



# Scaling and Enhancing LLM-based AVSR: A Sparse Mixture of Projectors Approach

Umberto Cappellazzo<sup>1</sup>, Minsu Kim<sup>2</sup>, Stavros Petridis<sup>1</sup>, Daniele Falavigna<sup>3</sup>, Alessio Brutti<sup>3</sup>

<sup>1</sup>Imperial College London, UK

<sup>2</sup>Meta AI, UK

<sup>3</sup>Fondazione Bruno Kessler, Italy

u.cappellazzo@imperial.ac.uk

## Abstract

Audio-Visual Speech Recognition (AVSR) enhances robustness in noisy environments by integrating visual cues. While recent advances integrate Large Language Models (LLMs) into AVSR, their high computational cost hinders deployment in resource-constrained settings. To address this, we propose Llama-SMoP, an efficient Multimodal LLM that employs a Sparse Mixture of Projectors (SMoP) module to scale model capacity without increasing inference costs. By incorporating sparsely-gated mixture-of-experts (MoE) projectors, Llama-SMoP enables the use of smaller LLMs while maintaining strong performance. We explore three SMoP configurations and show that Llama-SMoP DEDR (Disjoint-Experts, Disjoint-Routers), which uses modality-specific routers and experts, achieves superior performance on ASR, VSR, and AVSR tasks. Ablation studies confirm its effectiveness in expert activation, scalability, and noise robustness.

**Index Terms:** Audio-Visual Speech Recognition, Multimodal LLMs, Mixture of Experts, Soft Mixture of Projectors

## 1. Introduction

Automated speech recognition technologies have made significant progress and are widely employed in various real-world applications [1]. In particular, Auditory Speech Recognition (ASR) technology [2, 3], which uses audio as its input modality, is the most widely used and recognized by users. However, in real-world scenarios, audio can be corrupted by various background noises (e.g., speech captured in a crowded restaurant), leading recent research to focus on improving ASR robustness [4, 5]. One key approach is leveraging multimodal inputs by integrating both audio and visual modalities. This technology, known as Audio-Visual Speech Recognition (AVSR) [6–12], achieves robust recognition even when input audio is severely corrupted with background noise by leveraging the correlation between audio and visual cues.

One chief research direction focuses on pre-training both audio and visual encoders through Self-Supervised Learning (SSL) [13] to capture the intrinsic correlation between audio-visual modalities. This approach has demonstrated impressive performance even with limited labeled data (e.g., 30 hours), as reported in prior studies [14–18]. Building on this foundation, recent advances in generative AI and Large Language Models (LLMs) [19–21] have given rise to a new research line: aligning speech representations with LLMs [22–26]. Notably, several studies have successfully adapted LLMs for automated speech recognition, including [27–33]. However, despite their impressive performance, LLM-based speech recognition systems face

Only non-Meta authors conducted any of the dataset preprocessing (no dataset pre-processing took place on Meta’s servers or facilities).

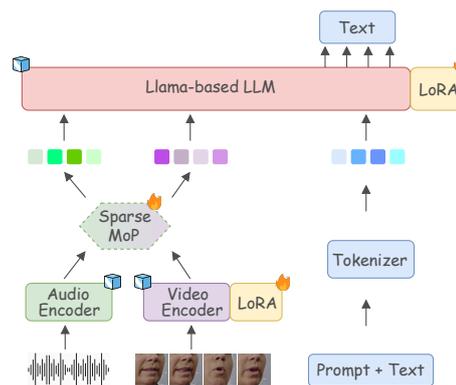


Figure 1: Illustration of the overall framework of the proposed Llama-SMoP model, where audio and video tokens are embedded using a sparsely-gated mixture-of-experts scheme. and represent whether the module is trained or kept frozen.

a significant challenge: they require a large number of parameters. Recent works, such as Llama-AVSR [33] and Llama-MTSK [34], have shown that larger LLMs generally achieve better speech recognition performance, which is why previous works tend to use LLMs with over 7 billion parameters. This poses significant challenges, as these large-scale models cannot be easily deployed on parameter-constrained settings.

