MMS-LLaMA: Efficient LLM-based Audio-Visual Speech Recognition with Minimal Multimodal Speech Tokens

Anonymous ACL submission

Abstract

Audio-Visual Speech Recognition (AVSR) achieves robust speech recognition in noisy environments by combining auditory and visual information. However, recent Large Language Model (LLM) based AVSR systems incur high computational costs due to the high temporal resolution of audio-visual speech processed by LLMs. In this work, we introduce an efficient multimodal speech LLM framework that minimizes token length while preserving essential linguistic content. Our approach employs an early av-fusion module for streamlined feature integration, an audio-visual speech Q-Former that dynamically allocates tokens based on input duration, and a refined query allocation strategy with a speech rate predictor to adjust token allocation according to speaking speed of each audio sample. Extensive experiments on the LRS3 dataset show that our method achieves state-of-the-art performance with a WER of 0.74% while using only 2.8 tokens per second. Moreover, our approach not only reduces token usage by 88.8% compared to the previous multimodal speech LLM framework, but also improves computational efficiency by reducing FLOPs by 36%.

1 Introduction

In human communication, watching lip movements and listening to sounds are essential for understanding speech. These multimodal cues, which combine visual and auditory information, enable people to communicate effectively in bustling cafes, crowded streets, and noisy factories. Thanks to these practical advantages and significant advances in deep learning, Audio-Visual Speech Recognition (AVSR) technology has made remarkable progress through numerous research efforts (Shi et al., 2022a; Ma et al., 2023; Afouras et al., 2018a; Ma et al., 2021; Serdyuk et al., 2022; Cappellazzo et al., 2024). Now, it is easy to find an AVSR model that can accurately predict what you have said, even in noisy environments.

Such rapid progress has been made possible by large-scale audio-visual datasets (Afouras et al., 2018b,a); advanced neural network architectures such as RNNs (Elman, 1990), Transformers (Vaswani, 2017), and Conformers (Gulati et al., 2020); improved multimodal learning strategies, including self-supervised learning (Shi et al., 2022a) and knowledge distillation using a pre-trained Automatic Speech Recognition (ASR) model (Ma et al., 2023); carefully designed training methods (Ma et al., 2022; Hong et al., 2023); and the utilization of Large Language Models' (LLMs) language understanding capabilities as sentence predictors (Cappellazzo et al., 2024). Among these approaches, multimodal speech LLM frameworks have achieved remarkable performance by directly integrating with LLMs and unlocking their enhanced context modeling capabilities. Despite significant advancements in integrating LLMs into the multimodal speech domain, these approaches still suffer from high computational costs. This is largely due to multimodal tokens having a higher temporal resolution compared to text tokens, which forces the self-attention mechanism in each LLM layer to process many more tokens, thereby significantly increasing the computational burden.

To address these issues, we aim to develop an efficient multimodal speech LLM framework, namely MMS-LLaMA, for AVSR that minimizes the length of multimodal speech tokens while preserving their linguistic content. To achieve this, we construct this framework with three primary components: 1) early av-fusion module shifts the fusion process to an earlier stage, prior to inputting multimodal speech tokens into the LLM. 2) audio-visual speech Q-former (AV Q-Former) is designed to dynamically allocate the number of learnable queries according to the duration of audio-visual input, where queries are transformed into multimodal speech tokens. 3) Going one step further, to more

effectively adjust the number of multimodal tokens, we propose a speech rate predictor. It accounts for differences in speaking speed across each audio sample, enabling the assignment of more tokens to faster speech.

Reflecting the growing trend toward efficient multimodal models and leveraging insights from efficient Large Multimodal Models (LMMs) that utilize minimal vision tokens during pre-fusion (Zhang et al., 2025), we incorporate their effective pre-fusion approach into our multimodal speech LLM framework. Concurrently, we explore and integrate existing audio-visual fusion strategies into our early-AV fusion module, which reduces the number of fused sequences by half.

Even though the number of fused sequences is reduced by half, their temporal resolution remains higher than that of text tokens. To minimize the audio-visual sequence length to similar level of text, AV Q-Former is designed. Since the language content in audio-visual inputs scales with their duration, a fixed-length output from a conventional Q-Former (Dai et al., 2023) may be inefficient for compressing variable-length audio and visual modalities. To address this, we propose a novel query allocation strategy that adjusts the number of multimodal speech tokens based on the duration of the audio-visual input. Using this dynamic strategy, we empirically determine the minimum number of multimodal speech tokens required for variable-length inputs and compress the sequence to a scale comparable to that of text tokens. Finally, to further enhance efficiency, we refine our query allocation strategy by incorporating a speech rate predictor. Even if two audio-visual inputs have the same duration, their language content may differ depending on the speaker's speaking rate. By taking this into account, we can assign more queries to faster speech segments, thereby generating a greater number of multimodal speech tokens and preserving their linguistic content.

Through extensive experiments, we demonstrate that using only 2.8 multimodal speech tokens per second can effectively preserve language content while maintaining performance. Moreover, our approach achieves state-of-the-art performance with a Word Error Rate (WER) of 0.74% on the LRS3 dataset. Compared to the previous LLM-based multimodal speech framework (Cappellazzo et al., 2024) that uses 25 tokens per second, our method reduces token usage by 88.8%, employing only 2.8 tokens per second. Additionally, our approach improves computational efficiency, reducing FLOPs by 36.6%.

