Vision-based Hand Gesture Virtual Keyboard–Mouse Framework

Sammam Mahdi · Reshad Ul Karim · Tareq Sujat · Syeda Maliha Tabassum · Abrar Samin · Aniqua Nusrat Zereen

Presenter: Dr. Aniqua Nusrat Zereen

Mahidol University

The 23rd International Joint Conference on Computer Science and Software Engineering (JCSSE 2026)



Introduction & Problem Statement

Motivation

- Touchless HCI is rising in AR/VR, wearables & accessibility
- Enables contactless, hygienic mid-air typing and pointing
- Needs to serve multiple languages / scripts, not just English

Problem — Limits of Prior Systems

- Small gesture vocabularies (often 5–7 gestures)
- Lighting sensitivity & cluttered-scene degradation
- High compute cost; many need special hardware
- Weak multilingual support & no cross-platform eval

Objective

A modular, non-probabilistic, real-time touchless keyboard-mouse with bilingual next-word prediction

01

Dual-Mode Interaction

QWERTY (EN) & phonetic (BN) keyboards + virtual mouse (move, click, drag-drop) from one gesture set

02

Multilingual Word Prediction

Two-layer LSTM for both language · top-1 > 97% · perplexity < 1.15 · real-time top-K suggestions

03

Cross-Platform Validation

5 hardware platforms · up to 37.7 FPS · CER 3.43% (EN) / 8.34% (BN) · 7 lighting conditions

Related Work & Research Gap

Related Work & Research Gap



RESEARCH GAP

- Limited gesture diversity
- Poor illumination robustness
- No real-time cross-platform benchmark
- Missing bilingual next-word prediction
- Reliance on special hardware

Proposed Methodology

Sequential Real-time Modules

- Camera feed → MediaPipe landmark localization (21 pts/hand)
- Non-probabilistic Gesture Detection Engine
- Unified Keyboard–Mouse Interface
- Next-Word Prediction Engine (LSTM)
- Render text in target application
- Performance Monitor: FPS, latency, CPU/RAM