In this paper, we investigate LLM-based AVSR systems, particularly focusing on the use of smaller LLMs (i.e., 1B and 3B parameters). We propose leveraging the Mixture of Experts (MoE) paradigm [35–40] within the audio-visual projector to enhance the model capacity while keeping inference costs low. Recently, the integration of MoE blocks into Multimodal LLMs (MLLMs) has gained attention as a strategy for scaling encoders and LLMs. However, in LLM-based AVSR models, pre-trained encoders and LLMs are typically frozen during training [33]. To address this, we propose scaling the projector, an MLP layer, aligning with the trend of transforming LLMs’ MLP layers in MoE-based models. While MoE has been applied to projectors in vision-language tasks [41] and chart understanding [42], these methods primarily emphasize single-modality inputs and efficient initialization through co-upcycling or task-specific alignment techniques.

In this work, we introduce a novel module called **Sparse Mixture of Projectors (SMoP)** for embedding multimodal speech representations into the LLM space. Specifically, we investigate three different SMoP variants for processing the audio-visual tokens, based on different router and expert projector configurations. We concentrate on audio-visual models that utilize small-scale pre-trained encoders and LLMs, which typically exhibit lower performance compared to their large-scale counterparts. SMoP is particularly advantageous in these

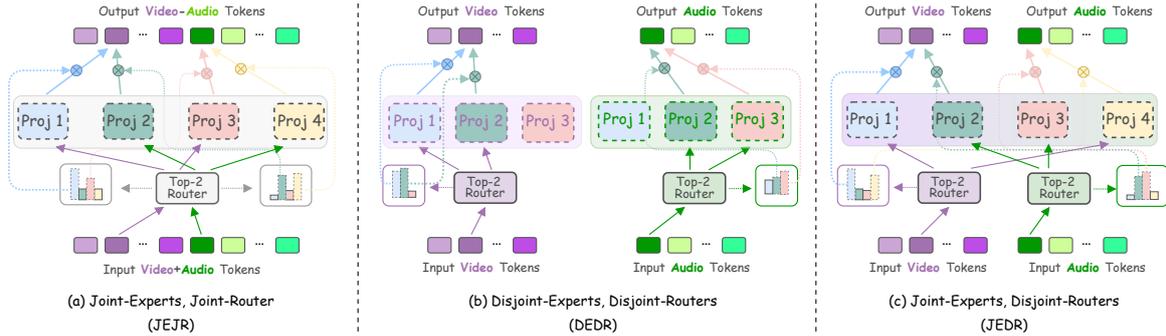


Figure 2: Detailed illustration of the three proposed SMoP configurations. (a) Joint-Experts, Joint-Router (JEJR) uses one multimodal router and one pool of expert for embedding audio-visual representations. (b) Disjoint-Experts, Disjoint-Routers (DEDR) uses modality-specific routers and experts for embedding modality-specific representations. (c) Joint-Experts, Disjoint-Routers (JEDR) uses modality-specific routers and one shared group of experts, so routers assign the Top-K experts considering the modality characteristics.

resource-constrained settings, as it improves performance using a small number of experts while incurring negligible additional parameter activation and computational overhead in inference.

We summarize the contributions of this work as follows:

- We propose Llama-SMoP, an MLLM employing Sparse MoE to enhance audio-visual capabilities. The SMoP module is simple, efficient, and model-agnostic, allowing seamless integration with various pre-trained encoders and LLMs.
- Among the three proposed configurations, Llama-SMoP DEDR (Disjoint-Experts, Disjoint-Routers) achieves the best performance, outperforming previous methods across different-sized Llama-based LLMs on the AVSR task. We also demonstrate that Llama-SMoP is effective for both ASR and VSR tasks.
- We conduct ablation studies on expert activation frequency and optimal expert configurations, as well as showing that Llama-SMoP remains robust also in noisy scenarios.

## 2. Llama-SMoP

We propose Llama-SMoP, an MLLM employing sparsely-gated mixture-of-experts [35, 36] to increase model capacity without a proportional increase in computational cost. This is crucial in resource-constrained LLM-based AVSR systems, as we aim to improve performance despite using smaller-scale LLMs and pre-trained encoders. Llama-SMoP computes audio and video tokens via modality-specific pre-trained encoders, which are then fed to the LLM as prefix tokens (together with the textual tokens). This approach, denoted as decoder-only, is adopted by several architectures due to its versatility [43–47]. Llama-SMoP consists of three main components: **1)** pre-trained audio and video encoders, **2)** a SMoP module, and **3)** an LLM parameter-efficiently fine-tuned via LoRA [48].

**Audio/Video Pre-Trained Encoders.** We use pre-trained audio and video encoders to project the input audio and video data into audio tokens  $\mathbf{X}^A$  and video tokens  $\mathbf{X}^V$ . The pre-trained encoders are maintained *frozen* during the training stage.