2 Related Work

2.1 Audio-Visual Speech Recognition

Audio-based Automatic Speech Recognition (ASR) is a well-studied and one of the most frequently used technology (Amodei et al., 2016; Kim et al., 2017; Prabhavalkar et al., 2023). However, as it primarily depends on acoustic inputs, its performance naturally degraded when the input audio is perturbed with other background noises (e.g., babble noise) (Varga and Steeneken, 1993). Visual Speech Recognition (VSR) aims to employ visual inputs only in speech recognition, using the fact that the visual inputs are not affected by the acoustic noises. A diverse range of literature has been explored on VSR techniques to avoid the impact of background noise on speech recognition performance (Petridis et al., 2017, 2018b; Martinez et al., 2020; Ma et al., 2021, 2022, 2023; Kim et al., 2022, 2023, 2024a,b; Yeo et al., 2024b,c,d).

Audio-Visual Speech Recognition (AVSR) integrates both audio and visual modalities as input, combining the strengths of ASR and VSR techniques to enhance the robustness and performance of speech recognition systems, particularly in noisy environments. Early AVSR models (Noda et al., 2015; Huang and Kingsbury, 2013; Mroueh et al., 2015; Stewart et al., 2013) established the foundation for multimodal speech recognition, demonstrating the effectiveness of fusing visual features such as lip movements with audio signals. Advancements in AVSR (Petridis et al., 2018a; Shi et al., 2022b; Hong et al., 2022, 2023) have been fueled by improvements in both data availability (Chung et al., 2017; Afouras et al., 2018b) and model architectures (Vaswani, 2017; Gulati et al., 2020). Self-supervised learning has also played a crucial role in advancing AVSR (Shi et al., 2022a; Lian et al., 2023; Haliassos et al., 2024b,a) to further enhance performance.

More recently, two research trends have emerged in the field. One focuses on leveraging pre-trained ASR models, with approaches like pseudo-labeling unlabeled audio-visual data for data augmentation (Ma et al., 2023; Yeo et al., 2024d) or integrating a pre-trained visual encoder with Whisper (Rouditchenko et al., 2024). The other trend integrates LLMs with speech features to harness the context modeling capabilities of LLMs, thereby enhancing

the recognition performance (Yu et al., 2024; Yeo et al., 2024a; Cappellazzo et al., 2024).

While recent efforts leveraging LLMs have achieved remarkable performance in speech recognition, they have primarily focused on further improving accuracy. In contrast, our goal is to limit the computational burden on our multimodal speech LLM without sacrificing accuracy, we introduce the AV Q-Former. By dynamically modifying the number of multimodal speech tokens, it effectively preserves essential linguistic details from both the audio and visual streams.

2.2 Speech Large Language Model

Early advancements in speech recognition began with monolingual English ASR models (Hsu et al., 2021; Radford et al., 2023). Building on these monolingual models, multilingual speech recognition systems (Fathullah et al., 2024; Rubenstein et al., 2023) have significantly improved performance, especially for low-resource languages, by leveraging the multilingual capabilities of LLMs. Building on the contextual understanding capabilities of LLMs, current Speech LLMs have significantly greatly improved the accuracy, robustness to noise, and adaptability to diverse accents and dialects. Beyond speech recognition, Speech LLMs have also expanded to support multitask learning, enabling a single model to perform a wide range of speech-related tasks (Chu et al., 2023; Tang et al., 2023). Qwen-Audio (Chu et al., 2023) and SALMONN (Tang et al., 2023) scale audio-language pre-training to cover various speech-related tasks and diverse audio inputs. Despite these significant progresses, the exploration of extending LLMs' capabilities to the audio-visual speech domain remains limited.

However, most current speech LLMs primarily focus on audio-based tasks and have not explored the effectiveness of adopting utilization of speech rate. In contrast, our work incorporates a speech rate predictor that dynamically allocates resources based on speaking speed of each audio, enabling more efficient and robust processing for audio-visual speech inputs.

3 Method

In this paper, our objective is to minimize the length of multimodal tokens while preserving their linguistic content, thereby enhancing the efficiency of our multimodal speech LLM framework. Recent works (Yeo et al., 2024a; Cappellazzo et al., 2024) have demonstrated that LLMs can serve as effective multimodal speech learners by leveraging their context modeling capabilities. Motivated by this finding, we adopt their framework as our baseline AVSR model.

As illustrated in Figure 1, the architecture consists of three main components: a visual encoder, an audio encoder, and an LLM decoder that predicts sentences from multimodal tokens. Building on these components, and with the goal of reducing the number of multimodal tokens while retaining essential linguistic information, we introduce three additional modules: the early av-fusion module, the AV Q-Former, and the speech rate predictor.

3.1 Early AV-Fusion Module

Although pre-fusion techniques that combine visual and text modalities have proven effective at reducing computational costs, their application in multimodal speech LLM frameworks remains unexplored. To extend their effectiveness to the multimodal speech domain, we propose an early av-fusion module that fuses visual and audio modalities before inputting them into the LLM, thereby halving the sequence length. To design this module effectively, we investigate three previously proposed fusion techniques for audio-visual speech: concatenation, addition, and multimodal attention.

Given the audio and video inputs, the visual encoder and audio encoder extract visual features $\mathbf{X}_v \in \mathbb{R}^{T_v \times D}$ and audio features $\mathbf{X}_a \in \mathbb{R}^{T_a \times D}$ that contain linguistic content from lip movements and sound, respectively. Since these features have different temporal resolutions (with audio features typically having a higher resolution than visual features), we employ a length adapter to resample the audio features so that they match the temporal scale of the visual features. We denote the resampled audio features as $\mathbf{X}'_a \in \mathbb{R}^{T_v \times D}$. Through this process, we align the audio-visual features along the time dimension and evaluate the effectiveness of three fusion methods based on these aligned features.

