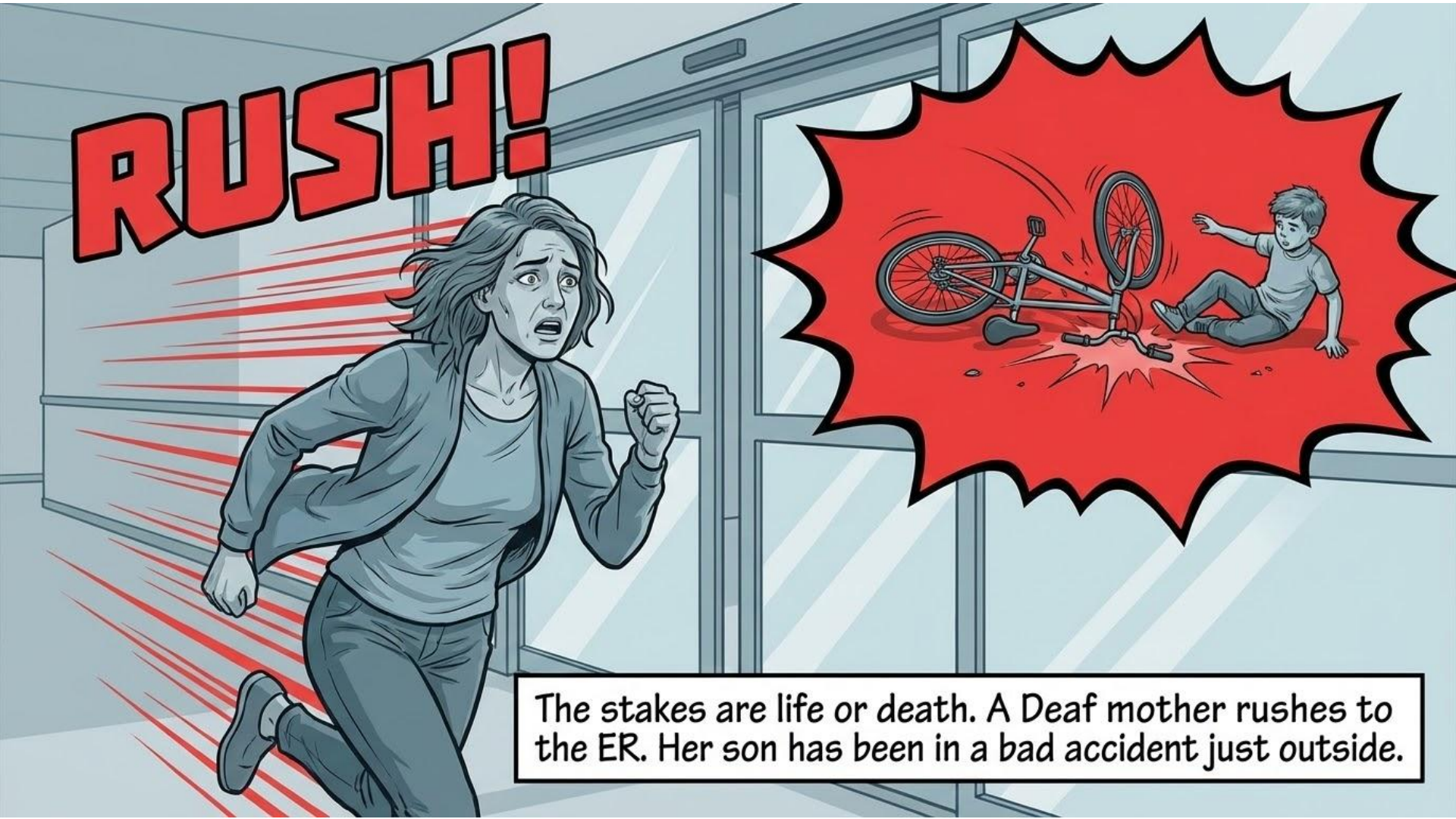
