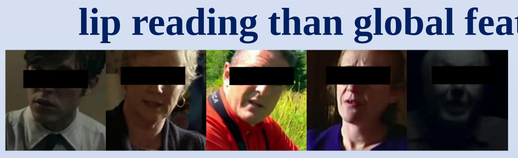
GLip: A Global-Local Integrated Progressive Framework for Robust Visual Speech Recognition

Tianyue Wang, Shuang Yang, Shiguang Shan, Xilin Chen*

Motivation

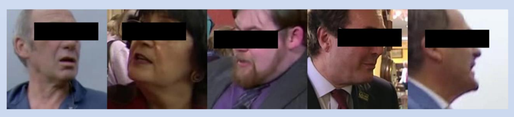
- **Existing Problems:** Current VSR methods struggle with real-world visual challenges (e.g., illumination variations, occlusions, blurring, and pose). Data collection and augmentation approaches are costly and lack generalization
- **Motivated by two key insights:**
  - Learning an initial coarse alignment between visual features across varying conditions and corresponding speech content facilitates the subsequent learning of precise visual-to-speech mappings in challenging environments;
  - Certain local regions often exhibit more discriminative cues for lip reading than global features under adverse conditions.



(a) Illumination

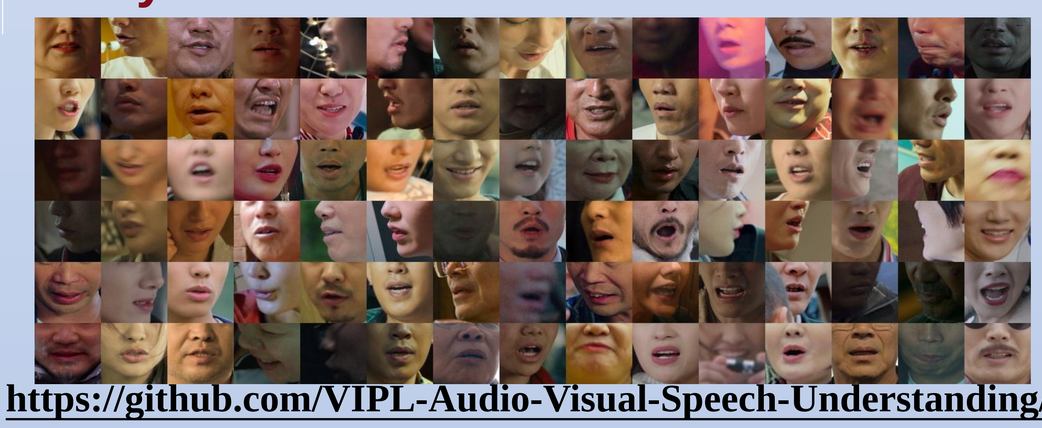
(b) Occlusions

(c) Motion blur



(d) Head poses

A Newly Introduced Dataset: CAS-VSR-MOV20



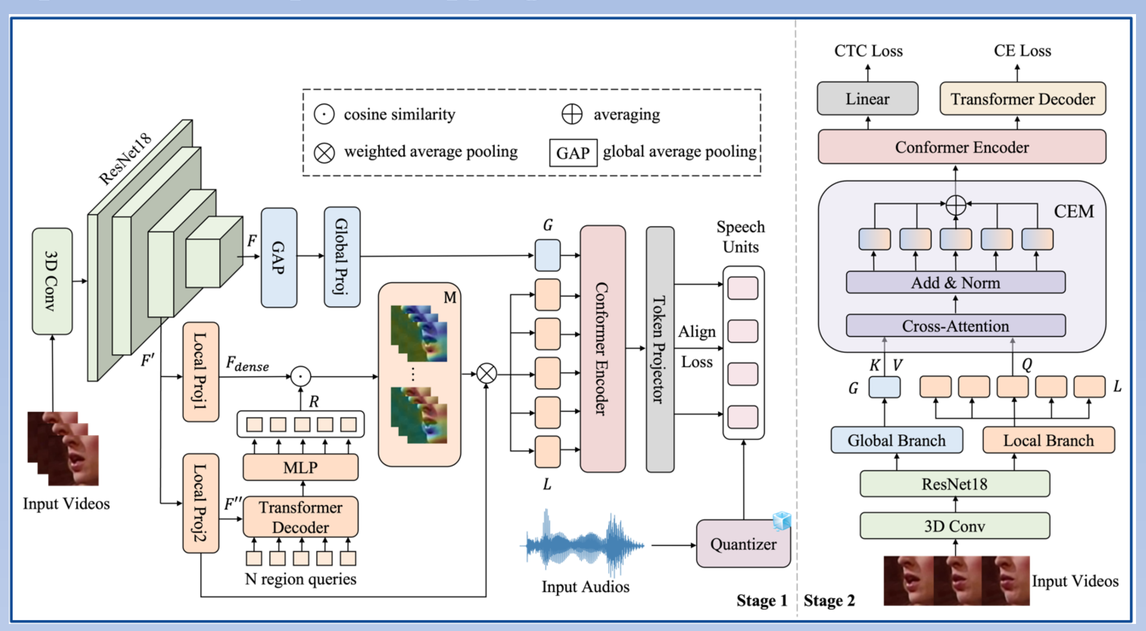
https://github.com/VIPL-Audio-Visual-Speech-Understanding/CAS-VSR-MOV20

Results on CAS-VSR-MOV20 datasets

Method	Setting	CER(%)	
		MOV20-Val	MOV20-Test
Baseline	From Scratch	93.05	91.73
GLip	From Scratch	88.04	87.78
GLip	Load Pre-trained Model	85.64	84.72

The Proposed GLip

- **GLip** is a novel two-stage framework for robust visual speech recognition:
  - **Stage 1:** Coarse global-local alignment establishes robust initial representations;
  - **Stage 2:** Context-aware refinement integrates multimodal cues for precise visual-speech mapping.