Concatenation. The audio and visual feature vectors are combined by simply appending one to the other along the feature dimension via concatenation approach. This can be expressed as follows:

$$\mathbf{X}_{av} = [\mathbf{X}'_a; \mathbf{X}_v] \in \mathbb{R}^{T_v \times 2D} \quad (1)$$

Addition. The addition method fuses audio and visual features by performing an element-wise sum.

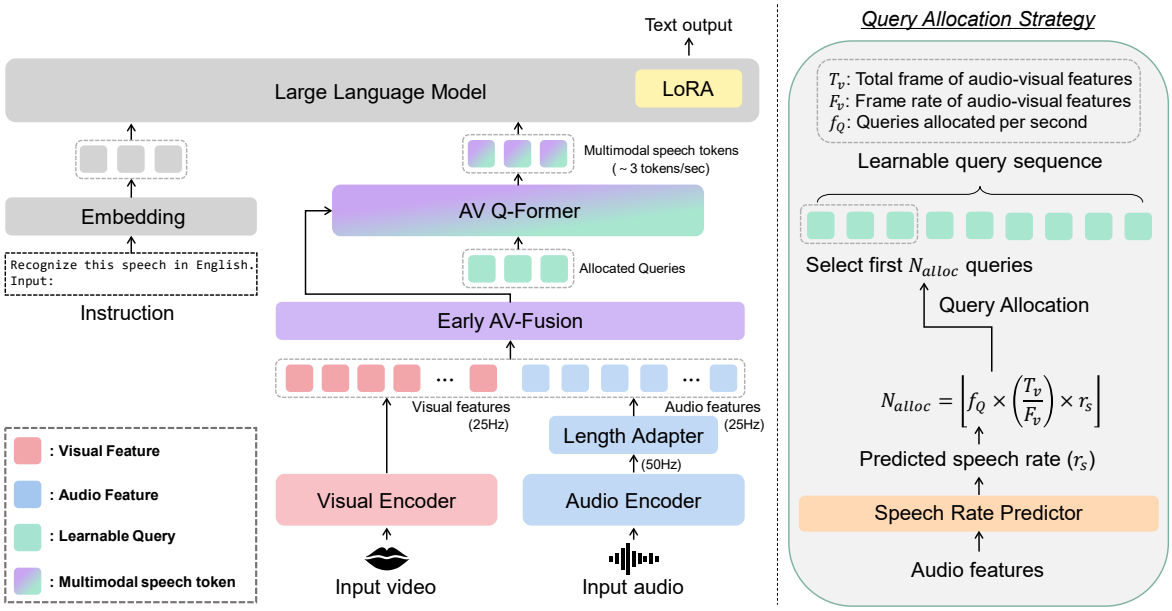

Figure 1: Illustration of the MMS-LLaMA framework. Audio and visual features are extracted by separate encoders, with a length adapter aligning audio frames. Early AV-Fusion merges the two modalities, while a speech rate predictor estimates speaking speed. Through the query allocation strategy, the appropriate number of queries is passed to the AV Q-Former, which produces multi-modal speech tokens. These tokens, combined with instruction embeddings, are then fed into the LLM to generate the output sentence. During training, the parameters of the encoders, the LLM's text embedding layer, and the speech rate predictor remain frozen.

It can be formulated as follows:

$$\mathbf{X}_{av} = \mathbf{X}'_a + \mathbf{X}_v \in \mathbb{R}^{T_v \times D} \quad (2)$$

where, $+$ indicate element-wise summation.

Multimodal Attention. The multimodal attention method fuses audio and visual features based on attention mechanism.

$$\mathbf{X}_{av} = \mathrm{MHCA}(\mathbf{X}_v W_Q, \mathbf{X}'_a W_K, \mathbf{X}'_a W_V) \quad (3)$$

where MHCA indicate multi-head cross attention, $\mathbf{X}_{av} \in \mathbb{R}^{T_v \times D}$, W_Q, W_K, and W_V are learnable projection matrices that transform the features into the query, key, and value.

3.2 AV Q-Former

While these early av-fusion modules reduce the length of the audio-visual feature sequence by half, there is still a gap compared to the number of text tokens. To bridge this gap between audio-visual speech and text modalities in terms of token count, we introduce a novel AV Q-Former.

To transform variable-length input sequences into fixed-length output queries, the Q-Former is introduced by Dai et al. (2023) in the vision-language domain. By employing a fixed-size window, Tang et al. (2023) demonstrates that the Q-Former effectively compresses audio-based speech tokens into text-level queries. Despite this progress, because the window-level Q-Former uses a fixed window size, it captures context information from only a portion of speech tokens in a single query. To address this limitation, we employ the conventional Q-Former with a novel query allocation strategy.

3.2.1 Query Allocation Strategy

While Q-Formers allocate a fixed number of queries regardless of input sequence length, the language content of audio-visual inputs is proportional to their duration. Therefore, our allocation strategy aims to dynamically adjust the number of queries based on the input length.