All modules run sequentially in a single real-time loop

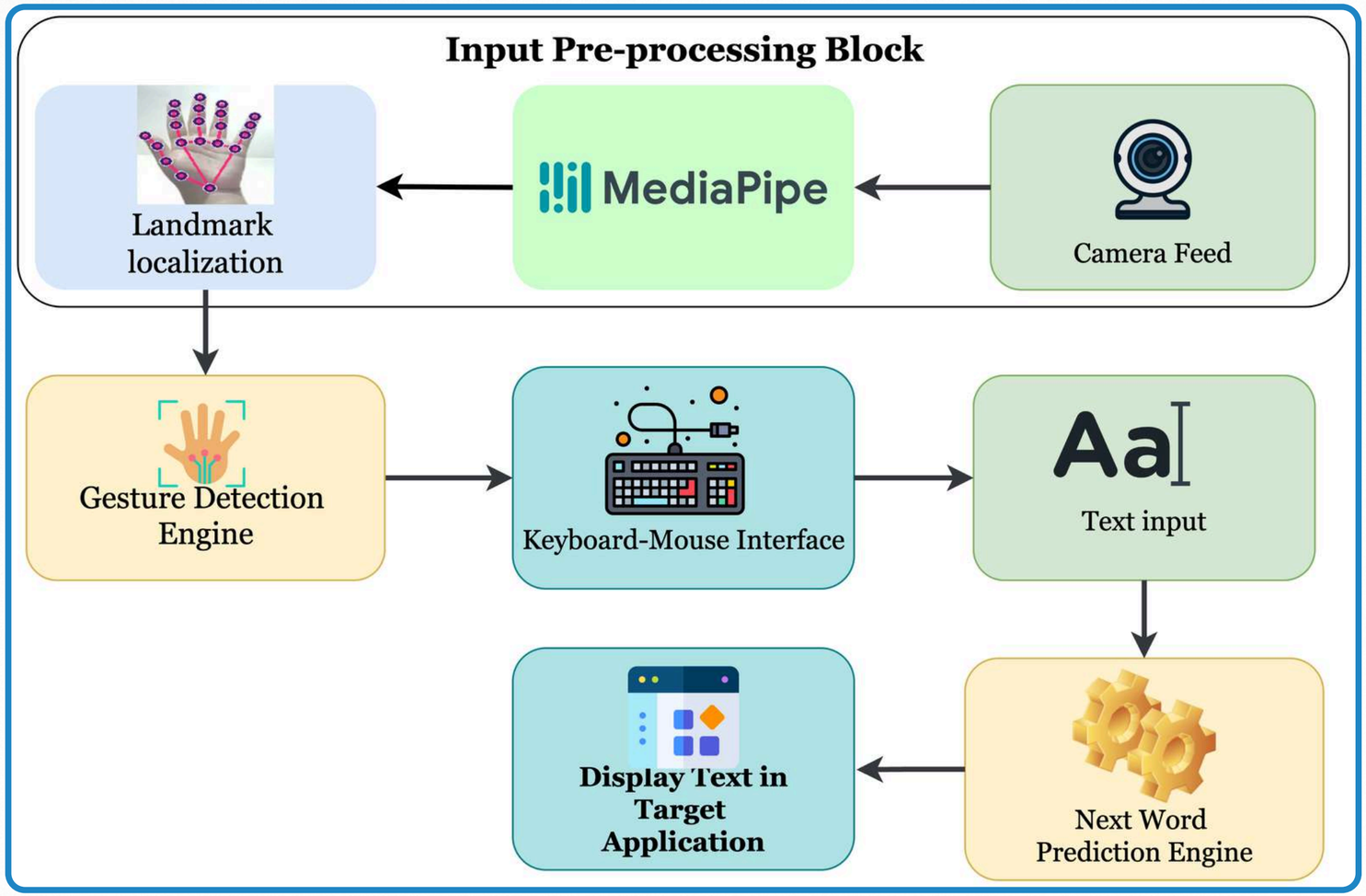
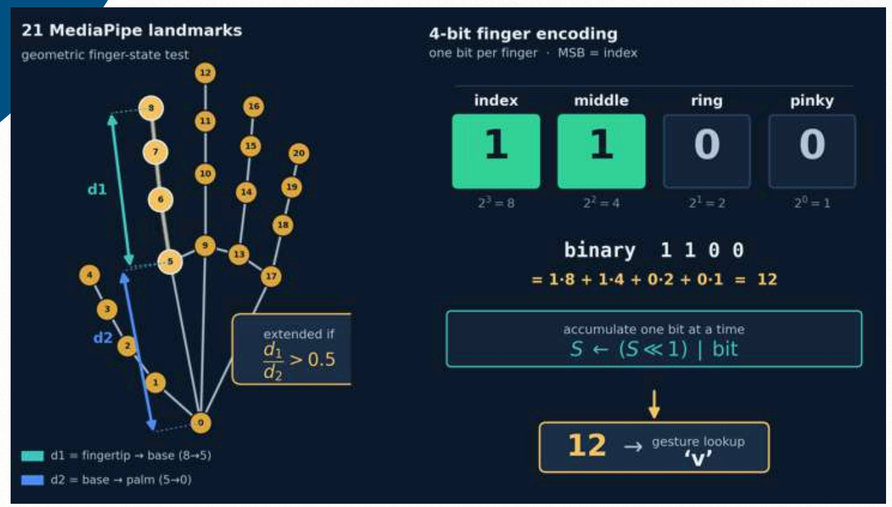
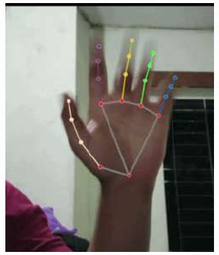


Fig. 1 — Modular pipeline of the virtual keyboard–mouse application

Gesture Detection Engine



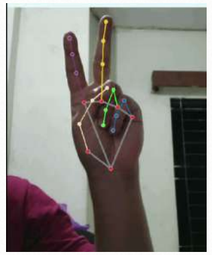
Gesture	Action
v (V-sign)	Move the cursor
left click	Left click
right click	Right click
index mid closed	Double-click
grab → open hand	Drag-and-drop
last3	switch number to letter in keyboard
first3	switch keyboard to mouse
Pinky down	Press a key on the keyboard



