5	0	-5
Llama-AVSR [14]	5.6	7.1	10.6	11.8	41.8
Llama-MTSK MS [23]	6.2	8.0	13.0	12.4	44.9
MoME-23/4-LAYER	**4.8**	**6.4**	**9.6**	**9.6**	**32.6**

MoME exhibits greater resilience to noise compared to prior methods, with particularly strong gains in highly degraded scenarios.

Expert Activation Analysis



Thanks to MoME, the same subset of experts tends to be activated across different token granularities, demonstrating strong alignment in routing behavior.

Further details and results can be found in our paper!

SCAN ME