**SMoP Module.** The SMoP module replaces the standard projector with sparsely-gated MoE projectors [35], each consisting of a two-layer MLP. Since we handle both audio and video modalities, we propose three router design strategies for processing the multimodal tokens. Before describing them in detail, we define how a generic SMoP module works.

To scale up the model with multiple projectors in parallel, a SMoP module encompasses a router network  $\mathcal{R}$  which chooses the top-K expert projectors out of the total N experts  $\{E_i\}_{i=1}^N$ , and thus learns the optimal distribution over the experts for each token. For a given token  $\mathbf{x}$ , the router  $\mathcal{R}$  picks the top-K experts

based on the highest scores obtained using a learnable gating function (in our case a linear layer parameterized by  $\mathbf{W}$ ), which are normalized via Softmax. The final output  $\mathbf{z}$  is the linearly weighted combination of each expert’s output scaled by the corresponding gate’s output as follows:

$$\mathbf{z} = \sum_{i=1}^N \mathcal{R}(\mathbf{x})_i \cdot E_i(\mathbf{x}), \quad (1)$$

$$\mathcal{R}(\mathbf{x}) = \text{Top-K}(\text{Softmax}(\mathbf{x} \cdot \mathbf{W}), \mathbf{K}), \quad (2)$$

$$\text{Top-K}(\mathbf{h}, \mathbf{K}) = \begin{cases} \mathbf{h}, & \text{if } \mathbf{h} \text{ is in the Top-K,} \\ 0, & \text{otherwise.} \end{cases} \quad (3)$$

The SMoP module processes the audio  $\mathbf{X}^A$  and video  $\mathbf{X}^V$  tokens, obtaining the output audio-visual tokens  $\mathbf{Z}^{AV}$ .

We investigate three different SMoP methods based on how we process the audio and video tokens, as shown in Fig. 2. The first approach, denoted *Joint-Experts, Joint-Router* (JEJR; Fig. 2a), defines a *single joint* audio-visual router  $\mathcal{R}^{AV}$  that dispatches the concatenation of audio and video tokens to a *shared* pool of N audio-visual experts  $\{E_i^{AV}\}_{i=1}^N$ . This method combines audio and video tokens into a single representation, allowing the model to learn cross-modal interactions directly. While JEJR can be effective for capturing these cross-modal interactions, it may lose some modality-specific details.

In contraposition to JEJR, we explore the *Disjoint-Experts, Disjoint-Routers* (DEDR; Fig. 2b) configuration. It utilizes *modality-specific* routers,  $\mathcal{R}^A$  for audio and  $\mathcal{R}^V$  for video, to direct each modality’s tokens to its *own dedicated* set of experts ( $\{E_i^A\}_{i=1}^{N_A}$  and  $\{E_i^V\}_{i=1}^{N_V}$ , respectively). Output audio and video tokens are then concatenated before being processed by the LLM. This modular approach enables each expert pool to specialize in processing a specific modality, thereby maximizing the extraction and utilization of modality-specific representations. However, because modalities are processed separately, this approach may hinder the model’s ability to learn cross-modal interactions. Finally, we also study a hybrid approach between the two previous ones: *Joint-Experts, Disjoint-Routers* configuration (JEDR; Fig. 2c). *Modality-specific* routers,  $\mathcal{R}^A$  and  $\mathcal{R}^V$ , assign each modality’s tokens to a *shared* pool of experts  $\{E_i^{AV}\}_{i=1}^N$ . This setup allows the model to leverage a common pool of expertise while maintaining the distinctiveness of each modality during routing.

For the JEJR and JEDR configurations, we must guarantee that audio and video tokens have the same token hidden dimension because they are processed by the same group of experts. If the hidden dimension of audio and video encoders does not match, we add a linear layer to compensate for this.

Table 1: ASVR results of our three proposed SMOp configurations and baselines on LRS3. SMOp-Y means that each pool of experts contains Y experts. **Bold** and underline numbers denote the best and second best scores, respectively.