Grab/Open hands



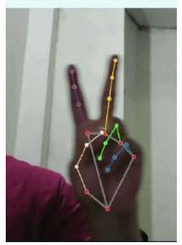
Right Click



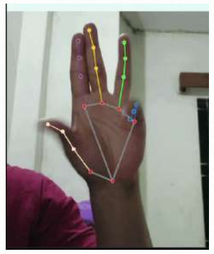
Left Click



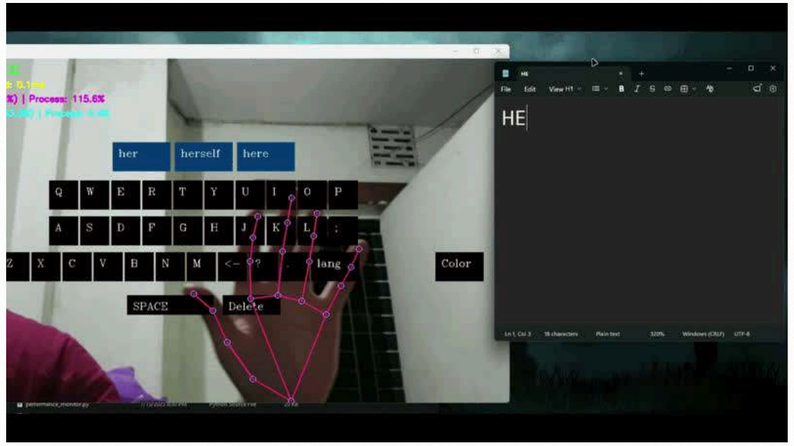
First3/Last3



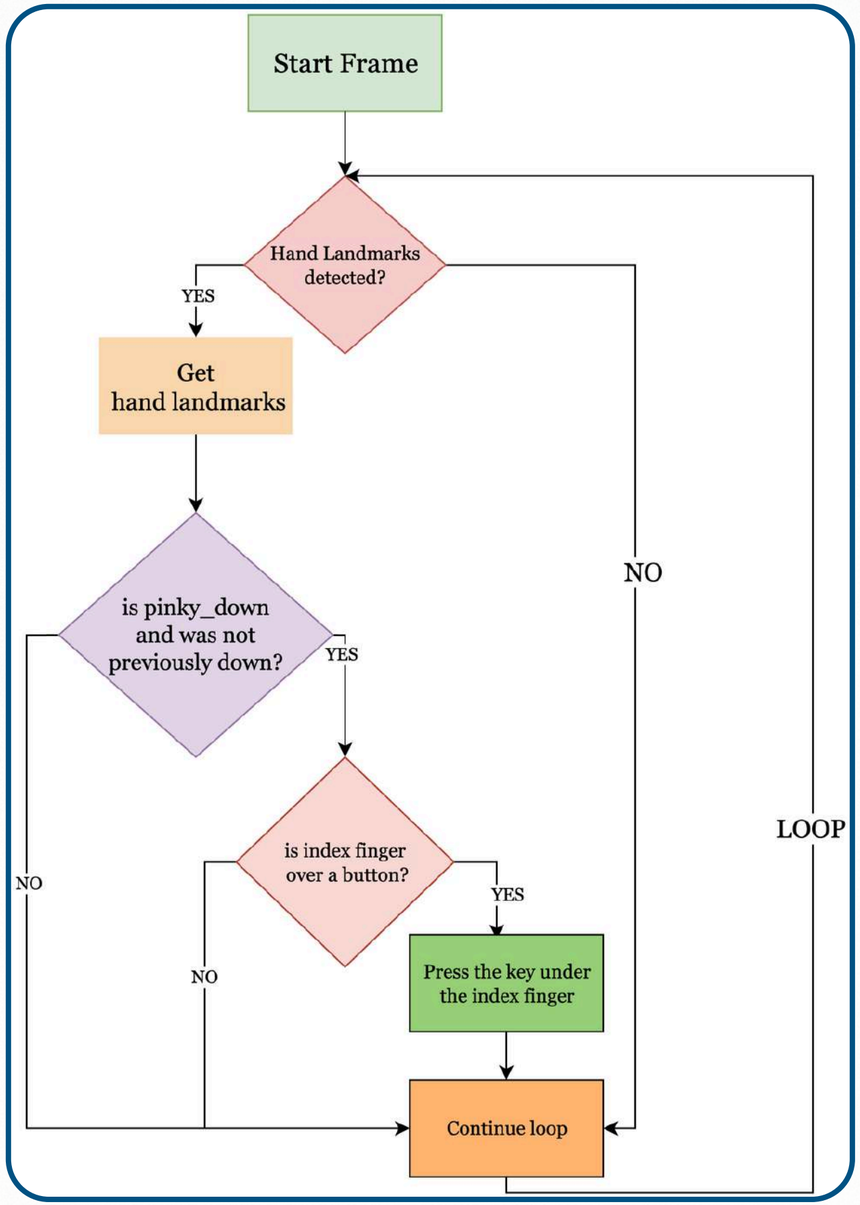
V



Index mid Closed



Typing (Pinky Down)



