Bridging the Temporal Gap in Multimodal LLMs: Deeply Stacking Temporal Tokens for Audio-Visual Speech Recognition

Liyong Wang[1], **Junliang Xing**[2†], **Tianyu Hu**[1], **Jianfei Jiang**[1], **Jihuai Zhao**[1], **Huimin Ma**[1†]
[1]University of Science and Technology Beijing [2]Tsinghua University
{wangly, jiangjf, jihuaizhao}@xs.ustb.edu.cn jlxing@tsinghua.edu.cn
{Tianyu, mhmpub}@ustb.edu.cn

Abstract

Audio-Visual Speech Recognition enhances speech recognition robustness in noisy conditions by leveraging visual cues. However, current Multimodal LLMs suffer from a fundamental temporal gap. This gap is characterized by limited fine-grained temporal modeling in vision encoders and progressive temporal semantic degradation throughout the deep layers of LLM decoders. To bridge this gap, we propose a novel framework that deeply stacks temporal tokens across both the encoding and decoding stages. Specifically, we enhance the vision encoder with a temporal-aware attention module and temporal rotary positional embeddings to precisely capture the sequential evolution and lip movement dynamics. Furthermore, we stack hierarchical temporal tokens that incorporate temporally enriched features into multiple layers of the LLM decoder in a bottom-up manner. Extensive experiments on the LRS2 and LRS3 benchmarks demonstrate that our approach achieves quite satisfactory results, outperforming existing supervised, self-supervised, and LLM-based methods by 6.1% on LRS2 and 7.8% on LRS3. Our codes are available at https://github.com/LyongW/Temporal-AVSR.

1 Introduction

Automatic speech recognition systems (Park et al., 2019; Gulati et al., 2020) have achieved remarkable progress in recent years. However, their performance remains highly vulnerable in noisy acoustic environments (Radford et al., 2022). Humans naturally leverage visual cues, particularly lip movements, to enhance speech perception in noisy conditions. This observation has motivated the development of Audio-Visual Speech Recognition (AVSR) systems (Shi et al., 2022; Rekesh et al., 2023; Yeo et al., 2025b) that integrate both auditory and visual modalities to achieve robust recognition under adverse acoustic conditions.

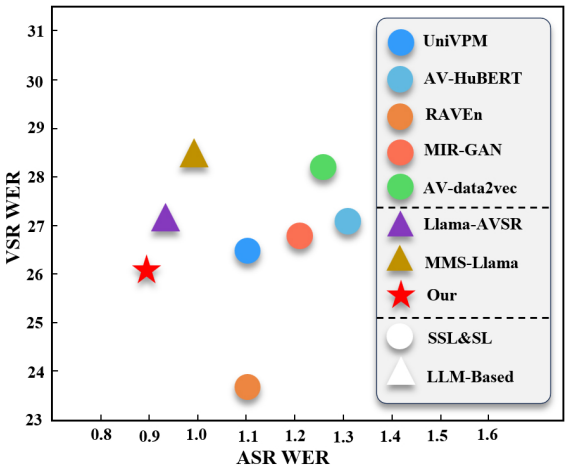


Figure 1: Comparison of ASR and VSR performance between traditional AVSR models based on self-supervised (SSL) or supervised learning (SL) (○ markers) and LLM-based AVSR frameworks (△ markers).

Visual speech (Liu et al., 2023) patterns captured from lip movements provide complementary information that remains inherently robust to acoustic noise. A fundamental characteristic of visual speech is its rich, highly structured temporal dependency. Unlike static visual concepts, lip movements convey linguistic information through continuous motion over multiple video frames, where a sequence of subtle articulatory changes often expresses a single word or phoneme . Accurately recognizing speech from lip movements, therefore, requires modeling fine-grained temporal dynamics across consecutive frames.

Recent advances in Multimodal LLMs (Cappellazzo et al., 2024; Yeo et al., 2025b) leverage the impressive generalization and instruction-following capabilities of LLMs. Most existing LLM-based approaches (Ye et al., 2024; Cappellazzo et al., 2025b; Yeo et al., 2024a) process non-

†Corresponding Author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 27748–27759

text modalities by incorporating large numbers of visual tokens and audio tokens as prefix inputs alongside language prompts. Compared to traditional self-supervised (Haliassos et al., 2022) or supervised learning (Ma et al., 2023) AVSR models, LLM-based methods capitalize on powerful contextual reasoning to achieve lower Word Error Rates (WER) on Automatic Speech Recognition (ASR) (Chen et al., 2024). Yet, they underperform on Visual Speech Recognition (VSR), as illustrated in Fig. 1. These observations reveal a fundamental temporal gap in multimodal LLM-based AVSR frameworks. Current LLM-based approaches (Cappellazzo et al., 2024, 2025a) lack fine-grained modeling of action sequences and temporal progression, leading to degraded performance in VSR.

Furthermore, even when temporal features are extracted at the visual encoder, a second challenge emerges: the progressive degradation of temporal semantics within deep LLM layers. Standard LLM-based methods (Yeo et al., 2025b; Cappellazzo et al., 2024, 2025c) typically incorporate visual and audio tokens only at the input embedding layer, treating them as a prefix to the text sequence. As shown in Fig. 2, when feeding the LLM with standard and shuffled video token sequences, we observe a progressive increase in similarity across decoder layers. This trend indicates that fine-grained temporal semantics are gradually lost as visual features propagate through deep autoregressive decoding layers, with the loss being more apparent in the top decoder layers. The fine-grained temporal information tends to be gradually diluted or overridden by the inherently strong linguistic priors of the pre-trained language model.

In this paper, we propose a novel LLM-based AVSR method that bridges the temporal gap by deeply stacking temporal tokens across both encoding and decoding stages. At the vision encoder level, we enhance temporal modeling to capture the fine-grained dynamics of visual speech, producing temporally coherent representations for downstream decoding. At the LLM decoding level, we stack hierarchical temporal tokens that facilitate the propagation of temporal features across decoder layers to mitigate temporal semantic degradation. Bridging the encoder and decoder, we employ a Q-Former to compress representations into learnable tokens, reducing the LLM's computational overhead while preserving the temporal structure.

