AI

Supplementary Materials

Supplementary Equation S1. Video Swin Transformer

Let the input video be represented as a 4D tensor $X \in R^{T\times H\times W\times C}$, where T, H, W, and C denote temporal frames, height, width, and channels. The input is divided into nonoverlapping 3D patches of size $P_T \times P_H \times P_W$, producing tokens that are linearly projected onto a higher-dimensional space.

Each Swin stage consists of 3D window-based self-attention:

$$Attention(Q,K,V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}} + B\right)V$$

where B is the 3D relative position bias.

Shifted windows were alternately used to enable cross-window communication. Patch merging combines 2×2 neighboring patches using a linear transformation:

$$\mathbf{z}'' = \mathbf{Concat}(\mathbf{z_1}, \mathbf{z_2}, \mathbf{z_3}, \mathbf{z_4}) \cdot \mathbf{W} \in \mathbb{R}^{\mathbf{4C\times 2C}}$$

Supplementary Equation S2. Conformer Encoder

Each Conformer block includes

- Feedforward layers:

$$FFN(z) = Dropout(\sigma(zW_1 + b_1)W_2 + b_2)$$

- Multi-head attention:

$$MHSA(Q,K,V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

- Depthwise convolution:

$$Conv(z) = GLU\big(DepthwiseConvD(z)\big)$$

All layers use residual connections and normalization.

Supplementary Equation S3. Transformer Decoder

Each decoder layer contains

- Masked multi-head attention (causal masking):

$$MaskedAttention(Q,K,V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}} + M\right)V$$

- Cross-attention with encoder outputs;
- Feedforward layers with GELU.

$$MaskedAttention(Q,K,V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}} + M\right)V$$

Supplementary Equation S4. CTC/Attention Hybrid Loss

Training objective:

$$L_{total} = \lambda L_{CTC} + (1-\lambda)L_{Attention}$$

Decoding uses joint beam search:

https://doi.org/10.3390/ai6090222

$$C^* = arg_c^{mac}[\lambda logP_{CTC}(C|X) + (1-\lambda) logP_{Attention}(C|X)]$$

Table S1. Comparison of our dataset with existing word-level audio–visual speech datasets across key metrics such as tasks, speakers, and resolution.

Dataset	Year	Task	Speakers	Classes	Utterances	Resolution	Pose
Tulips1 [35]	1996	Digits	12	4	96	100 × 75	Frontal
M2VTS [36]	1997	Digits	37	10	1K	768 × 576	Frontal
AVLetters [37]	1998	Alphabets	10	26	780	376× 288	Frontal
AVLetters2 [38]	2008	Alphabets	5	29	910	1920× 1080	Frontal
MIRACL-VC1 [39]	2011	Words	15	10	1.5K	640 × 480	Frontal
AVDigits [40]	2018	Digits	**53**	10	795	1280× 780	0,45,90
Our	2025	**9**	40	**185**	**740K**	**1920× 1080**	Frontal

Table S2. Overview of categorized keywords used for visual speech recognition dataset creation.

Alphabet 26	Number 10	System 20	Emotion 27	Daily 25	Food 25	Work 17	Education 25	Finance 15
A	One	Shutdown	Happy	Chair	Apple	Work	School	Money
B	Two	Restart	Sad	Table	Banana	Job	Teacher	Bank
C	Three	Execute	Angry	Pen	Orange	Task	Student	Cash
D	Four	Configure	Excited	Pencil	Bread	Office	Class	Account
E	Five	Download	Scared	Computer	Rice	Boss	Lesson	Budget
F	Six	Lock	Nervous	Phone	Pasta	Employee	Homework	Save
G	Seven	Run	Calm	Cup	Chicken	Team	Exam	Loan
H	Eight	Open	Surprised	Plate	Beef	Meeting	Test	Credit
I	Nine	Close	Confused	Spoon	Fish	Project	Grade	Debt
J	Ten	Save	Tired	Fork	Salad	Deadline	Subject	Payment
K		Delete	Lonely	Knife	Soup	Salary	Study	Interest
L		Copy	Hopeful	Glass	Pizza	Break	Book	Profit
N		Paste	Frustrated	Bottle	Hamburger	Overtime	Notebook	Expense
M		Move	Relaxed	Bag	Sandwich	Shift	Desk	Income
O		Search	Grateful	Shoe	Egg	Resume	Blackboard	Tax
P		Update	Jealous	Hat	Cheese	Promotion	Library	
Q		Install	Proud	Jacket	Milk	Hire	University	
R		Uninstall	Embarrassed	Key	Coffee		College	
S		Backup	Curious	Door	Tea		Classroom	
T		Restore	Disappointed	Window	Cake		Pencil	
U			Annoyed	Car	Cookie		Pen	
V			Loving	Bicycle	Yogurt		Quiz	
W			Confident	Ball	Ice cream		Diploma	
X			Guilty	Toy	Tomato		Lecture	
Y			Bored	Clock	Potato		Schedule	
Z			Hopeful					
			Shy					

Table S3. Performance comparison of visual front-end architectures and encode–decoder combinations based on Average CER (%).

Visual Front-end	Encoder + Decoder	Average CER (%)	Standard Deviation
3D-CNN	Conformer +	11.29	3.76

3D-CNN + DenseNet-121	Transformer + CTC/Attention	6.86	2.41
ViT-Base		5.05	1.76
(**Our**) Dual-Stream Former		**4.06**	**1.79**

Table S4. CER (%) and 95% confidence intervals (CIs) across models, computed via participant-wise bootstrap resampling (1,000 iterations).

Visual Front-end	CER mean (%)	95% CI low	95% CI high
3D-CNN	11.29	10.54	12.03
3D-CNN + DenseNet-121	6.86	6.38	7.34
ViT-Base	5.05	4.70	5.40
(**Our**) Dual-Stream Former	**4.06**	**3.70**	**4.41**