Method	Llama Model		
	3.2-1B	3.2-3B	3.1-8B
Llama-AVSR [33]	3.81	2.80	<u>1.09</u>
DCI [43]	3.46	2.60	1.28
MM-Fuser [50]	3.45	2.66	1.31
<b>SMoP-4</b> JEJR	3.97	2.63	1.53
<b>SMoP-4</b> JEDR	4.16	2.80	1.23
<b>SMoP-3</b> DEDR	<b>3.31</b>	<b>2.29</b>	<b>0.96</b>
<b>SMoP-4-V</b> + DCI-A	3.34	3.05	2.79
<b>SMoP-4-V</b> + MM-Fuser-A	3.45	<u>2.51</u>	1.86

**LLM.** The role of LLM in our work is generating the corresponding speech-recognized transcription  $\mathbf{Y} = \{y_l\}_{l=1}^L$  in an auto-regressive manner, conditioned on the input audio-visual tokens  $\mathbf{Z}^{\text{AV}}$  and prompt textual tokens  $\mathbf{X}^P$ , where  $L$  represents the number of tokens of the ground truth transcription. Accordingly, the probability of the target  $\mathbf{Y}$  is computed by:

$$p(\mathbf{Y}|\mathbf{Z}^{\text{AV}}, \mathbf{X}^P) = \prod_{l=1}^L p(y_l|\mathbf{Z}^{\text{AV}}, \mathbf{X}^P, y_{<l}), \quad (4)$$

where  $y_{<l}$  is the previous generated output sequence.

**Total Loss.** Besides the next-token prediction loss of the LLM, we include two auxiliary losses, which are commonly used to avoid the router to activate only a few experts, dispensing with expert imbalance issues. The auxiliary losses comprise the load balancing loss  $\mathcal{L}_b$  [35], which penalizes unequal assignment of the experts, and the router z-loss  $\mathcal{L}_z$  [49], which penalizes large logits in the router that may cause instabilities. Finally, the total loss is defined as:

$$\mathcal{L} = -\log p(\mathbf{Y}|\mathbf{Z}^{\text{AV}}, \mathbf{X}^P) + \alpha_b \mathcal{L}_b + \alpha_z \mathcal{L}_z, \quad (5)$$

with  $\alpha_b$  and  $\alpha_z$  set to 0.01 and 0.001 following [35, 40, 49].

## 3. Experiments and Results

### 3.1. Implementation Details

**Datasets.** We train and evaluate Llama-SMOp on LRS3 [51], the largest publicly available dataset for AVSR. LRS3 contains 433 hours of transcribed English video clips from TED talks.

**Pre-Processing.** We follow [11, 33] for the pre-processing of the dataset. For the video modality, we crop the mouth region of interests (ROIs) through a bounding box of  $96 \times 96$ . Each frame is normalised by subtracting the mean and dividing by the standard deviation of the training set. Audio data is preprocessed with z-normalisation per utterance.

**Tasks.** The AVSR task is studied for the main results, and we also report the results for the ASR/VSR tasks in Section 3.4.

**Llama-SMOp Details.** We use multiple Whisper models [52] (Tiny, Base, Small, Medium) as pre-trained audio encoders, whilst AV-HuBERT Large [14] is used for computing the video tokens for all experiments. Their weights remain frozen throughout the training phase, and only for the VSR task do we equip the video encoder with LoRA modules following [33]. Each expert projector is a two-layer MLP. We use 4 shared experts for the JEJR and JEDR configurations (e.g., **SMoP-4** JEJR in Table 1), while 3 experts for each modality-based SMOp module in DEDR, resulting in 6 experts overall. We use Token-Choice Top-K as routing strategy [35, 36, 40],

Table 2: ASR/AVSR results under different acoustic noise levels.

Method	Modality	SNR Level (dB)				
		7.5	5	2.5	0	-2.5
Llama-AVSR [33]	A	6.3	10.0	19.1	35.1	95.1
DCI [43]	A	6.2	10.1	18.7	33.3	93.0
MM-Fuser [50]	A	5.8	9.3	17.6	33.0	89.4
<b>SMoP-4</b>	A	6.1	9.8	18.3	31.4	87.5
Llama-AVSR [33]	A-V	5.2	7.1	10.4	11.3	27.5
DCI [43]	A-V	4.9	6.8	9.6	9.8	23.1
MM-Fuser [50]	A-V	4.7	6.1	9.3	9.8	24.5
<b>SMoP-3</b> DEDR	A-V	4.5	6.1	8.9	9.5	22.9

where we select and activate the top-K experts for each input token, with  $K = 2$ . As for the LLM, we experiment with 3 pre-trained models from the Llama 3 family of varying size [53]: Llama 3.2-1B, Llama 3.2-3B, and also Llama 3.1-8B to confirm the effectiveness of SMOp under a large-scale parameter setting. As in [33], we parameter-efficiently fine-tune the LLM via LoRA [48, 54, 55]. Following [29, 31, 33, 56], we reduce the number of tokens processed by the LLM by stacking multiple consecutive tokens along the hidden dimension. We apply a compression rate of 3 both for audio and video tokens.