The main contributions of this paper are summarized as follows:

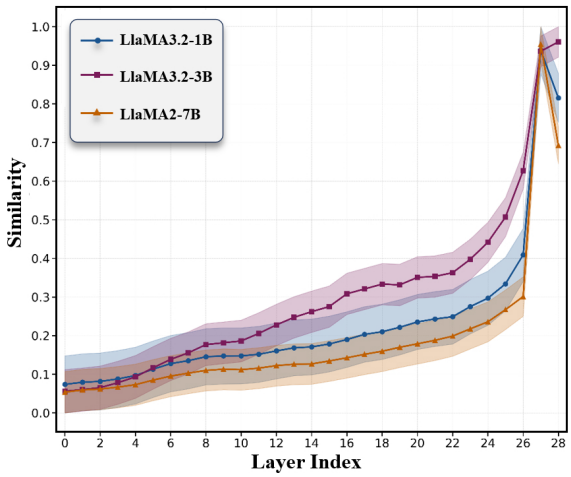


Figure 2: Similarity across different layers between normal and shuffled video tokens orders in LLM decoder.

- We analyze a temporal gap in multimodal LLM-based AVSR frameworks, revealing limited fine-grained temporal modeling and progressive temporal degradation.
- We bridge this temporal gap by deeply stacking temporal tokens at both the vision encoder and LLM decoder levels.
- Following a broad analysis of existing SSL/SL and LLM-based methods for ASR, VSR, and AVSR tasks, our method achieves superior performance on LRS2 and LRS3 while maintaining high computational efficiency.

2 Related Work

2.1 Supervised & Self-Supervised AVSR

AVSR has been widely studied to improve speech recognition robustness in noisy environments. Early supervised learning (Afouras et al., 2018a; Chung et al., 2016) utilized end-to-end architectures (Petridis et al., 2018) to jointly model multimodal streams, demonstrating the value of lip movements in noise-robust recognition. Subsequent works introduced temporal convolutional (Ma et al., 2021), and later transformer-based (Serdyuk et al., 2022; Hong et al., 2023) and Conformer-based (Rekesh et al., 2023; Chang et al., 2024) models to better capture long-range temporal dependencies across modalities.

Recently, SSL has emerged as a dominant paradigm in AVSR (Haliassos et al., 2022; Hsu and Shi, 2022), learning aligned representations from large-scale unlabeled data. While achieving strong performance (Haliassos et al., 2024b; Lian et al.,

2023), these frameworks demand substantial data and computational resources. Consequently, some work has pivoted toward sample-efficient strategies, including low-resource (Kim et al., 2023), synthetic-data (Liu et al., 2023), and knowledge-transfer (Yang et al., 2024) approaches. Parallel efforts address robustness via cross-lingual (Choi et al., 2023), cross-speaker (Yeo et al., 2024b), and zero-shot (Yeo et al., 2025a) settings. Despite their proficiency in VSR, these traditional architectures lack the flexibility and rich contextual reasoning of LLMs, often yielding suboptimal ASR results, thereby motivating a shift toward LLM-based frameworks.

2.2 LLM-based Audio-Visual Speech Recognition

Recent advances in LLMs (Ye Bai and Chen, 2024) have revolutionized speech recognition by offering robust linguistic priors and long-context modeling. Early attempts to incorporate LLMs into ASR primarily treated them as auxiliary language models for post-hoc error correction (Chen et al., 2024), demonstrating their effectiveness in improving linguistic consistency. Building on these successes, some work explores (Cappellazzo et al., 2024) integrating LLMs into AVSR to jointly exploit multimodal speech cues and strong language modeling. Recent work (Yeo et al., 2025b) applies LLMs to language-agnostic and zero-shot (Yeo et al., 2025a) AVSR by mapping audio–visual representations to text, and further incorporates visual speech cues into LLM-based frameworks via lightweight projection modules for end-to-end multimodal decoding. To reduce the cost of long audio–visual sequences, some methods (Cappellazzo et al., 2025a) adopt structured token compression to preserve linguistically salient information.

However, despite superior noise robustness compared to SSL methods, LLM-based approaches significantly underperform in VSR. This performance disparity reveals a fundamental temporal gap: by treating dynamic audio-visual tokens as static prefixes, existing methods neglect the sequential evolution of lip movements, presenting a critical challenge for the field. Therefore, closing this temporal gap is an urgent prerequisite for further advances in LLM-based AVSR.

3 Method

In this section, we propose a novel audio–visual speech recognition framework that explicitly addresses fine-grained temporal modeling in multimodal LLMs. As shown in Fig. 3, our approach enhances temporal understanding at two complementary levels: (1) temporal-aware attention in the vision encoder, and (2) stacking hierarchical temporal tokens in the LLM decode.

3.1 Overview Model Architecture

Multimodal Encoders: Given a video, we first decompose it into individual video frames $V \in \mathbb{R}^{B\times T\times H\times W}$ and audio segments $A \in \mathbb{R}^{B\times K\times M}$, where T represents the frame number, and K represents the segment number. For video frames, an AV-HuBERT (Shi et al., 2022) is employed to obtain frame-level embeddings $F_v \in \mathbb{R}^{B\times T\times D}$. For audio segments, we utilize Whisper (Radford et al., 2022) to extract the last hidden states as the audio embedding $F_a \in \mathbb{R}^{B\times K\times D_a}$, where D_a represents the embedding dimension for each audio segment: $F_a = \{a^1, a^2, a^3, ..., a^K\}$.

After obtaining the synchronised visual feature $F_v \in \mathbb{R}^{B\times T\times D}$ and audio feature $F_a \in \mathbb{R}^{B\times K\times D_a}$, the two modalities are aligned and concatenated along the feature dimension to form a unified multimodal sequence $F_{av} = Concat(F_v, F_a)$, where F_{av} denotes the fused audio–visual representation. Note that when only one input modality is present, the other modality is filled with a sequence of zero padding of the same sequence length. The fused multimodal sequence F_{av} is fed into the proposed TAA module, as described in detail in Section 3.2.

Q-Former: To compress the fused multimodal sequence while preserving salient audio–visual information, we adopt a causal Q-Former equipped with the same set of N trainable input query tokens $\mathbf{Q} = \{\mathbf{q}_1, \ldots, \mathbf{q}_N\}$.

$$\boldsymbol{F}_N^Q = \text{QFormer}\left(\boldsymbol{F}_{t:t+T}^{AV}; \mathbf{Q}\right), \tag{1}$$

the resulting query representations $\boldsymbol{F}_N^Q$ are then projected to the LLM input dimension before being fed into the language model.