Interface, Dataset & Pre-processing

77.5K+
English docs (Gutenberg)

11.5K
Bangla news articles

10K
Words per language

2,000
Vocab tokens (+PAD/UNK)

Clean text | Remove punctuation | Lowercase (EN) | Tokenize | 80:20 split

Rendering & Interaction

- MediaPipe normalizes hand scale/position across frames
- English glyphs via OpenCV; Bangla via HarfBuzz + PyQt5
- Hover = key highlight; pinky-down = select
- Special keys: lang, pred-3 words, Color for UI

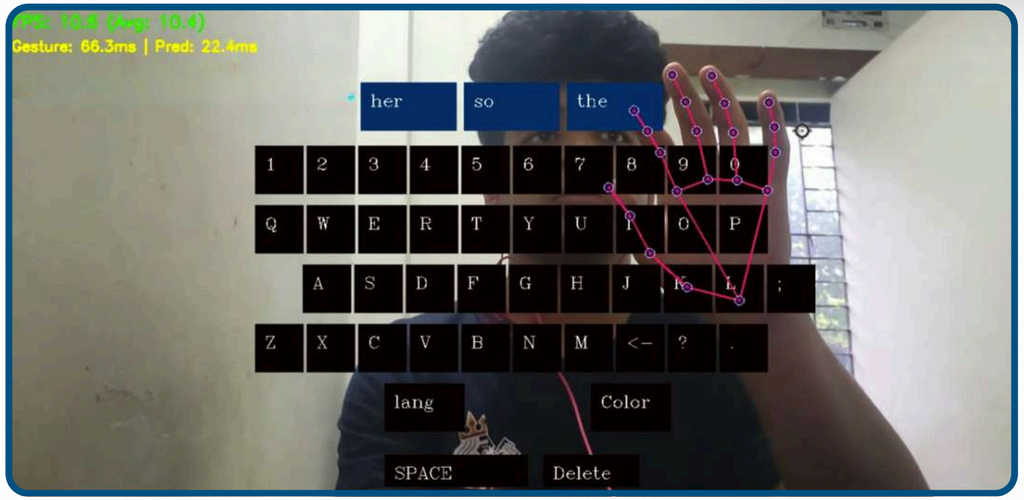


Fig. 4 — English QWERTY virtual keyboard

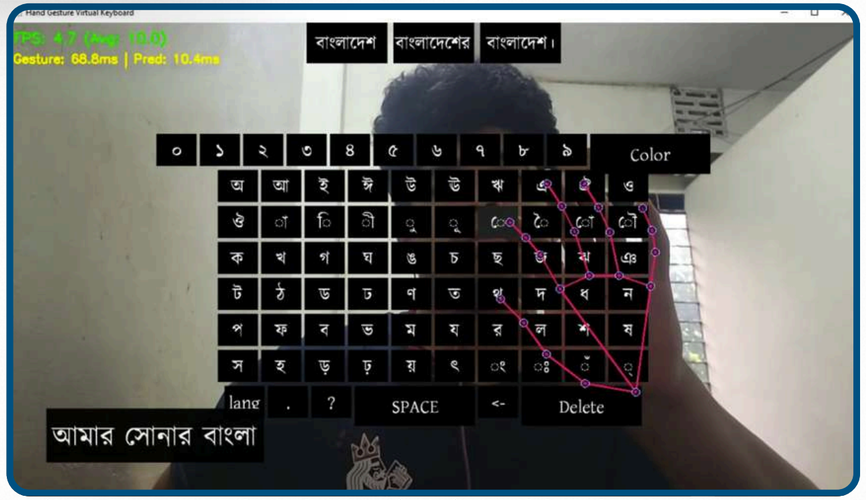


Fig. 5 — Bangla phonetic virtual keyboard

Language Model & Experimental Setup