**Baselines.** We compare Llama-SMOp with Llama-AVSR [33], which serves as a baseline using a single projector without any scaling method. Additionally, we compare it with two recent scaling methods that integrate deep and shallow intermediate features from pre-trained encoders (in our case, audio and video encoders) by implementing them into Llama-AVSR. Specifically, Dense Channel Integration (DCI) [43] incorporates features from all layers before concatenating them with the final layer’s features across the hidden dimension, whereas MM-Fuser [50] fuses features from different layers using an attention-based module. Both DCI and MM-Fuser have been shown to enhance the visual representations of existing vision-language MLLMs. To the best of our knowledge, we are the first to investigate their efficacy in AVSR.

**Training/Inference Details.** Following [11, 33], we augment visual inputs through horizontal flipping, random cropping, and adaptive time masking, while for audio we only apply adaptive time masking. We define the textual prompts as: “Transcribe {**task prompt**} to text.”, where **task prompt**  $\in$  {“speech”, “video”, “speech and video”}. We train our model for 10 epochs with the AdamW optimizer with cosine annealing scheduler and weight decay set to 0.1 using NVIDIA A100 GPUs. The learning rate is set to  $1e-3$  for ASR and AVSR tasks, and  $5e-4$  for VSR. For decoding, we use beam search with a beam width of 15 and temperature of 0.6. The evaluation metric is the Word Error Rate (WER, %).

### 3.2. AVSR Performance Evaluation

In Table 1, we compare AVSR results on LRS3 across our three SMOp strategies and the baseline models. Based on the ablation studies provided in Section 3.4 for ASR/VSR, we also report results for a hybrid approach that applies SMOp for video and DCI/MM-Fuser for audio. For Llama 3.2-1B and 3.2-3B, we use Whisper Base as the audio encoder, whereas for Llama 3.1-8B, we use Whisper Medium. Among the three SMOp variants, DEDR achieves the best results, improving the baseline Llama-AVSR across all LLM configurations and outperforming both DCI and MM-Fuser. SMOp-4 JEJR and JEDR provide little or no improvement over the Llama-AVSR baseline, suggesting that learning experts and routers independently is more

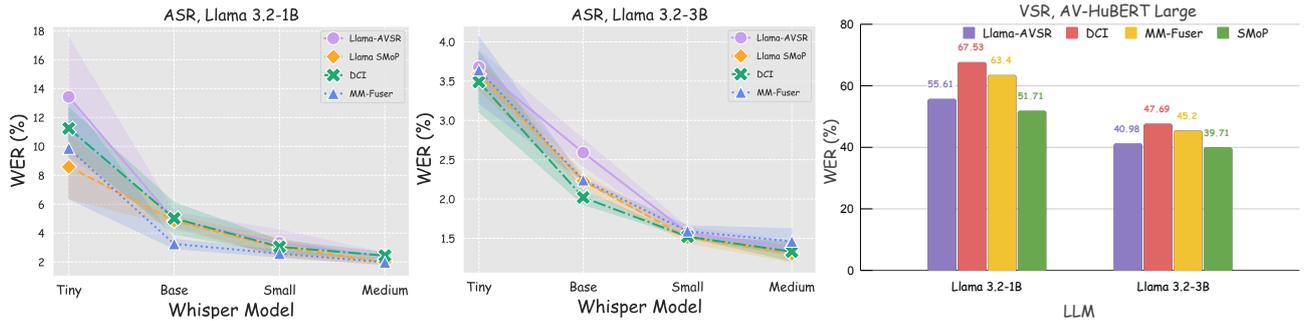


Figure 3: (Left). ASR results for Llama-SMoP using different-size Whisper models with Llama 3.2-1B. (Middle). ASR results for Llama-SMoP using different-size Whisper models with Llama 3.2-3B. (Right). VSR results for Llama-SMoP with Llama 3.2-1B/3.2-3B.