Multimodal Large Language Models: LLMs are typically pre-trained on a vast amount of unlabeled text corpus using a transformer decoder-only architecture. MLLMs extend pre-trained LLMs by conditioning text generation on input images and audio via visual tokens as prefix inputs. Formally, the learning objective can be formulated as:

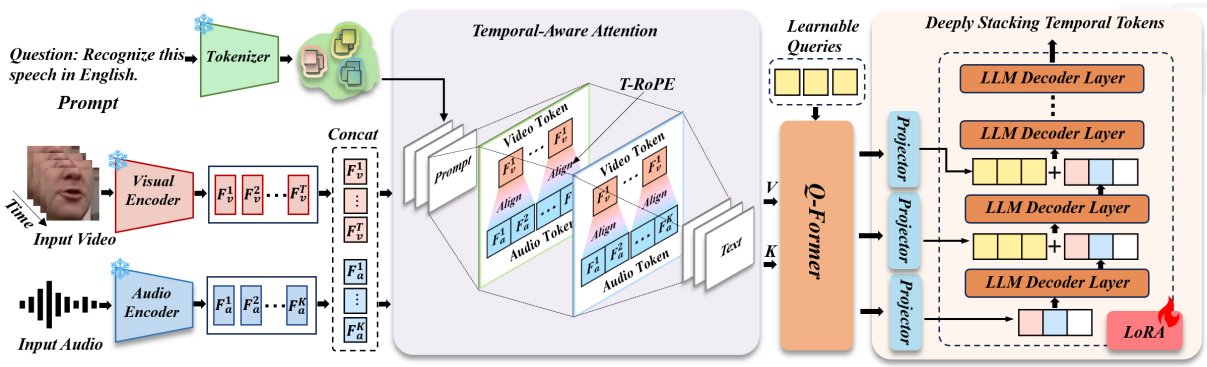


Figure 3: Overview of the proposed AVSR framework. Visual and audio features are first encoded and temporally aligned, and then enhanced by the temporal-aware attention module with T-RoPE. The resulting representations are compressed by a Q-Former and hierarchically integrated into multiple LLM decoder layers via temporal token stacking, enabling robust fine-grained temporal modeling for AVSR.

$$\mathcal{L} = \sum_{t=1}^{N} \log \mathcal{P}_\theta(x_{t+1} \mid x_{1:t}, \boldsymbol{F}_N^Q), \tag{2}$$

where $\mathcal{P}$ represents the LLM and θ is the trainable parameters of the model. $\boldsymbol{F}_N^Q \in R^{N \times C}$ represents the sequence of visual and audio tokens.

3.2 Temporal-Aware Attention

In this section, we explicitly address the challenge of temporal understanding by deeply stacking temporal tokens across both encoding and decoding stages. At the vision encoder level, we introduce a Temporal-Aware Attention (TAA) module enhanced with Temporal Rotary Positional Embeddings (T-RoPE) to better capture inter-frame dependencies and the sequential evolution of visemes. By enriching the visual tokens with temporal cues, we provide a compact yet informative temporal representation for the subsequent decoding process. At the LLM decoder level, we stack hierarchical temporal tokens that incorporate temporally enriched visual representations into multiple decoder layers in a bottom-up manner via lightweight projectors, thereby mitigating progressive temporal semantic degradation during autoregressive decoding.

3.2.1 Temporal Rotary Position Embeddings

To explicitly model the temporal correspondence between visual and acoustic modalities, which often exhibit different temporal resolutions, we introduce a Temporal Rotary Positional Embedding (T-RoPE) strategy based on RoFormer (Su et al., 2021). This approach ensures that visual and audio tokens corresponding to the same timestamp share identical temporal position IDs, allowing the attention mechanism to naturally associate temporally

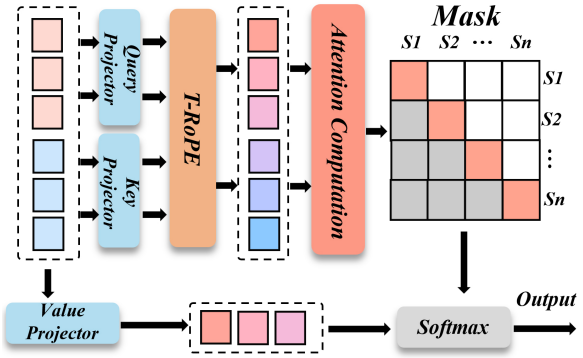


Figure 4: Temporal-Aware Attention with T-RoPE. T-RoPE and causal masking enable fine-grained temporal attention over audio–visual tokens.

aligned cross-modal information. Simultaneously, the temporal order is maintained across different frames, with position IDs incrementing accordingly. For visual feature $F_v \in \mathbb{R}^{B \times T \times D}$ and audio feature $F_a \in \mathbb{R}^{B \times K \times D_a}$. The proposed temporal position id is defined as follows:

$$\mathbf{I}_t(n) = \begin{cases} n, & \text{if } n < v_s^a, \\ v_s^a + T\left\lfloor \dfrac{n - v_s^a}{K} \right\rfloor, & \text{if } v_s^a \le n \le v_e^a, \\ n - \Big(v_e^a - v_s^a + 1 \\ \qquad - T\left\lfloor \dfrac{n - v_s^a}{K} \right\rfloor\Big), & \text{if } n > v_e^a, \end{cases} \tag{3}$$

where v_s^a and v_e^a are the starting and ending position ids of the audio tokens within the global RoPE position id n. $\lfloor . \rfloor$ denotes the floor function, which rounds down to the nearest integer. By scaling the position ids, temporal information is introduced through the adjusted position $\hat{n}$, defined as:

$$\hat{n} = n + \mathbf{I}_t(n). \tag{4}$$

This adjustment ensures that temporal information is effectively incorporated into the original position embedding. When computing the attention map, the T-RoPE technique introduces the multiplication of Euler's formula $e^{i\theta}$ to the query and key vectors as a relative position embedding.

3.2.2 Temporal Attention

As shown in Fig. 4, we process the audio-visual features for each sentence i across all frames, defined as $F_i^Q = [F_{1,i}^Q, \ldots, F_{N,i}^Q]^\top \in \mathbb{R}^{N\times D}$. The temporal attention mechanism computes:

$$\begin{cases} Q_i = \mathrm{LN}\left(F_i^Q\right) W_Q, \\ K_i = \mathrm{LN}\left(F_i^Q\right) W_K, \\ V_i = \mathrm{LN}\left(F_i^Q\right) W_V, \end{cases} \tag{5}$$

where $W_Q, W_K, W_V \in \mathbb{R}^{D\times d_t}$ are learnable projection matrices, and d_t is the dimension of queries and keys.