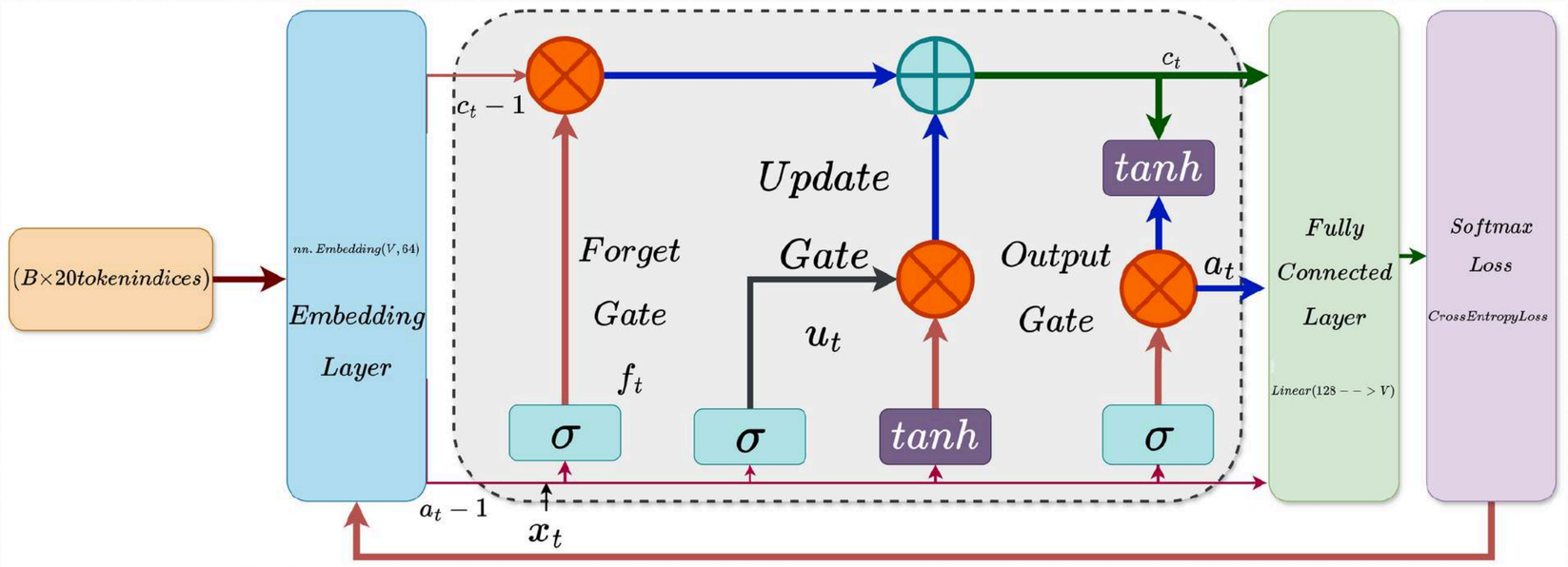


Fig. 6 — Two-layer LSTM cell: embedding → forget/update/output gates → softmax

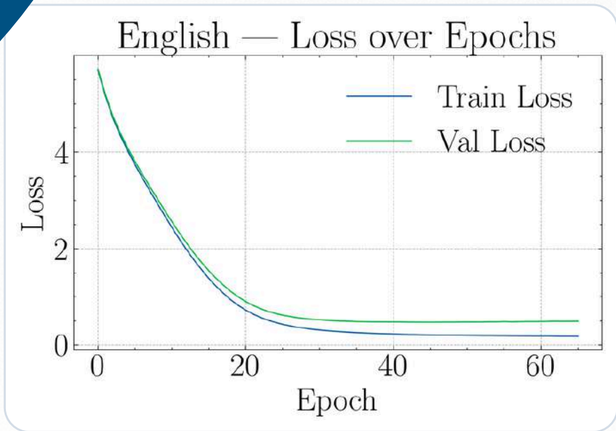
LSTM Configuration

- 2-layer LSTM, 128 units, 64-dim embedding
- Dropout 0.5 + batch normalization
- Adam ($\alpha=10^{-3}$, $L_2=10^{-5}$), ≤100 epochs, early stop (patience 5)
- Auto EN/BN detection (Unicode vs ASCII); top-K softmax suggestions

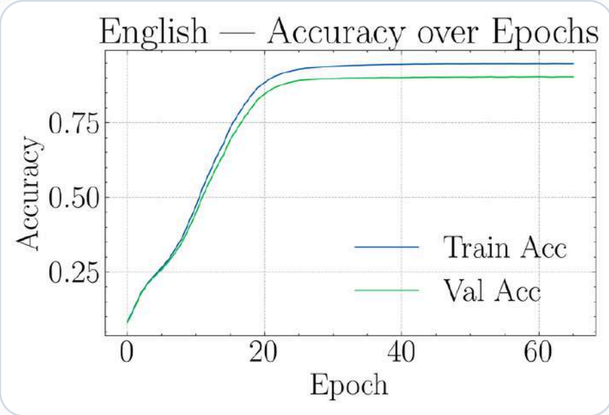
Evaluation Protocol

- 5 platforms: Ryzen 7 6800H · MacBook M2 · i5-12th · i5-4th · Ryzen 5 5600G
- Metrics: loss, perplexity, top-K acc, FPS, latency, CPU/RAM, CER
- 7 lighting levels (50–2000 lux) · 3 Fitzpatrick skin types (II/IV/V)