As shown in Figure 1, the AV Q-Former dynamically assigns a number of queries proportional to the length of the audio-visual feature sequence. To achieve this, we define a learnable query sequence $\mathbf{Q} \in \mathbb{R}^{N \times D_q}$, where N denotes the number of queries and D_q represents the embedding dimension of each query. Then, depending on the duration of the input audio and video, the number of queries is allocated proportionally to their respective durations. Let F_v denote the frame rate (in Hz) of the audio-visual feature sequence and assume a query rate f_Q (i.e., the number of queries per second), the allocated number of queries is given by

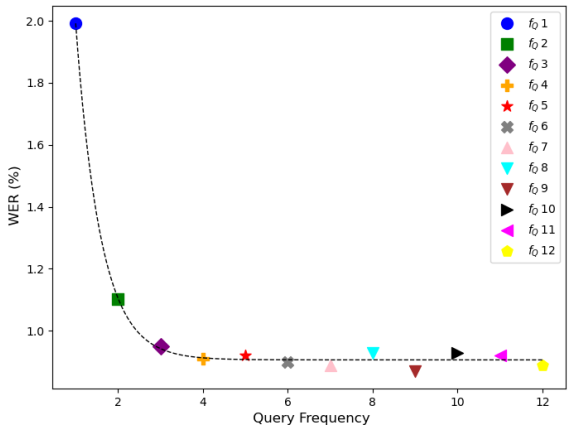

Figure 2: Impact of Query Frequency Variations on AVSR Performance. This results demonstrates that compressing the audio-visual feature sequence to as low as query frequency 4 maintains performance.

$$N_{alloc} = \lfloor f_Q \times \frac{T_v}{F_v} \rfloor \tag{4}$$

We subsequently select the first N_{alloc} queries from the learnable query sequence. It can be formulated as: $\mathbf{Q}_{alloc} = \mathbf{Q}[:N_{alloc}] \in \mathbb{R}^{N_{alloc} \times D_q}$. These queries with audio-visual feature sequence, are fed into the Q-Former to generate multimodal speech tokens $\mathbf{M} \in \mathbb{R}^{N_{alloc} \times D_q}$. This process can be expressed as follows:

$$\mathbf{M} = \text{Q-Former}(\mathbf{Q}_{\text{alloc}}; \mathbf{X}_{\text{av}}) \tag{5}$$

Then, we apply two linear layers to project the multimodal speech tokens into the LLM's embedding space. Next, we concatenate these projected tokens with the text instruction embeddings along the temporal axis, and then provide the resulting sequence as input to the LLM to predict the sentence. With this allocation strategy, we explore the minimum query frequency required to effectively compress the audio-visual feature sequence while maintaining performance.

3.3 Speech Rate Predictor

Through our AV Q-Former, as shown in Figure 2, we have confirmed that that performance remains robust even when compressing the audio-visual feature sequence by leveraging a query frequency of 4, which corresponds to 2.8 multimodal speech tokens per second (detailed in Section 5.2.1).

However, performance begins to degrade below a query frequency of 4 Hz. This is likely due to variations in the speech rate across audio samples, which rate usually measured in words per minute. Faster speech may contain more linguistic content and thus require additional queries, even if the total duration is the same as slower speech. To address this, we propose a speech rate predictor that allocates queries more effectively by considering each audio sample's speech rate. Our goal with this predictor is to optimize the query allocation strategy for enhanced efficiency.

To train the speech rate predictor, we first compute the average speech rate across the training set and normalize each sample's speech rate based on this reference. The normalized values serve as target labels, and we train the predictor using Mean Squared Error (MSE) loss with only audio features as input. This training process is performed before training our multimodal speech LLM framework. The pre-trained predictor then estimates the speech rate r_s and allocates more queries to higher speech rates (i.e., faster speech). This process can be formulated as follows:

$$N_{alloc} = \lfloor f_Q \times \frac{T_v}{F_v} \times r_s \rfloor. \tag{6}$$

The illustration of the speech rate predictor can be found in the right side of Figure 1.

4 Experimental Setup

4.1 Dataset

Lip Reading Sentences 3 (LRS3) as detailed in (Afouras et al., 2018b), is a widely used dataset designed for audio-visual speech recognition. It includes 433 hours of audio-visual data sourced from TED and TEDx talks, accompanied by human-annotated text transcriptions.

VoxCeleb2 as detailed in (Chung et al., 2018), is a dataset designed for speaker recognition. It consists of 2,442 hours of multilingual audio-visual data. Following (Shi et al., 2022a), we utilize only the English portion of this dataset, which amounts to 1,326 hours. Moreover, we also use the Whisper ASR model to generate pseudo text transcriptions, which we combine with the LRS3 dataset for training our model. This combined dataset amounts to 1,756 hours.

4.2 Implementation Details

4.2.1 Pre-processing

Following (Ma et al., 2023), we resample all audio and video from the LRS3 and VoxCeleb2 datasets to 25 fps and 16 kHz, respectively. Using RetinaFace (Deng et al., 2020), we crop the mouth region from the face video to a size of 96x96. The