We adopt a T-RoPE and apply it to the multimodal domain for both video and audio tokens. This design explicitly encodes relative temporal relationships within the audiovisual modality while facilitating effective cross-modal interactions. For instance, when considering the n-th and m-th query and key vectors Q_i^n and K_i^m in $\mathbb{R}^{1\times d_{head}}$, T-RoPE is applied as follows:

$$Q_i'^n = Q_n e^{i\hat{n}\theta} = Q_n e^{i(n+\gamma\cdot\mathbf{I}_t(n))\theta}, \tag{6}$$

$$K_i'^m = K_m e^{i\hat{m}\theta} = K_m e^{i(m+\gamma\cdot\mathbf{I}_t(m))\theta}. \tag{7}$$

These transformed representations are used to compute the attention matrix, thereby capturing inter-frame temporal dynamics via relative position modulation. The module output is obtained as:

$$Z_i = F_i^Q + \mathrm{softmax}(Q_i' K_i') V_i. \tag{8}$$

The feature representation Z_i from the last transformer block is projected into the LLM's embedding space, concatenated with the prompt's language embeddings, and fed to the LLM. Our key innovation lies in the parameter-efficient temporal-aware attention. By integrating temporal rotary positional embeddings within the temporal attention block, we explicitly encode the relative temporal structure of the multimodal sequence. This lightweight design significantly enhances the modeling of fine-grained temporal dynamics with minimal computational overhead, thereby improving the quality of audio-visual features for temporally challenging AVSR tasks.

3.3 Hierarchical Temporal Token Stacking

As the autoregressive decoding process progresses, fine-grained temporal cues introduced at the input stage often degrade gradually in multimodal settings, where visual and audio signals are usually overshadowed by the strong linguistic priors inherent in deep LLM layers. To leverage the hierarchical nature of Transformer decoders, which encode low-level patterns in early layers and abstract semantics in deeper layers, we propose reinforcing temporal modeling through explicit hierarchical token stacking. Specifically, we employ multiple projection layers to map temporally enriched tokens into the LLM's embedding space, stacking them across specific decoder layers to mitigate semantic degradation.

Given an LLM decoder with L layers, we denote the l-th decoder layer as $\mathcal{D}_l$, and its output hidden states as $\mathcal{H}_l = \mathcal{D}_l(\mathcal{H}_{l-1})$. We partition it bottom-up into an early block $\mathcal{D}^{bottom}$ for stacking temporal tokens and a late block $\mathcal{D}^{up}$ for standard prefix-based sequential modeling. The early decoder block is defined as $\mathcal{D}^{\mathrm{bottom}} = \{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_{L_b}\}$ where L_b denotes the number of layers in the early block. We project temporally enriched features into each early decoder layer in a hierarchical manner via layer-wise projection modules.

$$\mathcal{H}_l^{stack} = \mathrm{Projector}(\mathcal{Z}), \tag{9}$$

where $\mathcal{H}_l^{\mathrm{stack}} = \{\mathcal{H}_1^{\mathrm{stack}}, \mathcal{H}_2^{\mathrm{stack}}, \ldots, \mathcal{H}_{L_b}^{\mathrm{stack}}\}$ represents temporally projected tokens aligned with the l-th early decoder layer. The original hidden states $\mathcal{H}_{l-1}$ produced by the decoder layer $\mathcal{D}_{l-1}$ are combined with the stacked temporal representations $\mathcal{H}_l^{\mathrm{stack}}$ via a residual connection, which is formulated as:

$$\mathcal{H}_l = \mathcal{D}_l(\mathcal{H}_{l-1}) + \mathcal{H}_l^{\mathbf{stack}}. \tag{10}$$

Specifically, temporal tokens are stacked only within the early decoder block $\mathcal{D}^{bottom}$, while the first decoder layer operates without temporal token stacking. Hierarchical temporal token stacking preserves and propagates fine-grained temporal information across early decoder layers, mitigating

Method	Backbone	Trainable (Par. M)	Labeled Hours	WER (%) ↓		
				VSR	ASR	AVSR
Supervised Learning						
Auto-AVSR(Ma et al., 2023)	Conformer	425	818	33.0	1.5	1.0
Hyb-Conformer (Ma et al., 2021)	Conformer	412	433	27.9	2.3	2.3
MIR-GAN (Hu et al., 2023a)	Transformer	–	433	26.6	1.3	1.2
Fast Conformer(Rekesh et al., 2023)	Conformer	197	433	43.8	1.7	1.6
AVEC (Burchi and Timofte, 2023)	Conformer	61	818	37.5	2.0	1.8
ViT 3D(Serdyuk et al., 2022)	Transformer	310	YouTube-90k	17.0	–	1.6
LP Conformer(Chang et al., 2024)	Conformer	570	YouTube-100k	12.8	–	0.9
Self-Supervised Learning						
AV-data2vec(Lian et al., 2023)	Transformer	325	1,759	28.5	1.2	1.3
AV-HuBERT(Shi et al., 2022)	Transformer	325	1,759	26.9	–	–
RAVEn(Haliassos et al., 2022)	Transformer	328	1,759	23.1	1.4	–
USR(Haliassos et al., 2024a)	Transformer	503	1,759	22.3	1.2	1.1
AV-HuBERT(Shi et al., 2022)	Transformer	325	433	28.6	1.5	1.4
VATLM(shi Zhu et al., 2022)	Transformer	332	433	28.4	–	1.2
RAVEn(Haliassos et al., 2022)	Transformer	328	433	28.2	1.4	–
BRAVEn(Haliassos et al., 2024b)	Transformer	328	433	26.6	1.2	–
u-HuBERT(Hsu and Shi, 2022)	Transformer	325	433	27.2	1.5	1.3
UniVPM (Hu et al., 2023b)	Transformer	478	433	26.7	1.42	1.18
LLM-Based						
Llama-AVSR(Cappellazzo et al., 2024)	Whisper + AV-HuBERT V	57	224	37.20	2.40	2.21
Llama-AVSR(Cappellazzo et al., 2024)	Whisper + AV-HuBERT V	57	433	27.10	1.10	0.95
Llama-AVSR(Cappellazzo et al., 2024)	Whisper + AV-HuBERT V	57	1,759	25.90	0.97	0.85
+ Ours	Whisper + AV-HuBERT V	65.3	433	26.10	1.01	0.91
MMS-Llama*(Yeo et al., 2025b)	Whisper + AV-HuBERT V	39.6	224	38.91	3.50	3.11
MMS-Llama*(Yeo et al., 2025b)	Whisper + AV-HuBERT V	39.6	433	29.83	1.50	1.14
MMS-Llama*(Yeo et al., 2025b)	Whisper + AV-HuBERT V	39.6	1,759	28.50	1.20	0.91
MoME*(Cappellazzo et al., 2025a)	Whisper + AV-HuBERT V	12.7	433	29.0	1.70	1.50
Ours*	Whisper + AV-HuBERT V	48	224	36.50	2.50	2.84
Ours*	Whisper + AV-HuBERT V	48	433	27.50	1.32	1.02
Ours*	Whisper + AV-HuBERT V	48	1,759	26.30	1.10	0.89

Table 1: Comparison of WER (%) under VSR, ASR, and AVSR settings in supervised learning, self-supervised learning, and LLM-based models. Our method is highlighted in green. * denotes token compression methods.

the dominance of linguistic priors during autoregressive decoding. In contrast, the late decoder block $\mathcal{D}^{\text{up}}$ operates without temporal token stacking and focuses on high-level semantic reasoning and language generation based on the temporally enriched representations from the early block.