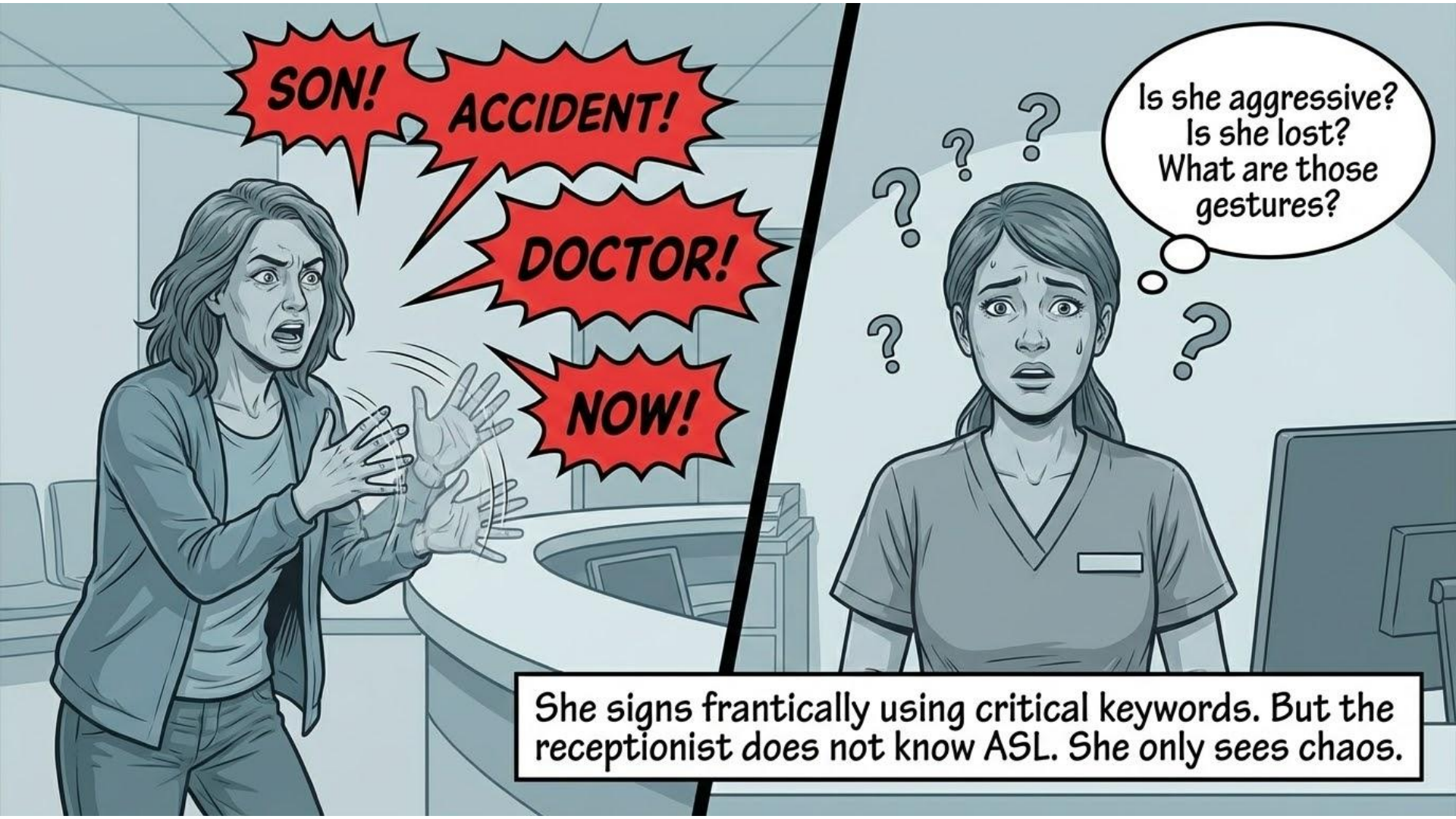
UNIVERSITI MALAYA