Method	Audio Encoder	Visual Encoder	Decoder	Training Data(hrs)	WER ↓	
					Noisy	Clean
CM-seq2seq (Ma et al., 2021)	Conformer		Transformer	433	-	2.3
ViT3D-CM (Serdyuk et al., 2022)	Conformer		LSTM	90K	2.9	1.6
CMA (Kim et al., 2024c)	Transformer		Transformer	433	4.4	1.5
AV-data2vec (Lian et al., 2023)	Transformer		Transformer	433	6.7	2.5
auto-avsr (Ma et al., 2023)	Conformer		Transformer	1902/3448	-	1.0/0.9
LP Conformer (Chang et al., 2024)	Conformer		LSTM	100K	**1.9**	0.9
Whisper-Flamingo (Rouditchenko et al., 2024)	Whisper	AV-HuBERT	Whisper	433/1759	5.6/5.6	1.1/0.76
Multi-modal Speech LLM Framework						
LLaMA-AVSR (Cappellazzo et al., 2024)	Whisper	AV-HuBERT	LLaMA 3.1 8B	433	4.2	0.95
	Whisper	AV-HuBERT	LLaMA 3.1 8B	1759	-	0.77
MMS-LLaMA	Whisper	AV-HuBERT	LLaMA 3.2 3B	433	2.8	0.92
	Whisper	AV-HuBERT	LLaMA 3.2 3B	1759	**1.9**	**0.74**

Table 1: Comparisons with state-of-the-art methods on the LRS3 dataset. We report each method's architecture (audio encoder, visual encoder, decoder), the amount of training data, token throughput, and WER under both clean and noisy conditions. The clean condition is evaluated on the original test set, while the noisy condition is evaluated on a test set with babble noise added at 0-SNR

cropped mouth clips are then converted to grayscale and flipped horizontally for data augmentation during the training stage. Audio at a 16 kHz sampling rate is mixed with babble noise from the NOISEX dataset (Varga and Steeneken, 1993) at a 75% probability. After Whisper processing, the audio is padded to 30 seconds and converted into an 80-dimensional Mel spectrogram, which is then fed into the Whisper encoder

4.2.2 Architectures

We adopt AV-HuBERT (Shi et al., 2022a) as the visual encoder and Whisper (Radford et al., 2023) as the audio encoder. For the large language model (LLM), we use LLaMA variants: LLaMA 3.2 1B, 3B, and LLaMA 3.1 8B (Dubey et al., 2024). Our AV Q-Former is based on a BERT-large model with two Transformer layers, each having an embedding dimension of 1024, 16 attention heads, and a feed-forward dimension of 4096. Finally, the speech rate predictor consists of two Transformer layers with a 256-dimensional embedding, 4 attention heads, and a feed-forward dimension of 1024.

4.2.3 Training and evaluation

We use the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.98$, alongside a cosine learning rate scheduler. The initial learning rate is set to $1e^{-4}$, with 0.5k warm-up steps out of a total 30,000 steps. We also employ a minimum learning rate of $1e^{-5}$ and a final learning rate scale of 0.05. For fine-tuning the LLM, we adopt the QLoRA (Dettmers et al., 2024) approach with a LoRA rank of 16, an alpha (scaling factor) of 32, and a dropout rate of 0.05, applying LoRA to the query, key, value, and output projection layers. We evaluate performance using beam search decoding (beam size = 5) with a temperature of 0.3. All experiments are conducted on 8 RTX 3090 GPUs.

5 Experimental Results

5.1 Comparison with the state-of-the-art methods

In order to validate the effectiveness of the proposed method, we compare the proposed MMS-LLaMA with the previous state-of-the-art AVSR methods on LRS3 dataset, as shown in Table 1.

Traditional approaches, such as CM-seq2seq (Ma et al., 2021), ViT3D-CM (Serdyuk et al., 2022), and auto-avsr (Ma et al., 2023), commonly employ Conformer- or Transformer-based backbones and rely on large-scale datasets to improve AVSR performances. Notably, LP Conformer (Chang et al., 2024) achieved a 0.9% WER using 100K hours of training data. More recent models, including Whisper-Flamingo (Rouditchenko et al., 2024) and LLaMA-AVSR (Cappellazzo et al., 2024), leverage large-scale pretrained models (e.g., Whisper and LLMs) and have demonstrated superior performance with WERs of 0.77% and 0.76%,

Method	# MMS Tokens per second	GPU Memory Usage(GB)	FLOPs (T)	WER(%)
Baseline* (Cappellazzo et al., 2024)	25	18.2	2.24	0.97
+ Early AV Fusion	12.5	14.7	1.81	0.92
+ AV Q-Former	0.826	11.1	1.35	1.99
	1.793	12.1	1.39	1.10
	2.796	12.2	1.42	0.95
+ Speech Rate Predictor	0.832	11.2	1.35	1.61
	1.794	12.2	1.39	0.97
	2.797	12.2	1.42	0.92

Table 2: Effectiveness of each proposed component on LRS3. We report token throughput, gpu memory usage, FLOPs, and WER. MultiModal Speech (MMS) tokens indicates the number of tokens derived from one second of audio-visual input that are fed into the LLM. Note that in the rows corresponding to AV Q-Former and Speech Rate Predictor, query frequencies of 1, 2, and 3 are applied from top to bottom. *We re-implemented it to have the same LLM parameters with ours.

respectively. Our MMS-LLaMA surpasses previous state-of-the-art methods, achieving a WER of 0.74% when trained on 1,759 hours of data. When trained on only 433 hours of data, MMS-LLaMA obtains a WER of 0.92%, outperforming both Whisper-Flamingo and LLaMA-AVSR. Under noisy conditions, our model also achieves state-of-the-art performance, with a WER of 1.9%. Notably, our model significantly improves performance in noisy conditions, addressing a major weakness of previous multimodal speech LLM framework that showed a substantial gap between noisy and clean conditions. Please note that the proposed MMS-LLaMA not only achieves the superior performances but also effectively reduces the number of tokens with the proposed AV Q-Former, which will be evaluated in the following subsections.