4 Experiments

4.1 Datasets

LRS2: (Chung et al., 2016) consists of 224 hours of video collected from BBC television programs. It contains approximately 144k utterances, with 1.2k reserved for testing. **LRS3** (Afouras et al., 2018b) comprises 439 hours of TED and TEDx talks with varied visual conditions. The dataset includes 120k pre-training and 32k training utterances, with a test set of 1,321 samples.

VoxCeleb2: (Chung et al., 2018) is employed to scale up our training. We specifically utilize its English subset, which provides approximately 1,326 hours of audio-visual data from over 6,000 speakers. To enable joint training with labeled datasets, we use pseudo-text transcriptions generated by a pre-trained ASR system.

4.2 Implementation Details

Pre-processing: Following (Ma et al., 2023), audio and video are resampled to 16 kHz and 25 fps. For the visual stream, RetinaFace (Deng et al., 2019) detects and crops 96×96 mouth ROIs, followed by random augmentation and normalization. To improve robustness, babble noise (Varga and Steeneken, 1993) is added during training.

Training and evaluation: We implement our framework on the Llama3.2-3B backbone, fine-tuned with LoRA (Hu et al., 2021) (rank $r = 16$,

TAA	T-RoPE	HSTT	LRS2 WER ↓	LRS3 WER ↓
✗	✗	✗	38.91	29.83
✓	✗	✗	37.13	28.23
✓	✓	✗	36.94	28.10
✓	✓	✓	36.57	27.50

Table 2: Effect of different components (TAA, T-RoPE, and HSTT) on WER performance for LRS2 and LRS3 datasets.

Blocks	Para. (M)	FLOPs (G)	LRS2 ↓	LRS3 ↓
0	-	-	38.91	29.83
2	8.3	7.56	36.57	27.53
4	16.6	15.12	36.85	27.72
6	24.9	22.68	36.33	27.56
8	33.2	30.24	36.12	27.21

Table 3: Effect of different numbers of temporal attention blocks on WER for LRS2 and LRS3 datasets.

$\alpha = 32$, dropout 0.05). A Q-Former (Li et al., 2023) with token compression processes both visual and audio inputs. We train using Adam (Kingma and Ba, 2014) ($\beta_1 = 0.9, \beta_2 = 0.98$) with a cosine schedule on RTX 4090 GPUs with gradient accumulation, and decode with beam search (beam size 5, temperature 0.37).

4.3 Experimental Results

4.3.1 Comparison Across Different Supervision Methods

Table 1 compares our method with state-of-the-art supervised, self-supervised, and LLM-based baselines in terms of WER on VSR, ASR, and AVSR. Our approach surpasses conventional supervised models on all three tasks while using substantially fewer trainable parameters. Although ViT 3D (Serdyuk et al., 2022) and LP Conformer (Chang et al., 2024) further improve VSR by relying on larger models and more labeled data, our framework achieves superior ASR and AVSR performance with only 1,759 labeled training samples. Compared with self-supervised learning methods, our model demonstrates superior performance. Using 433 hours of labeled data, our method achieves 27.5% VSR WER, outperforming AV-HuBERT (28.6%) (Ma et al., 2023), VATLM (28.4%) (Burchi and Timofte, 2023), and RAVEn (28.2%) (Haliassos et al., 2024b) while using significantly fewer trainable parameters. For AVSR tasks, our model achieves 1.02% WER with 433 hours of training data, outperforming all self-supervised models.

We reproduce Llama-AVSR (Cappellazzo et al., 2024) and MMS-Llama (Yeo et al., 2025b) and compare them with our framework. For fairness, we group LLM-based methods into token-compression models (MMS-Llama, MoME (Cappellazzo et al., 2025a)) and non-compression models (Llama-AVSR). Our approach surpasses MMS-Llama and MoME, achieving WERs of 26.3%, 1.1%, and 0.89% on VSR, ASR, and AVSR, respectively. In VSR, it significantly outperforms MMS-Llama, with gains of 6.1% on LRS2 and 7.8% on LRS3. Compared with Llama-AVSR without token compression, our model attains comparable WER while compressing both visual and audio inputs by 6×.

4.3.2 Ablation Study

In this ablation, we evaluate the contributions of Temporal-Aware Attention (TAA), Temporal Rotary Positional Embeddings (T-RoPE), and Hierarchical Temporal Token Stacking (HSTT) on LRS2 and LRS3. Adding TAA reduces WER from 38.91% to 37.13% on LRS2 and from 29.83% to 28.23% on LRS3, indicating improved temporal dependency modeling. Adding T-RoPE on top of TAA further lowers WER to 36.94% and 28.10% on LRS2 and LRS3, respectively, by enhancing temporal position encoding. The full model with TAA, T-RoPE, and HSTT achieves the best results, with WERs of 36.57% on LRS2 and 27.50% on LRS3, highlighting the significance of hierarchical temporal token stacking.

4.3.3 Different Effects of Temporal Attention Blocks

This experiment evaluates the impact of varying the number of temporal attention blocks on LRS2 and LRS3, as shown in Table 3. Without TAA, the model attains baseline WERs of 38.91% (LRS2) and 29.83% (LRS3), while adding just 2 TAA blocks reduces WER to 36.57% and 27.53%, corresponding to 6.0% and 7.7% relative improvements. These gains, achieved with only 8.3M additional parameters and 7.56 GFLOPs, demonstrate the parameter efficiency of our TAA module.

4.3.4 Token Compression and Robustness Analysis

Table 4 summarizes AVSR performance under different audio-visual token rates and amounts of labeled data. Compared with non-compression baselines (CM-seq2seq (Ma et al., 2021), Eff. Conf. (Burchi and Timofte, 2023), USR (Haliassos et al., 2024a), auto-avsr (Ma et al., 2023)), which rely on more labeled data and parameters, our method achieves comparable WER on LRS2 and LRS3.