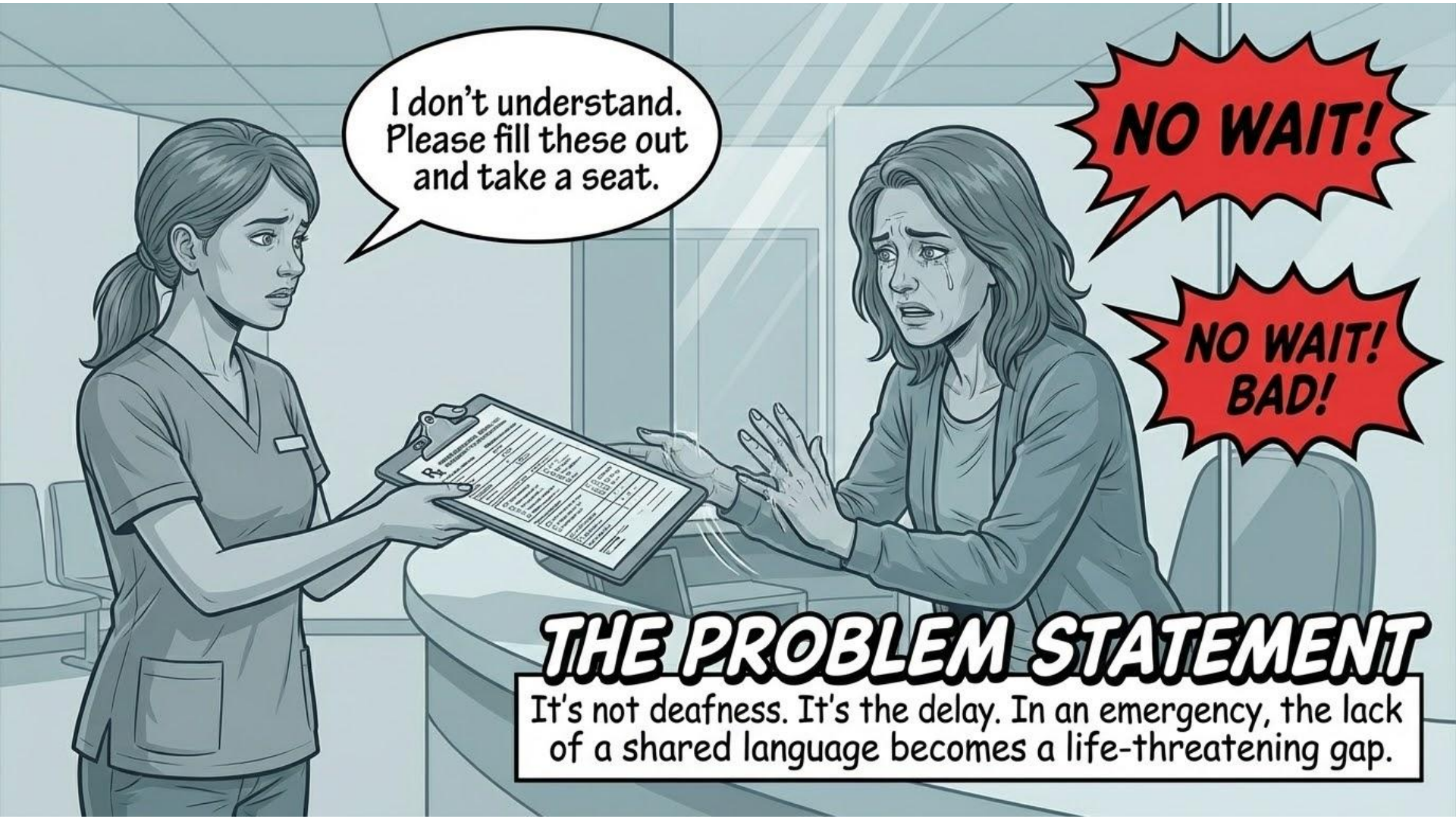
SIGN-BRIDGE

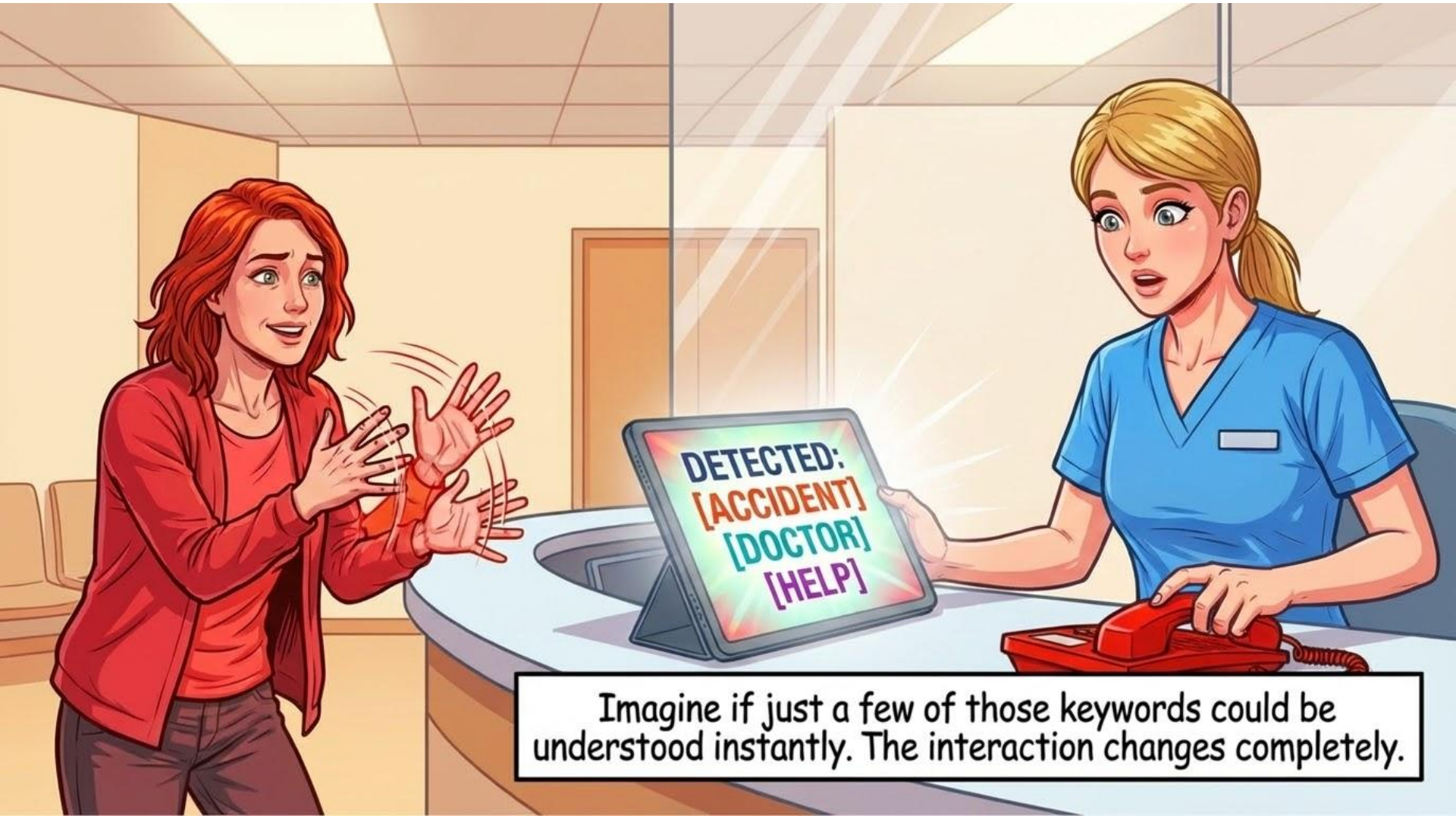
Zoe Low Pei Ee (24209891)
Aw Kai Le (24209888)
Hong Jia Herng (U2005313)
Lee Zhi Yang (22104663)
Chee Zen Yu (24088354)

Real-time ASL Video-to-Text Translation for Critical Moments.