5.2 Ablation study

5.2.1 Validation of the Effectiveness of Each Component via Sequential Integration

To verify the effectiveness of the proposed components in terms of computation cost and WER, we have conducted 8 experiments through sequential integration. These all models except for baseline, are applied concatenation early av-fusion, trained using 433 hours training data. Moreover, for fair comparison, we re-implemented baseline model* (Cappellazzo et al., 2024) using Llama 3.2 3B instead of LLaMA 3.1 8B, to mitigate the effects of varying LLM parameter sizes. Table 2 presents a detailed performance comparisons on the LRS3 dataset.

Types of AV-Fusion	FLOPs(T)	WER ↓	
		Noisy	Clean
Concatenation	1.50	2.77	0.92
Addition	1.49	3.02	0.97
Multimodal Attention	1.50	3.03	0.87

Table 3: Comparison of different audio-visual fusion strategies in terms of computational cost (FLOPs), and WER under noisy and clean conditions.

The baseline uses 25 multimodal speech tokens and achieves a WER of 0.97% with 2.24T FLOPs and 18.2GB gpu memory usage. With the addition of the early av fusion component, token throughput is reduced to 12.5 multimodal speech tokens per second, FLOPs decreases to 1.81T, while the WER improves slightly to 0.92%.

Next, we add the AV Q-Former component. Its impact is evaluated under three different query frequency configurations of 1, 2, and 3. Please note that the results using query frequencies of 1, 2, and 3 are shown from top to bottom in the table, respectively. The query frequency of 1 yields the lowest FLOPs of 1.35T and gpu memory usage of 11.1 GB, but the WER increases significantly to 1.99%. Increasing the query frequency to 2 and 3 leads to a modest rise in FLOPs of 1.39T and 1.42T, and gpu memory usage of 12.1GB and 12.2GB, while the WER improves to 1.10% and 0.95%, respectively. This demonstrates a trade-off where higher query frequency counts can help recover recognition accuracy at the cost of slightly higher computational demands.

Similarly, the speech rate predictor is added and evaluated with the same query frequency configurations. In the first row, FLOPs are maintained at 1.35T, and gpu memory usage slightly incrased 0.1 GB, but the WER is reduced to 1.61%. For the 2 and 3 query frequency settings, FLOPs are remained as 1.39T and 1.42T, and gpu memory usage are almost identical, while the WER are improved to 0.97% and 0.92%, respectively. One can find that the number of tokens per second is only slightly improved when the speech rate predictor is employed, while bringing huge WER improvements, indicating that the test set's average speech rate is close to 1, but individual samples exhibit varying speech rates.

Overall, the sequential integration of the proposed components demonstrates that both early av

Types of LLM	GPU Memory Usage(GB)	Flops(T)	WER ↓	
			Noisy	Clean
LLaMA3.2-1B	9.8	1.19	3.11	1.11
LLaMA3.2-3B	12.3	1.50	2.77	0.92
LLaMA3.1-8B	16.7	2.17	2.61	1.02

Table 4: Comparison of gpu memory usage, FLOPs and WER based on the model size of LLMs.

fusion and the additional modules (AV Q-Former and speech rate predictor) can effectively reduce computational costs and, under appropriate configurations, maintain or improve recognition accuracy.

5.2.2 Evaluation of Different Audio-Visual Fusion Strategies

To reduce the computational cost in LLMs, we have introduced an early av-fusion module that shifts the fusion process to an earlier stage. To determine the most effective approach, we conduct ablation study by using three different av-fusion techniques: Concatenation, Addition, and Multimodal Attention.

The results, presented in Table 3, indicate that all three fusion techniques exhibit similar FLOPs. Under noisy conditions, concatenation achieves the best WER of 2.77%, while multimodal attention performs best on clean speech with 0.87% WER, followed by concatenation of 0.92% WER. Because the performance gap in noisy settings is more significant than in clean conditions, we adopt concatenation as our primary fusion strategy in the other experiments.

5.2.3 Impact of LLM Model Size

To investigate how the model size of LLMs affects AVSR performance, computational cost, and gpu memory usage, we conducted experiments using three models based on LLaMA with 1B, 3B, and 8B parameters. We trained each model on 433 hours of data, applied early audio visual fusion through concatenation, and incorporated an AV Q Former with a query frequency of 5, without including the speech rate predictor.

As shown in Table 4, the LLaMA3.2 3B model achieved the best WER of 0.92% under clean conditions, outperforming both the LLaMA3.2 1B and LLaMA3.1 8B models. However, under noisy conditions, the LLaMA3.1 8B model achieved the best WER of 2.61%. These results suggest that larger model might perform better under the challenging environment. As expected, the larger model incurs higher GPU memory usage and computational costs. Given its comparable performance to the 8B model, we report results using the LLaMA3.2-3B model for other experiments due to its significantly lower computational requirements.

Query Frequency	Visual Modality	SNR Level (dB), WER(↓)					
		∞	5	2	0	-2	-5
3	-	1.10	1.58	2.66	4.17	6.30	13.54
1	✓	1.99	2.36	3.15	4.30	6.62	12.25
2	✓	1.10	1.54	2.13	3.34	4.72	9.31
3	✓	0.95	1.30	1.83	2.66	4.16	7.44
4	✓	0.91	1.32	1.84	2.72	3.95	7.42
5	✓	0.92	1.22	1.74	2.91	4.26	7.28

Table 5: WER results at various SNR levels (∞, 5, 2, 0, -2, -5 dB), where ∞ indicates clean audio, comparing different query frequencies with and without the visual modality.