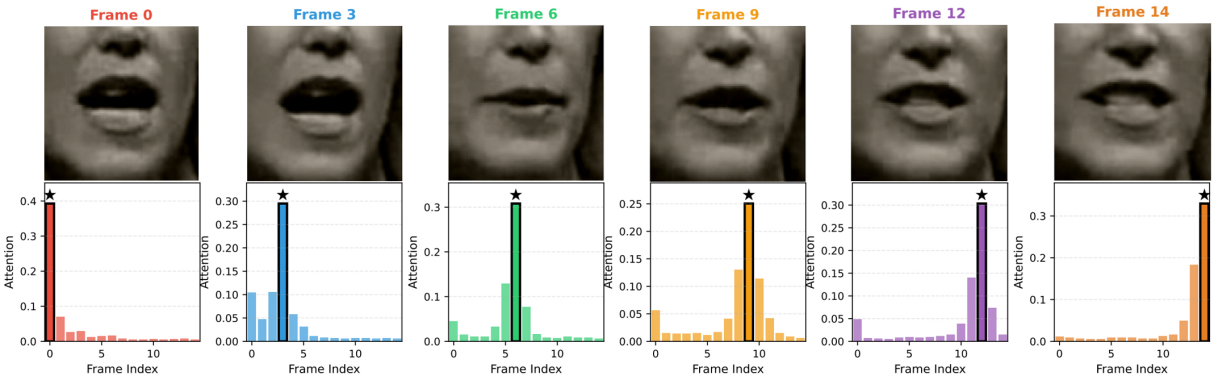


Figure 5: Visualization of temporal attention in a continuous video sequence. ★ marks the current frame.

Method	Rates	Lab.	Dataset	
	(A, V)	Hrs.	LRS2 ↓	LRS3 ↓
CM-seq2seq	(1, 1)	380/433	3.7	2.3
Eff. Conf.	(1, 1)	818/818	2.3	1.8
USR	(1, 1)	1982/1759	1.9	1.1
auto-avsr	(1, 1)	3448/1902	1.5	1.0
MMS-Llama	(6, 6)	223/433	3.11	1.14
MoME	(4, 2)	223/433	2.70	1.50
Our	(6, 6)	223/433	**2.50**	**1.02**

Table 4: Comparison of AVSR methods on LRS2 and LRS3 datasets under different audio-visual token rates and labeled hours.

Method	SNR (dB)				
	12.5	7.5	2.5	-2.5	-7.5
Auto-AVSR	1.0	1.0	1.5	2.2	5.6
Ours	0.89	0.97	1.4	2.5	6.3

Table 5: WER (%) comparison of Auto-AVSR and our method under different SNR levels.

Among token-compression approaches, our Q-Former with 6× compression for both audio and visual tokens attains the best results, with WERs of 2.50% on LRS2 and 1.02% on LRS3.

Table 5 reports the WER of Auto-AVSR and our method under SNRs from 12.5 to −7.5 dB. At moderate and high SNRs, our framework consistently yields lower WER, demonstrating improved robustness under mild noise. As SNR drops to −2.5 and −7.5 dB, both systems degrade, but our method remains competitive. Overall, these results indicate that the temporal-aware framework is robust to noise and generally superior to the baseline in realistic acoustic conditions.

4.4 Temporal Attention Visualization

To better understand how the proposed TAA operates over time, we visualize the temporal attention rollout on a continuous video sequence, as shown in Fig. 5. We select 15 consecutive frames and sample 6 key frames, for which the middle row plots the attention distribution over all frames for each queried frame. The visualization clearly reveals a localized and temporally coherent attention pattern. For each queried frame, TAA concentrates on the current frame and its immediate neighbors, with attention weights rapidly decaying for distant frames. As the utterance naturally progresses, the attention peak smoothly shifts along the temporal axis, indicating that the model dynamically tracks evolving lip motion. Our method thus focuses on a compact temporal neighborhood to capture fine-grained visual dynamics while producing temporally consistent representations for AVSR.

5 Conclusion

In this paper, we identify and analyze a critical temporal gap in existing multimodal LLM-based AVSR frameworks, where fine-grained temporal dynamics of lip movements are inadequately captured and progressively lost across deep decoder layers. To address this, we propose a novel temporal-aware attention and T-ROPE to capture fine-grained visual dynamics. Furthermore, we introduce hierarchical temporal token stacking to preserve temporal information across decoder layers effectively. Extensive experiments on the LRS2 and LRS3 benchmarks demonstrate that our method achieves superior performance, significantly outperforming existing supervised, self-supervised, and LLM-based baselines.

Acknowledgments

This work was supported by the Beijing Natural Science Foundation (No. L257003), the National Natural Science Foundation of China (No.U25B2073).

Limitations

Despite the demonstrated efficiency and robustness of our method, several limitations remain. First, our evaluations are currently restricted to English audio-visual benchmarks (LRS2 and LRS3). Consequently, the generalization capability of the proposed temporal-aware attention and T-RoPE mechanisms to languages with distinct phoneme-viseme mappings or complex tonal characteristics requires further verification. Second, while our model excels in offline, sentence-level decoding, practical deployment often demands real-time capabilities. Adapting the hierarchical temporal stacking mechanism to a continuous streaming setup, where strict latency constraints are critical, presents a significant challenge for future optimization.

References

Triantafyllos Afouras, Joon Son Chung, Andrew W. Senior, Oriol Vinyals, and Andrew Zisserman. 2018a. Deep audio-visual speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44:8717–8727.

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

Maxime Burchi and Radu Timofte. 2023. Audio-visual efficient conformer for robust speech recognition. *IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2257–2266.

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. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–5.

Umberto Cappellazzo, Minsu Kim, Pingchuan Ma, Honglie Chen, Xubo Liu, Stavros Petridis, and Maja Pantic. 2025a. MoME: Mixture of matryoshka experts for audio-visual speech recognition. *ArXiv*.

Umberto Cappellazzo, Minsu Kim, and Stavros Petridis. 2025b. Adaptive audio-visual speech recognition via matryoshka-based multimodal llms. *ArXiv*, abs/2503.06362.

Umberto Cappellazzo, Minsu Kim, Stavros Petridis, Daniele Falavigna, and Alessio Brutti. 2025c. Scaling and enhancing llm-based avsr: A sparse mixture of projectors approach. *ArXiv*, abs/2505.14336.

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

Chen Chen, Ruizhe Li, Yuchen Hu, Sabato Marco Siniscalchi, Pin-Yu Chen, Ensiong Chng, and Chao-Han Huck Yang. 2024. It's never too late: Fusing acoustic information into large language models for automatic speech recognition. *ArXiv*, abs/2402.05457.

Jeong Yun Choi, Se Jin Park, Minsu Kim, and Yong Man Ro. 2023. Av2av: Direct audio-visual speech to audio-visual speech translation with unified audio-visual speech representation. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27315–27327.

Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. 2018. Voxceleb2: Deep speaker recognition. In *Interspeech*.

Joon Son Chung, Andrew W. Senior, Oriol Vinyals, and Andrew Zisserman. 2016. Lip reading sentences in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3444–3453.