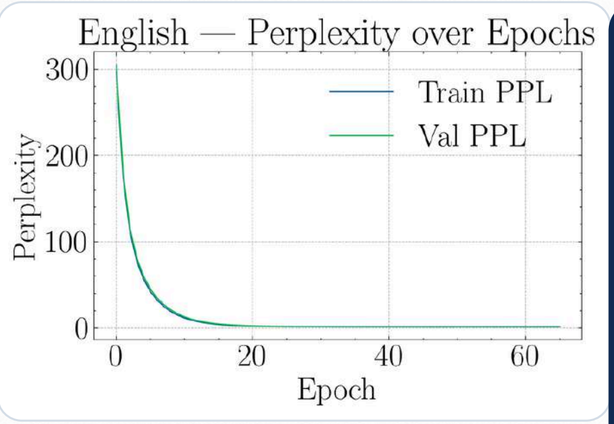
Results — Language Model Training



(a) English — Loss



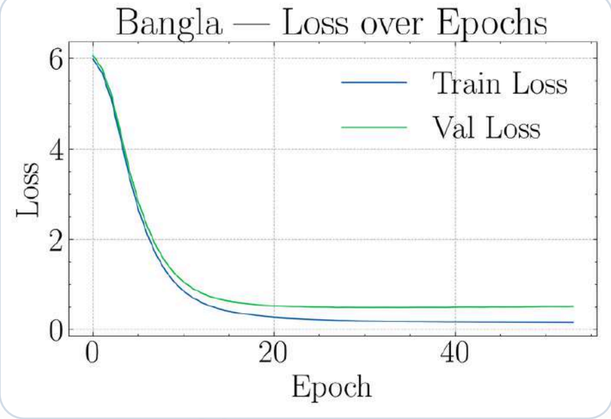
(b) English — Accuracy



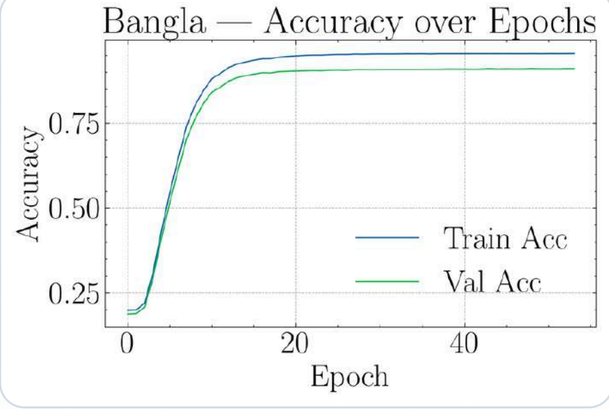
(c) English — Perplexity

English LSTM

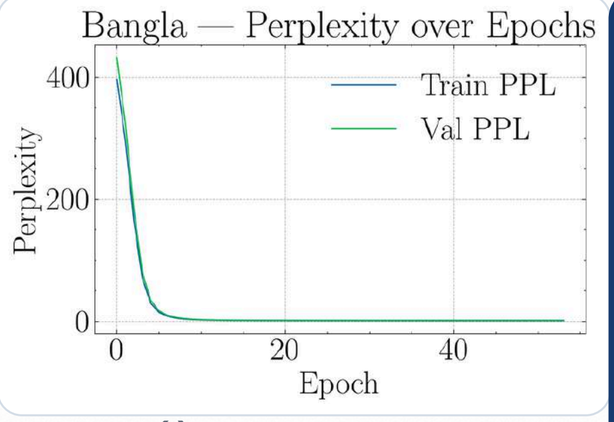
- Top-1 **97.00%**
- Top-3 **99.11%**
- Perplexity **1.146**
- Loss **0.137**