5.2.4 Evaluation of Performance Across Various SNR Levels with Different Query Frequencies

In this section, we aim to validate the effectiveness of the visual modality and evaluate AVSR performances across various query frequencies at different SNR levels. Table 5 presents the WER results spanning from clean settings (∞ dB) to severely noisy conditions (-5 dB).

The results indicate that incorporating the visual modality leads to a significant improvement in performance, especially in noisy environments. For example, while a query frequency of 3 without the visual modality obtain a WER of 1.10% in clean setting and 13.54% at -5 dB, adding the visual modality with the same query frequency reduces the WERs to 0.91% and 7.44%, respectively. Moreover, as the query frequency increases from 1 to 5, there is a general tendency for performance to improve across all SNR levels, aligning with the trends observed in Figure 2.

6 Conclusion

We have demonstrated that integrating efficient token compression strategies into multimodal speech LLM frameworks can dramatically reduce computational costs while preserving high-level language content. By employing an early AV-fusion module, a dynamically adaptive AV Q-Former, and a refined query allocation strategy with a speech rate predictor, we achieve state-of-the-art AVSR performance, attaining a WER of 0.74% while using only 2.8 multimodal speech tokens per second. Moreover, our extensive experiments on the LRS3 dataset confirm that our method not only achieves a remarkable WER but also reduces FLOPs by 36% compared to the previous LLM-based AVSR method.

7 Limitation

While we have introduced an efficient multimodal speech LLM framework, namely MMS-LLaMA, its current focus is constrained to the AVSR task. This narrow scope may limit the immediate applicability of our framework to other domains, such as multimodal speech dialogue systems. Nonetheless, the proposed method efficiently processes audio-visual inputs by leveraging the proposed tokens reducing scheme. Building on this efficient framework, we expect that MMS-LLaMA can be extended to real-world communication scenarios by training the system on large-scale multimodal dialogue corpora.

References

Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. 2018a. Deep audio-visual speech recognition. *IEEE transactions on pattern analysis and machine intelligence*, 44(12):8717–8727.

Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. 2018b. Lrs3-ted: a large-scale dataset for visual speech recognition. *arXiv preprint arXiv:1809.00496*.

Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. 2016. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning*, pages 173–182. PMLR.

Umberto Cappellazzo, Minsu Kim, Honglie Chen, Pingchuan Ma, Stavros Petridis, Daniele Falavigna, Alessio Brutti, and Maja Pantic. 2024. Large language models are strong audio-visual speech recognition learners. *arXiv preprint arXiv:2409.12319*.

Oscar Chang, Hank Liao, Dmitriy Serdyuk, Ankit Shahy, and Olivier Siohan. 2024. Conformer is all you need for visual speech recognition. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10136–10140. IEEE.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*.

Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. 2018. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*.

Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. 2017. Lip reading sentences in the wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3444–3453. IEEE Computer Society.

Wenliang Dai, Junnan Li, D Li, AMH Tiong, J Zhao, W Wang, B Li, P Fung, and S Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. arxiv 2023. *arXiv preprint arXiv:2305.06500*, 2.

Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. 2020. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5203–5212.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Jeffrey L Elman. 1990. Finding structure in time. *Cognitive science*, 14(2):179–211.

Yassir Fathullah, Chunyang Wu, Egor Lakomkin, Junteng Jia, Yuan Shangguan, Ke Li, Jinxi Guo, Wenhan Xiong, Jay Mahadeokar, Ozlem Kalinli, et al. 2024. Prompting large language models with speech recognition abilities. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 13351–13355. IEEE.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. 2020. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*.

Alexandros Haliassos, Rodrigo Mira, Honglie Chen, Zoe Landgraf, Stavros Petridis, and Maja Pantic. 2024a. Unified speech recognition: A single model for auditory, visual, and audiovisual inputs. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Alexandros Haliassos, Andreas Zinonos, Rodrigo Mira, Stavros Petridis, and Maja Pantic. 2024b. Braven: Improving self-supervised pre-training for visual and auditory speech recognition. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11431–11435. IEEE.

Joanna Hong, Minsu Kim, Jeongsoo Choi, and Yong Man Ro. 2023. Watch or listen: Robust audio-visual speech recognition with visual corruption modeling and reliability scoring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18783–18794.

Joanna Hong, Minsu Kim, Daehun Yoo, and Yong Man Ro. 2022. Visual context-driven audio feature enhancement for robust end-to-end audio-visual speech recognition. *arXiv preprint arXiv:2207.06020*.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460.

Jing Huang and Brian Kingsbury. 2013. Audio-visual deep learning for noise robust speech recognition. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 7596–7599. IEEE.

Minsu Kim, Hyung-Il Kim, and Yong Man Ro. 2024a. Prompt tuning of deep neural networks for speaker-adaptive visual speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Minsu Kim, Jeong Hun Yeo, Jeongsoo Choi, and Yong Man Ro. 2023. Lip reading for low-resource languages by learning and combining general speech knowledge and language-specific knowledge. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15359–15371.

Minsu Kim, Jeong Hun Yeo, and Yong Man Ro. 2022. Distinguishing homophenes using multi-head visual-audio memory for lip reading. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 1174–1182.

Minsu Kim, Jeonghun Yeo, Se Jin Park, Hyeongseop Rha, and Yong Man Ro. 2024b. Efficient training for multilingual visual speech recognition: Pre-training with discretized visual speech representation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 1311–1320.