Jiankang Deng, J. Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou. 2019. Retinaface: Single-stage dense face localisation in the wild. *ArXiv*, abs/1905.00641.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented transformer for speech recognition. *ArXiv*, abs/2005.08100.

Alexandros Haliassos, Pingchuan Ma, Rodrigo Mira, Stavros Petridis, and Maja Pantic. 2022. Jointly learning visual and auditory speech representations from raw data. *ArXiv*, abs/2212.06246.

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. *ArXiv*, abs/2411.02256.

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 *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 11431–11435.

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

Wei-Ning Hsu and Bowen Shi. 2022. u-HuBERT: Unified mixed-modal speech pretraining and zero-shot transfer to unlabeled modality. In *Neural Information Processing Systems*.

J. Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *ArXiv*, abs/2106.09685.

Yuchen Hu, Chen Chen, Ruizhe Li, Heqing Zou, and Chng Eng Siong. 2023a. MIR-GAN: Refining frame-level modality-invariant representations with adversarial network for audio-visual speech recognition. *ArXiv*, abs/2306.10567.

Yuchen Hu, Ruizhe Li, Cheng Chen, Chengwei Qin, Qiu shi Zhu, and Eng Siong Chng. 2023b. Hearing lips in noise: Universal viseme-phoneme mapping and transfer for robust audio-visual speech recognition. In *Annual Meeting of the Association for Computational Linguistics*.

Minsu Kim, Jeong Hun Yeo, Jeong Yun Choi, and Yong Man Ro. 2023. Lip reading for low-resource languages by learning and combining general speech knowledge and language-specific knowledge. In *IEEE/CVF International Conference on Computer Vision*, pages 15313–15325.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*.

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. *IEEE Automatic Speech Recognition and Understanding Workshop*, pages 1–8.

Xubo Liu, Egor Lakomkin, Konstantinos Vougioukas, Pingchuan Ma, Honglie Chen, Rui-Cang Xie, Morrie Doulaty, Niko Moritz, Jáchym Kolár, Stavros Petridis, Maja Pantic, and Christian Fuegen. 2023. Synthvsr: Scaling up visual speech recognitionwith synthetic supervision. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18806–18815.

Pingchuan Ma, Alexandros Haliassos, Adriana Fernandez-Lopez, Honglie Chen, Stavros Petridis, and Maja Pantic. 2023. Auto-avsr: Audio-visual speech recognition with automatic labels. pages 1–5.

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

Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin Dogus Cubuk, and Quoc V. Le. 2019. Specaugment: A simple data augmentation method for automatic speech recognition. In *Interspeech*.

Stavros Petridis, Themos Stafylakis, Pingchuan Ma, Feipeng Cai, Georgios Tzimiropoulos, and Maja Pantic. 2018. End-to-end audiovisual speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6548–6552.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*.

Dima Rekesh, Samuel Kriman, Somshubra Majumdar, Vahid Noroozi, He Juang, Oleksii Hrinchuk, Ankur Kumar, and Boris Ginsburg. 2023. Fast conformer with linearly scalable attention for efficient speech recognition. *IEEE Automatic Speech Recognition and Understanding Workshop*, pages 1–8.

Dmitriy Serdyuk, Otavio Braga, and Olivier Siohan. 2022. Transformer-based video front-ends for audio-visual speech recognition. In *Interspeech*.

Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdel rahman Mohamed. 2022. Learning audio-visual speech representation by masked multimodal cluster prediction. *ArXiv*, abs/2201.02184.

Qiu shi Zhu, Long Zhou, Zi-Hua Zhang, Shujie Liu, Binxing Jiao, J. Zhang, Lirong Dai, Daxin Jiang, Jinyu Li, and Furu Wei. 2022. Vatlm: Visual-audio-text pre-training with unified masked prediction for speech representation learning. *IEEE Transactions on Multimedia*, 26:1055–1064.

Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. 2021. Roformer: Enhanced transformer with rotary position embedding. *ArXiv*, abs/2104.09864.

Andrew Varga and Herman J. M. 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 Commun.*, 12:247–251.

Xiaoda Yang, Xize Cheng, Jiaqi Duan, Hongshun Qiu, Minjie Hong, Minghui Fang, Shengpeng Ji, Jialong Zuo, Zhiqing Hong, Zhimeng Zhang, and Tao Jin. 2024. Audiovsr: Enhancing video speech recognition with audio data. In *Conference on Empirical Methods in Natural Language Processing*.

Qilang Ye, Zitong Yu, Rui Shao, Xinyu Xie, Philip H. S. Torr, and Xiaochun Cao. 2024. Cat: Enhancing multimodal large language model to answer questions in dynamic audio-visual scenarios. *ArXiv*, abs/2403.04640.

Jitong Chen Ye Bai, Jingping Chen and Wei Chen. 2024. Seed-asr: Understanding diverse speech and contexts with llm-based speech recognition. *ArXiv*, abs/2407.04675.

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. In *Conference on Empirical Methods in Natural Language Processing*.

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. In *AAAI Conference on Artificial Intelligence*.

Jeong Hun Yeo, Minsu Kim, Chae Won Kim, Stavros Petridis, and Yong Man Ro. 2025a. Zero-avsr: Zero-shot audio-visual speech recognition with llms by learning language-agnostic speech representations. *ArXiv*, abs/2503.06273.

Jeong Hun Yeo, Hyeongseop Rha, Se Jin Park, and Yong Man Ro. 2025b. Mms-llama: Efficient llm-based audio-visual speech recognition with minimal multimodal speech tokens. *ArXiv*.

A Supplementary materials

A.1 Qualitative Analysis of Error Cases

Error Case : Substitution

Prompt: Recognize this video in English.

Reference - now my own speciality is combining digital technology and magic

Prediction - now my own specialization is combining digital technology and magic

---

Reference - what difference does it make if they talk like jerry seinfeld

Prediction - what difference does it make if they talk like george seinfeld

---

Reference - one thousand eight hundred and thirty six people died

Prediction - 1836 people died

Error Case : Omission

Prompt: Recognize this video in English.

Reference - it's shocking to realize that only 28 percent of american

Prediction - it's shocking to realize that only 28 of american

---

Reference - we do train them however to adjust caution according to where they are

Prediction - we do train them however to adjust according to where they are

Error Case : Insertion

Prompt: Recognize this video in English.