RUSH!
The stakes are life or death. A Deaf mother rushes to the ER. Her son has been in a bad accident just outside.


SON!
ACCIDENT!
DOCTOR!
NOW!
Is she aggressive?
Is she lost?
What are those gestures?
She signs frantically using critical keywords. But the receptionist does not know ASL. She only sees chaos.


I don't understand. Please fill these out and take a seat.
NO WAIT!
NO WAIT! BAD!
THE PROBLEM STATEMENT
It's not deafness. It's the delay. In an emergency, the lack of a shared language becomes a life-threatening gap.


DETECTED:
[ACCIDENT]
[DOCTOR]
[HELP]
Imagine if just a few of those keywords could be understood instantly. The interaction changes completely.

This isn't one story. It's a systemic failure.

Lack of immediate, autonomous access to communication

JUSTIFICATION: BRIDGING THE GAP WITH TECHNOLOGY

PILLAR 1: THE HUMAN RESOURCE SHORTAGE

Severe global shortage of qualified ASL interpreters. 24/7 physical presence everywhere is impossible.

PILLAR 2: THE NEED FOR IMMEDIACY

Emergencies don't wait for Video Relay Service connections. Immediate keyword recognition saves critical time.

PILLAR 3: UBIQUITY OF HARDWARE

Leveraging existing smartphone cameras makes the solution scalable and accessible without new hardware.

Global Impact: SDG Alignment

10 REDUCED INEQUALITIES

Breaking down communication barriers to ensure equal access to services.

3 GOOD HEALTH & WELL-BEING

Ensuring language barriers do not prevent timely, accurate urgent care.

UNIVERSITI MALAYA

Dataset

Dataset Overview: WLASL (Word-Level ASL)

Total Initial Videos

21,083

(Source: WLASL)

Unique Words (Classes)

2,000

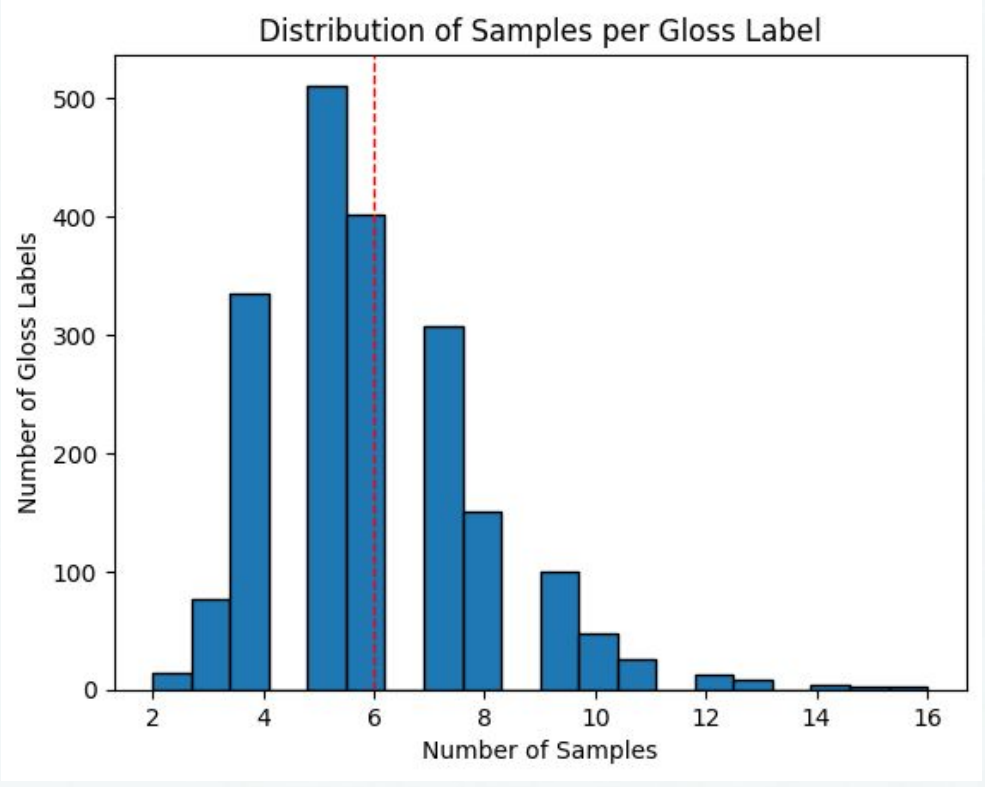
(Target vocabulary scope)