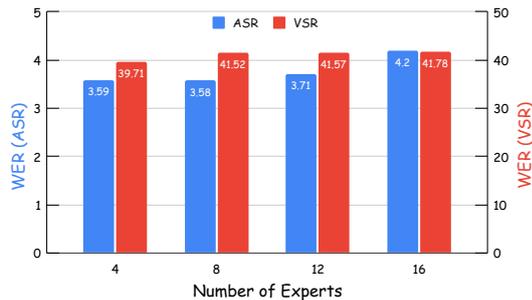


Figure 4: Ablation analysis on the number of expert projectors for the ASR and VSR tasks.

effective. We also observe that our proposed SMoP module can be effectively combined with MM-Fuser and DCI, achieving improvements over Llama-AVSR. Finally, in the best configuration with Llama 3.1-8B and Whisper Medium (last column), neither exploiting intermediate features (i.e., DCI and MM-Fuser) nor using SMoP leads to improvements, except for SMoP-3 DEDR. This demonstrates that these methods are highly effective when using smaller encoders and LLMs, making them well-suited for resource-constrained scenarios.

### 3.3. Robustness against Noise

In Table 2, we evaluate Llama-SMoP under noisy conditions. Following [33], we inject babble noise from the NOISEX dataset at various SNR levels in inference. We report results for the ASR and AVSR task, using Whisper Base and Llama 3.2-3B. For both tasks, Llama-SMoP proves to be the most resilient method, particularly as the noise level increases. Overall, both Llama-SMoP and DCI/MM-Fuser outperform Llama-AVSR.

### 3.4. Ablation on Model Parameter Scaling for ASR/VSR

In this section, we study the ASR and VSR tasks by varying the sizes of the audio encoder and LLM. In this case, we use a single modality-specific router and a single pool of experts.

**ASR Results.** For ASR, we investigate Llama-SMoP and DCI/MM-Fuser with Whisper models of different sizes for Llama 3.2-1B (Fig. 3, Left) and Llama 3.2-3B (Fig. 3, Middle). We observe that the WER improvement achieved by these methods over the baseline Llama-AVSR diminishes as the size of the Whisper encoder and LLM increases. This suggests that these methods are particularly beneficial when the model size is smaller, aligning with the AVSR trend observed in Table 1.

**VSR Results.** For the VSR task, Llama-SMoP achieves a nearly 4-point reduction in WER when using Llama 3.2-1B, demonstrating superior performance (Fig. 3, Right). In contrast,

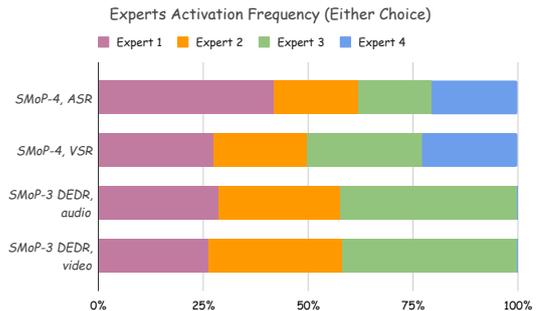


Figure 5: Proportion of tokens assigned to each expert, either as first or second choice.

both DCI and MM-Fuser degrade performance in both configurations. We attribute this to the inherently higher WER in VSR, which may exacerbate the ambiguity of visual speech (i.e., lip movements) when integrating intermediate video features.

### 3.5. Optimal Expert Configuration for SMoP

We study the optimal number of expert projectors for each SMoP module in ASR and VSR tasks, evaluating configurations with 4, 8, 12, and 16 experts. As shown in Fig. 4, increasing the number of experts for ASR and VSR does not improve performance; instead, it slightly degrades it, likely due to redundant learning of similar tokens by additional experts. Recent MoE-based AVSR models also use 4 [57] or 8 [58] experts. Finally, Fig. 5 presents the activation ratio of experts during inference, whether as the first or second choice. The router evenly activates the experts, ensuring all contribute to the computation of the output tokens. For ASR, one expert exhibits slightly higher activation, likely due to the relative simplicity of the task.

## 4. Conclusion

We present Llama-SMoP, an MLLM optimized for improved audio-visual processing. Its key innovation is replacing the linear projector with a Top-K sparse MoE module. This approach allows for more efficient processing of multimodal audio-visual tokens, and we investigate three SMoP designs based on varying router and expert configurations. Llama-SMoP<sub>DEDR</sub> achieves superior WER results across multiple tasks, pre-trained encoders and LLMs. The SMoP module offers a simple and effective way to scale LLM-based AVSR under resource-constrained settings, incurring minimal additional inference overhead while boosting performance.