Reference - for africans homegrown science fiction can be a will to power

Prediction - for africans home grown science fiction can be a way to power

---

Reference - of course migration will become even more important

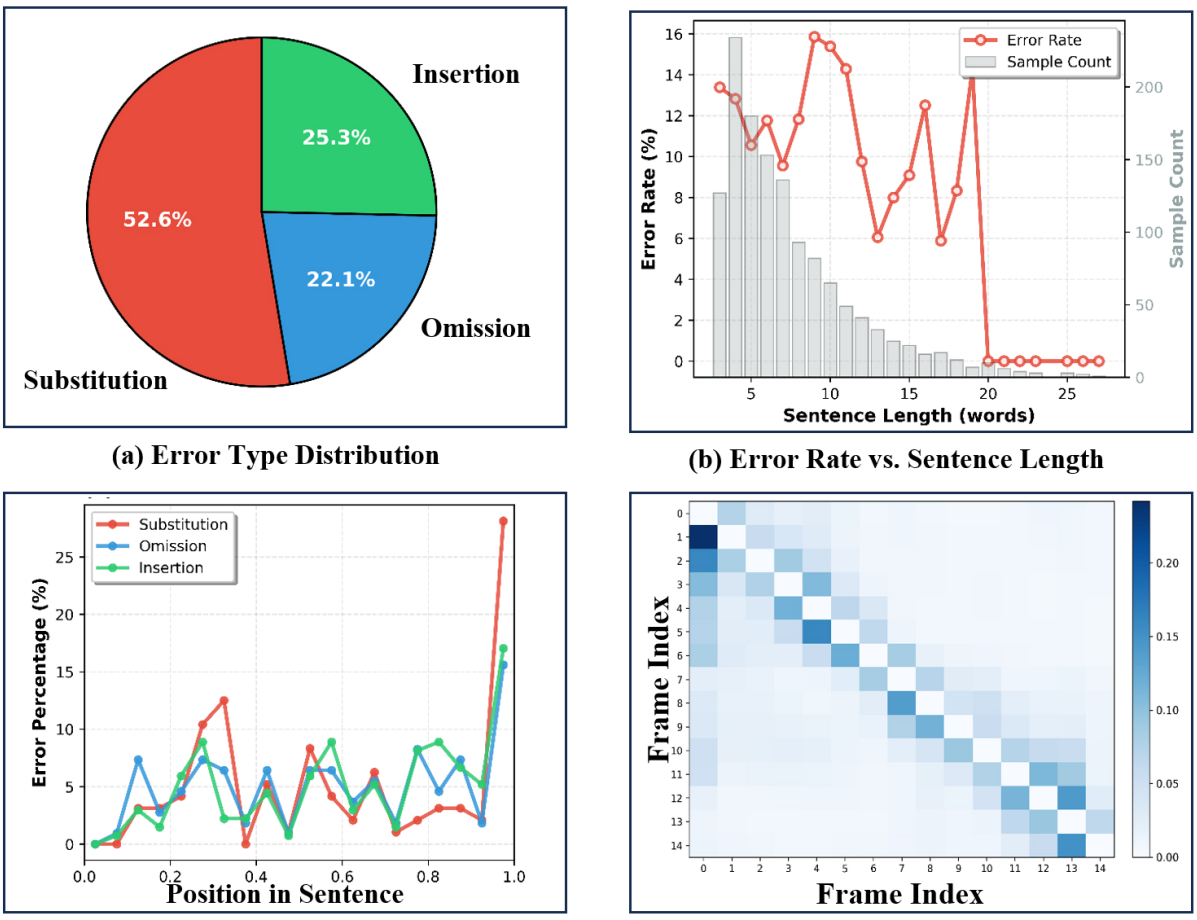
Prediction - and of course migration would become even more important

The qualitative error cases in the supplementary material, drawn from the VSR task on the LRS3 dataset, show that our model mainly fails at fine-grained lexical realization rather than global semantics. We categorize errors into three types: substitutions, omissions, and insertions. The qualitative analysis of error cases reveals that the model's deviations are often driven by its strong semantic reasoning and linguistic priors rather than simple recognition failures.

Substitution errors demonstrate the model's tendency towards semantic equivalence ("speciality" $\rightarrow$ "specialization"). However, visual ambiguity in proper nouns can occasionally lead to context-driven hallucinations (confusing "Jerry" $\rightarrow$ "George"). Omissions are primarily observed in words with weak visual cues or reduced lip motion amplitude during rapid speech, while insertions reflect the generative nature of the underlying LLM, which may introduce grammatically plausible connectives (adding "and") or alter tokenization ("home grown") to enhance syntactic fluency. Collectively, these patterns suggest the model functions as a semantic interpreter, prioritizing coherent understanding over rigid verbatim alignment.

A.2 Statistical Analysis of Recognition Errors

To investigate the performance bottlenecks of the VSR model on the LRS3 dataset, we conducted a detailed statistical analysis of recognition errors, as illustrated in Fig. 6. First, the error type distribution in Fig. 6 (a) indicates that substitution errors are dominant. Unlike audio-based recognition, substitution errors in VSR primarily stem from the visual ambiguity of homophenes, where multiple phonemes map to identical lip movements (e.g., /p/ and /b/). This high proportion suggests that



(a) Error Type Distribution

(b) Error Rate vs. Sentence Length

(c) Error Rate vs. Position in Sentence

(d) Temporal Attention Heatmap

Figure 6: Analysis of recognition errors and temporal attention patterns of the proposed AVSR model.

while the model successfully detects articulatory actions, it faces challenges in fine-grained viseme discrimination.

The relationship between error rate and sentence length in Fig. 6 (b) highlights the critical role of context in VSR. For short sentences (<10 words), the lack of sufficient semantic constraints makes visual ambiguities hard to rectify, resulting in higher and more volatile error rates. Conversely, as sentence length increases (>20 words), the error rate drops significantly. This demonstrates the model's ability to leverage long-range contextual dependencies to resolve local lip-reading ambiguities, yielding greater robustness in long-sentence recognition.

Furthermore, Fig. 6 (c) illustrates the error distribution relative to sentence position. We observe a sharp increase in error rates at the end of sentences (position > 0.9). This "boundary effect" is common in VSR, likely due to reduced lip motion amplitude or premature return to a neutral expression as the speaker finishes, which weakens visual features and complicates the prediction of the end of sentence token.

Finally, from a dynamic tracking perspective, the temporal attention heatmap in Fig. 6 (d) exhibits a clear diagonal alignment, demonstrating the model's ability to consistently follow the temporal evolution of the input sequence. Observed high attention scores along the diagonal reveal that the model focuses intensely on the current frame and its surrounding window, indicating an effective utilization of short-term temporal dependencies for feature extraction. Conversely, the regions off the main diagonal (representing distant frames) are predominantly suppressed with near-zero weights, indicating that the model has effectively learned to filter out irrelevant long-range temporal noise.