Sungnyun Kim, Kangwook Jang, Sangmin Bae, Hoirin Kim, and Se-Young Yun. 2024c. Learning video temporal dynamics with cross-modal attention for robust audio-visual speech recognition. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 447–454. IEEE.

Suyoun Kim, Takaaki Hori, and Shinji Watanabe. 2017. Joint ctc-attention based end-to-end speech recognition using multi-task learning. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4835–4839. IEEE.

Jiachen Lian, Alexei Baevski, Wei-Ning Hsu, and Michael Auli. 2023. Av-data2vec: Self-supervised learning of audio-visual speech representations with contextualized target representations. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE.

Pingchuan Ma, Alexandros Haliassos, Adriana Fernandez-Lopez, Honglie Chen, Stavros Petridis, and Maja Pantic. 2023. Auto-avsr: Audio-visual speech recognition with automatic labels. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Pingchuan Ma, Stavros Petridis, and Maja Pantic. 2021. End-to-end audio-visual speech recognition with conformers. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7613–7617. IEEE.

Pingchuan Ma, Stavros Petridis, and Maja Pantic. 2022. Visual speech recognition for multiple languages in the wild. *Nature Machine Intelligence*, 4(11):930–939.

Brais Martinez, Pingchuan Ma, Stavros Petridis, and Maja Pantic. 2020. Lipreading using temporal convolutional networks. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6319–6323. IEEE.

Youssef Mroueh, Etienne Marcheret, and Vaibhava Goel. 2015. Deep multimodal learning for audio-visual speech recognition. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2130–2134. IEEE.

Kuniaki Noda, Yuki Yamaguchi, Kazuhiro Nakadai, Hiroshi G Okuno, and Tetsuya Ogata. 2015. Audio-visual speech recognition using deep learning. *Applied intelligence*, 42:722–737.

Stavros Petridis, Zuwei Li, and Maja Pantic. 2017. End-to-end visual speech recognition with lstms. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 2592–2596. IEEE.

Stavros Petridis, Themos Stafylakis, Pingchuan Ma, Georgios Tzimiropoulos, and Maja Pantic. 2018a. Audio-visual speech recognition with a hybrid ctc/attention architecture. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 513–520. IEEE.

Stavros Petridis, Themos Stafylakis, Pingehuan Ma, Feipeng Cai, Georgios Tzimiropoulos, and Maja Pantic. 2018b. End-to-end audiovisual speech recognition. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 6548–6552. IEEE.

Rohit Prabhavalkar, Takaaki Hori, Tara N Sainath, Ralf Schlüter, and Shinji Watanabe. 2023. End-to-end speech recognition: A survey. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Andrew Rouditchenko, Yuan Gong, Samuel Thomas, Leonid Karlinsky, Hilde Kuehne, Rogerio Feris, and James Glass. 2024. Whisper-flamingo: Integrating visual features into whisper for audio-visual speech recognition and translation. *arXiv preprint arXiv:2406.10082*.

Paul K Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, et al. 2023. Audiopalm: A large language model that can speak and listen. *arXiv preprint arXiv:2306.12925*.

Dmitriy Serdyuk, Otavio Braga, and Olivier Siohan. 2022. Transformer-based video front-ends for audio-visual speech recognition for single and multi-person video. *arXiv preprint arXiv:2201.10439*.

Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. 2022a. Learning audio-visual speech representation by masked multimodal cluster prediction. *arXiv preprint arXiv:2201.02184*.

Bowen Shi, Wei-Ning Hsu, and Abdelrahman Mohamed. 2022b. Robust self-supervised audio-visual speech recognition. *arXiv preprint arXiv:2201.01763*.

Darryl Stewart, Rowan Seymour, Adrian Pass, and Ji Ming. 2013. Robust audio-visual speech recognition under noisy audio-video conditions. *IEEE transactions on cybernetics*, 44(2):175–184.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2023. Salmonn: Towards generic hearing abilities for large language models. *arXiv preprint arXiv:2310.13289*.

Andrew Varga and Herman JM Steeneken. 1993. Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems. *Speech communication*, 12(3):247–251.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Jeong Hun Yeo, Seunghee Han, Minsu Kim, and Yong Man Ro. 2024a. Where visual speech meets language: Vsp-llm framework for efficient and context-aware visual speech processing. *arXiv preprint arXiv:2402.15151*.

Jeong Hun Yeo, Chae Won Kim, Hyunjun Kim, Hyeongseop Rha, Seunghee Han, Wen-Huang Cheng, and Yong Man Ro. 2024b. Personalized lip reading: Adapting to your unique lip movements with vision and language. *arXiv preprint arXiv:2409.00986*.

Jeong Hun Yeo, Minsu Kim, Jeongsoo Choi, Dae Hoe Kim, and Yong Man Ro. 2024c. Akvsr: Audio knowledge empowered visual speech recognition by compressing audio knowledge of a pretrained model. *IEEE Transactions on Multimedia*.

Jeong Hun Yeo, Minsu Kim, Shinji Watanabe, and Yong Man Ro. 2024d. Visual speech recognition for languages with limited labeled data using automatic labels from whisper. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10471–10475. IEEE.

Wenyi Yu, Changli Tang, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2024. Connecting speech encoder and large language model for asr. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12637–12641. IEEE.

Shaolei Zhang, Qingkai Fang, Zhe Yang, and Yang Feng. 2025. Llava-mini: Efficient image and video large multimodal models with one vision token. *arXiv preprint arXiv:2501.03895*.