(d) Bangla — Loss



(e) Bangla — Accuracy

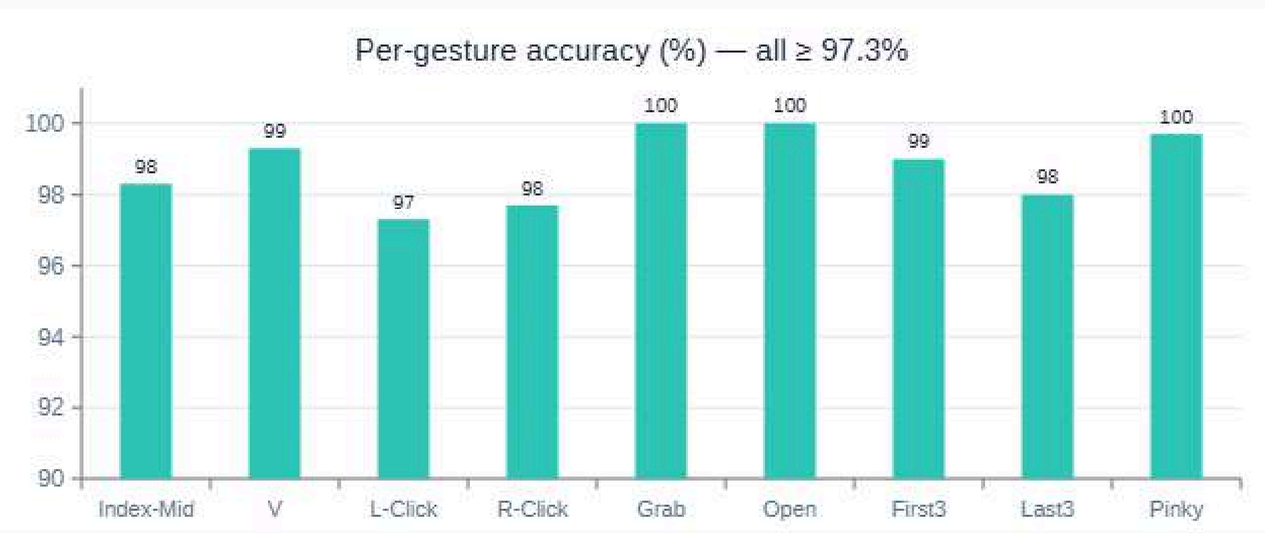


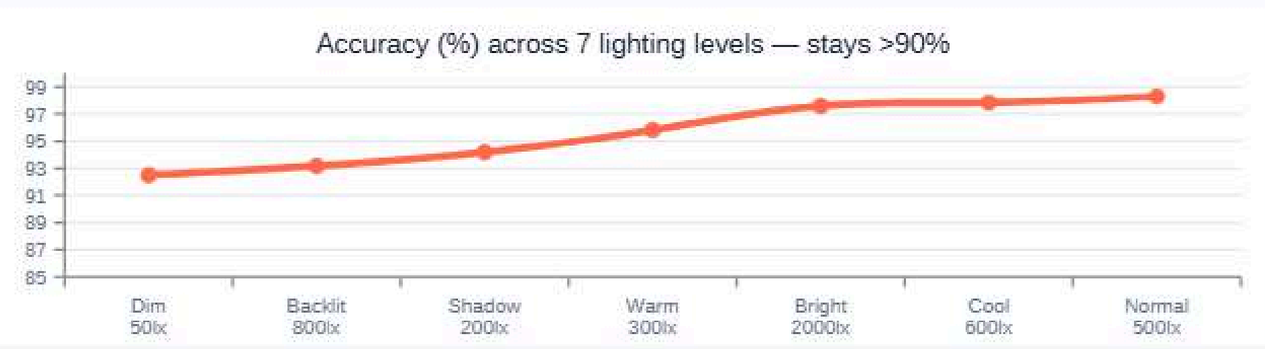
(f) Bangla — Perplexity

Bangla LSTM

- Top-1 **97.16%**
- Top-3 **99.26%**
- Perplexity **1.100**
- Loss **0.095**

Recognition Accuracy, Speed & Robustness





<0.03 ms

Recognition latency

3.43%

CER — English

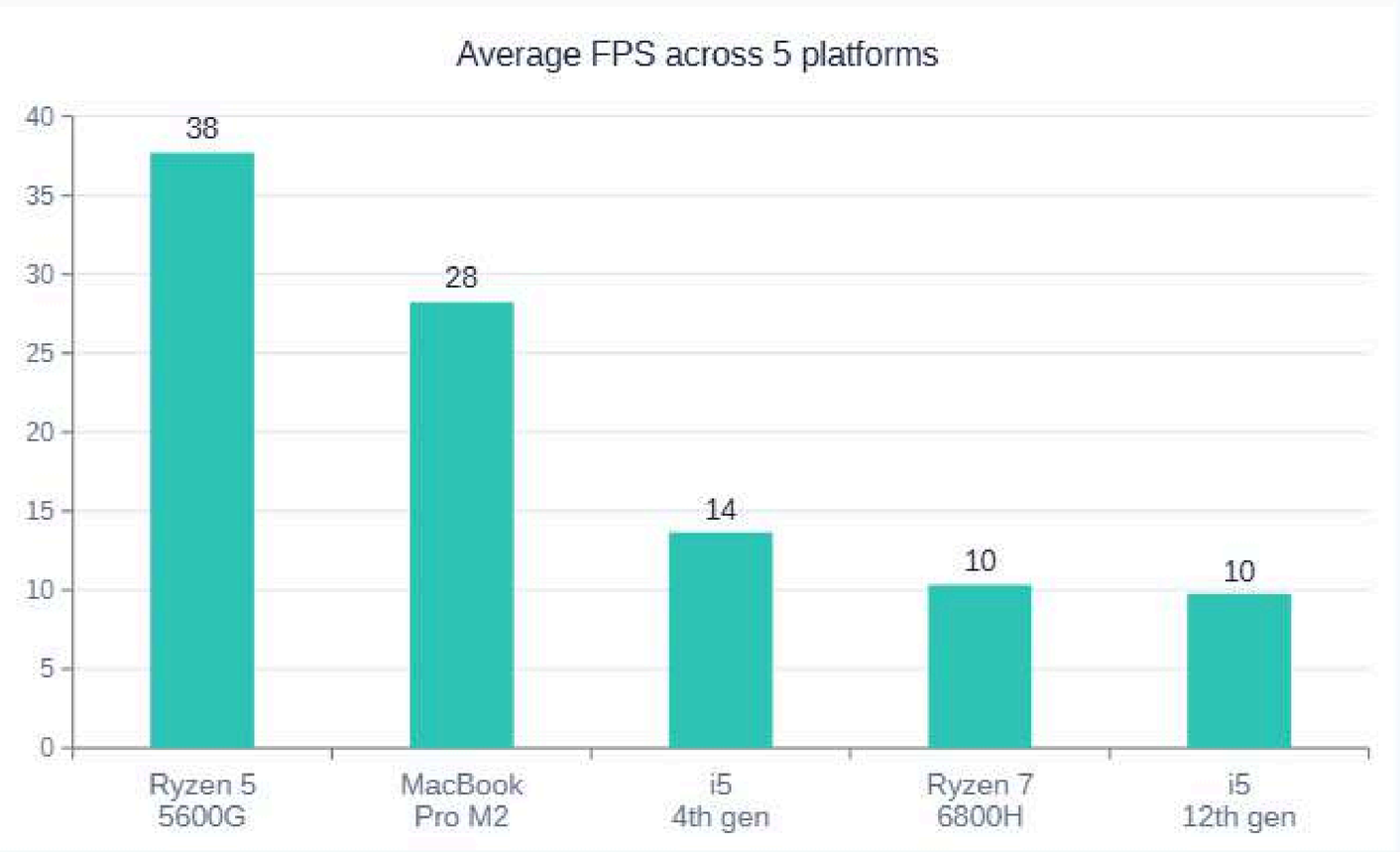
8.34%

CER — Bangla

Recognition time

Gesture	Time (ms)
index mid closed	0.0256
v	0.0166
left click	0.0048
right click	0.0047
grab	0.0049
open hand	0.0056
first3	0.0053
last3	0.0040
pinky down	0.0042

Cross-Platform Comparative Analysis



OBSERVATIONS

- Ryzen 5 5600G fastest: 37.7 FPS, lowest total latency
- i5-4th gen: 93% CPU yet still real-time
- $p < 0.05$ in nearly all metrics → strong hardware effect

Conclusion & Future Work

Key Findings

- 9 gestures ≥ 97.4%, <0.03ms latency
- Bilingual LSTM: >97% top-1
- Real-time up to 37.7 FPS

Contributions

- Unified keyboard + mouse, one gesture set
- Non-probabilistic geometric engine
- 5-platform statistical validation

Limitations

- Accuracy dips in extreme light
- Desktop-class hardware only
- Smaller training corpus

Future Work

- Dynamic & two-hand gestures
- Lightweight transformers for edge NWP
- AR/VR & mobile deployment

Thank You