**Acknowledgment:** this work was partially funded by the European Union’s Horizon 2020 project ELOQUENCE (grant 101070558).

## 5. References

- [1] Y. Zhang *et al.*, “Google usm: Scaling automatic speech recognition beyond 100 languages,” *arXiv preprint arXiv:2303.01037*, 2023.
- [2] D. Amodei *et al.*, “Deep speech 2: End-to-end speech recognition in english and mandarin,” in *ICML*. PMLR, 2016, pp. 173–182.
- [3] R. Prabhavalkar *et al.*, “End-to-end speech recognition: A survey,” *TASLP*, 2023.
- [4] S. Dupont *et al.*, “Audio-visual speech modeling for continuous speech recognition,” *IEEE transactions on multimedia*, 2000.
- [5] A. Narayanan *et al.*, “Investigation of speech separation as a front-end for noise robust speech recognition,” *TASLP*, 2014.
- [6] K. Noda *et al.*, “Audio-visual speech recognition using deep learning,” *Applied intelligence*, vol. 42, pp. 722–737, 2015.
- [7] T. Afouras *et al.*, “Deep audio-visual speech recognition,” *IEEE TPAMI*, vol. 44, no. 12, pp. 8717–8727, 2018.
- [8] S. Petridis *et al.*, “Audio-visual speech recognition with a hybrid ctc/attention architecture,” in *SLT*, 2018.
- [9] P. Ma *et al.*, “End-to-end audio-visual speech recognition with conformers,” in *ICASSP*, 2021.
- [10] J. Hong *et al.*, “Visual context-driven audio feature enhancement for robust end-to-end audio-visual speech recognition,” in *Interspeech*, 2022.
- [11] P. Ma *et al.*, “Auto-avs: Audio-visual speech recognition with automatic labels,” in *ICASSP*, 2023.
- [12] A. Rouditchenko *et al.*, “Whisper-flamingo: Integrating visual features into whisper for audio-visual speech recognition and translation,” in *Interspeech*, 2024.
- [13] J. Gui *et al.*, “A survey on self-supervised learning: Algorithms, applications, and future trends,” *TPAMI*, 2024.
- [14] B. Shi *et al.*, “Learning audio-visual speech representation by masked multimodal cluster prediction,” in *ICLR*, 2022.
- [15] A. Haliassos *et al.*, “Jointly learning visual and auditory speech representations from raw data,” in *ICLR*, 2023.
- [16] W.-N. Hsu and B. Shi, “u-hubert: Unified mixed-modal speech pretraining and zero-shot transfer to unlabeled modality,” *NeurIPS*, vol. 35, pp. 21 157–21 170, 2022.
- [17] A. Haliassos *et al.*, “Braven: Improving self-supervised pre-training for visual and auditory speech recognition,” in *ICASSP*, 2024.
- [18] —, “Unified speech recognition: A single model for auditory, visual, and audiovisual inputs,” in *NeurIPS*, 2024.
- [19] J. Achiam *et al.*, “Gpt-4 technical report,” *arXiv preprint arXiv:2303.08774*, 2023.
- [20] H. Touvron *et al.*, “Llama: Open and efficient foundation language models,” *arXiv preprint arXiv:2302.13971*, 2023.
- [21] H. Liu *et al.*, “Improved baselines with visual instruction tuning,” in *CVPR*, 2024.
- [22] K. Lakhotia *et al.*, “On generative spoken language modeling from raw audio,” *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 1336–1354, 2021.
- [23] R. Huang *et al.*, “Audiogpt: Understanding and generating speech, music, sound, and talking head,” in *AAAI*, 2024.
- [24] S. Park *et al.*, “Let’s go real talk: Spoken dialogue model for face-to-face conversation,” in *ACL*, 2024.
- [25] K. Lu *et al.*, “Developing instruction-following speech language model without speech instruction-tuning data,” in *ICASSP*, 2025.
- [26] W. Tan *et al.*, “Ssr: Alignment-aware modality connector for speech language models,” *arXiv preprint arXiv:2410.00168*, 2024.
- [27] C. Chen *et al.*, “It’s never too late: Fusing acoustic information into large language models for automatic speech recognition,” in *ICLR*, 2024.
- [28] Y. Hu *et al.*, “Large language models are efficient learners of noise-robust speech recognition,” in *ICLR*, 2024.