Workflow of the Glip

Comparison with Others

- **GLip** achieves state-of-the-art performance across various benchmarks

Comparison on LRS2 and LRS3 datasets

Method	Input Scale	Unlab hours	Lab hours	WER	
				LRS2	LRS3
MV-WAS [8]	224 × 224	–	223	70.4	–
TDNN [35]	112 × 112	–	223	48.9	–
CM-Seq2Seq [25]	96 × 96	–	223/438	39.1	46.9
CM-Aux [26]	96 × 96	–	223/438	32.9	37.9
LiRA [24]	96 × 96	438	223	38.8	–
RAVEn [13]	96 × 96	433	223/433	32.1	39.1
AutoAVSR [27]	96 × 96	–	438	–	36.3
ES^3 [36]	96 × 96	433	223/433	28.7	37.9
SyncVSR [3]	128 × 128	–	223/438	28.9	31.2
GLip (Ours)	96 × 96	223	223	**28.1**	–
GLip (Ours)	96 × 96	433	223/433	**27.4**	**30.1**

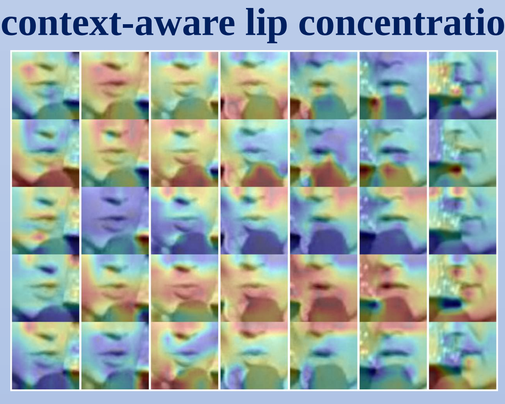
Ablation Study

- The progressive learning, the dual-path architecture, and the CEM module collectively boost performance

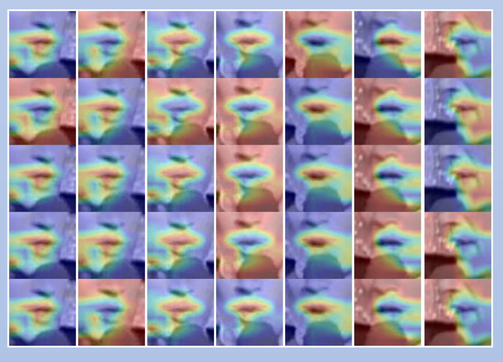
Method	Branches		Fusion	WER
	Stage1	Stage2		
Baseline	×	Global	×	30.01
Ours	Global	Global	×	28.45
	Global&Local	Global	×	28.24
	Global&Local	Local	×	28.30
	Global&Local	Global&Local	AVG	28.94
	Global&Local	Global&Local	CEM	**28.10**

Qualitative Results

- **GLip** evolves attention from a diffuse facial focus to precise, context-aware lip concentration



(a) Stage 1



(b) Stage 2

Under Diverse Visual Conditions

- **GLip** demonstrates superior robustness across various challenging visual conditions

Dataset	Method	Illumination			Occlusion		Blur			Pose		
		B	M	D	N	Y	C	M	B	S	M	L
LRS3	RAVEn[13]	48.82	37.97	41.79	38.52	41.77	32.84	40.61	52.34	37.42	42.77	44.04
	AutoAVSR[27]	37.05	34.25	37.50	34.01	40.99	30.21	35.03	47.25	33.22	39.28	37.61
	GLip (Ours)	**32.05**	**29.35**	**33.50**	**30.51**	**35.10**	**24.64**	**30.53**	**44.15**	**28.79**	**33.63**	**31.19**
LRS2	AutoAVSR[27]	30.44	28.33	34.70	29.48	38.68	30.16	29.64	32.89	29.51	29.42	49.16
	SyncVSR[3]	28.74	26.87	33.87	28.07	38.68	28.05	28.89	31.91	28.08	28.30	47.49
	GLip (Ours)	**28.40**	**26.82**	**31.69**	**27.58**	**36.84**	**27.98**	**27.96**	**30.92**	**27.46**	**28.04**	**44.13**

Tianyue Wang received the B.S. degree in Artificial Intelligence science from Beijing Institute of Technology, Beijing China, in 2025. She is currently pursing the Ph.D. degree majoring in AI with the University of Chinese Academy of Science. **E-mail:** wangtianyue33@163.com

Corresponding author: Shuang Yang is an associate professor with the Institute of Computing Technology, CAS. Her research interests include computer vision, lipreading, pattern recognition etc. shuang.yang@ict.ac.cn; https://vipl.ict.ac.cn/people/~syang