- [29] Z. Ma *et al.*, “An embarrassingly simple approach for llm with strong asr capacity,” *arXiv preprint arXiv:2402.08846*, 2024.
- [30] W. Yu *et al.*, “Connecting speech encoder and large language model for asr,” in *ICASSP*, 2024.
- [31] Y. Fathullah *et al.*, “Prompting large language models with speech recognition abilities,” in *ICASSP*, 2024.
- [32] J. Yeo *et al.*, “Where visual speech meets language: Vsp-llm framework for efficient and context-aware visual speech processing,” in *Findings of the EMNLP*, 2024, pp. 11 391–11 406.
- [33] U. Cappellazzo *et al.*, “Large language models are strong audio-visual speech recognition learners,” in *ICASSP*, 2025.
- [34] U. Cappellazzo, M. Kim, and S. Petridis, “Adaptive audio-visual speech recognition via matryoshka-based multimodal llms,” *arXiv preprint arXiv:2503.06362*, 2025.
- [35] N. Shazeer *et al.*, “Outrageously large neural networks: The sparsely-gated mixture-of-experts layer,” in *ICLR*, 2016.
- [36] D. Lepikhin *et al.*, “Gshard: Scaling giant models with conditional computation and automatic sharding,” in *ICLR*, 2021.
- [37] U. Cappellazzo *et al.*, “Efficient fine-tuning of audio spectrogram transformers via soft mixture of adapters,” in *Interspeech*, 2024.
- [38] A. Liu *et al.*, “Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model,” *arXiv preprint arXiv:2405.04434*, 2024.
- [39] X. O. He, “Mixture of a million experts,” *arXiv preprint arXiv:2407.04153*, 2024.
- [40] N. Muennighoff *et al.*, “Olmoe: Open mixture-of-experts language models,” in *ICLR*, 2025.
- [41] J. Li *et al.*, “Cummo: Scaling multimodal llm with co-upcycled mixture-of-experts,” in *NeurIPS*, 2024.
- [42] Z. Xu *et al.*, “Chartmoe: Mixture of expert connector for advanced chart understanding,” in *ICLR*, 2025.
- [43] H. Yao *et al.*, “Dense connector for mllms,” in *NeurIPS*, 2024.
- [44] H. Liu *et al.*, “Visual instruction tuning,” in *NeurIPS*, 2023.
- [45] J. Lin *et al.*, “Vila: On pre-training for visual language models,” in *CVPR*, 2024.
- [46] E. Fini *et al.*, “Multimodal autoregressive pre-training of large vision encoders,” *arXiv preprint arXiv:2411.14402*, 2024.
- [47] B. Lee *et al.*, “Meteor: Mamba-based traversal of rationale for large language and vision models,” in *NeurIPS*, 2024.
- [48] E. Hu *et al.*, “Lora: Low-rank adaptation of large language models,” in *ICLR*, 2021.
- [49] B. Zoph *et al.*, “St-moe: Designing stable and transferable sparse expert models,” *arXiv preprint arXiv:2202.08906*, 2022.
- [50] Y. Cao *et al.*, “Mmfuser: Multimodal multi-layer feature fuser for fine-grained vision-language understanding,” *arXiv preprint arXiv:2410.11829*, 2024.
- [51] T. Afouras *et al.*, “Lrs3-ted: a large-scale dataset for visual speech recognition,” *arXiv preprint arXiv:1809.00496*, 2018.
- [52] A. Radford *et al.*, “Robust speech recognition via large-scale weak supervision,” in *ICML*, 2023.
- [53] A. Dubey *et al.*, “The llama 3 herd of models,” *arXiv preprint arXiv:2407.21783*, 2024.
- [54] J. He *et al.*, “Towards a unified view of parameter-efficient transfer learning,” in *ICLR*, 2022.
- [55] U. Cappellazzo *et al.*, “Parameter-efficient transfer learning of audio spectrogram transformers,” in *IEEE MLSP*, 2024.
- [56] Q. Fang *et al.*, “Llama-omni: Seamless speech interaction with large language models,” *arXiv preprint arXiv:2409.06666*, 2024.
- [57] Y. Cheng, Y. Li, J. He *et al.*, “Mixtures of experts for audio-visual learning,” *NeurIPS*, vol. 37, pp. 219–243, 2024.
- [58] S. Kim *et al.*, “Mohave: Mixture of hierarchical audio-visual experts for robust speech recognition,” *arXiv preprint arXiv:2502.10447*, 2025.