Label Integrity

100%

(No missing video IDs or Gloss labels)

The Challenge: Data Distribution Imbalance



Key Takeaways

The "Long Tail" Problem:
The distribution is highly skewed.

Average samples per gloss is 6.
The most frequent gloss has 16 samples, and the least frequent has 2.

Impact: The model will struggle to learn the rare words without specific interventions (like data augmentation).

Technical Variabilities & Quality Issues

Inconsistent Inputs (Resolution & FPS)

Huge variety in video resolutions and Frame Rates (mostly 30fps, but outliers exist).

Decoding Errors (Missing Frames)

1,408 videos (6.6%) contain unreadable frames, causing decoding errors.

Unusable Data (Black Frames)

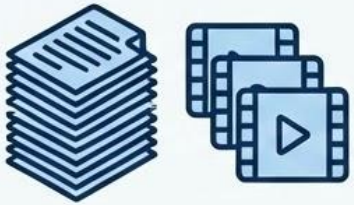
80 videos were completely blank or black screen recordings.

UNIVERSITI
MALAYA

Data Preprocessing

Data Cleaning Pipeline

Original Dataset

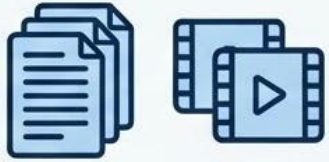


2000 Glosses,
21083 Videos

Missing Video Exclusion & Selection

Excluding Missing Videos, Selecting 35 Glosses

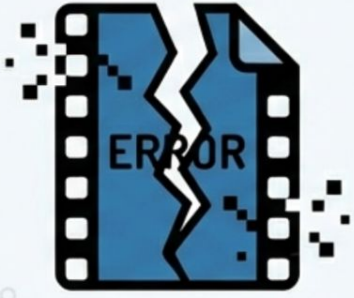
Final Cleaned Dataset



35 Glosses,
395 Total Samples

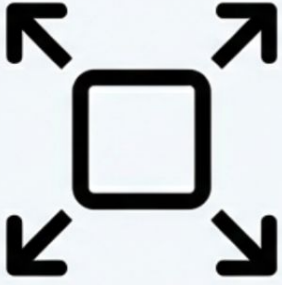
STABILIZING THE INPUT STREAM

1. Garbage Collection



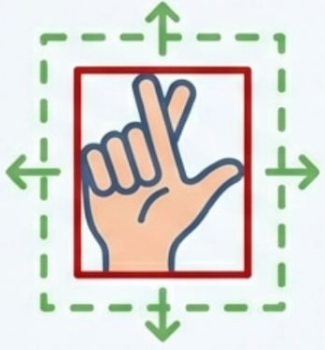
Removed 1,408 videos with decoding errors and 80 “black screen” recordings.

2. Normalization



All videos resized to uniform 256x256 resolution for consistent input.

3. BBox Expansion



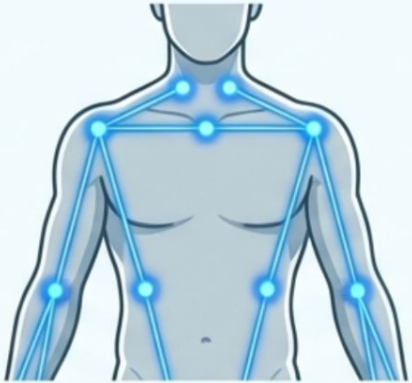
Applied 10-20% bounding box expansion to prevent cropping hand gestures during motion.

UNIVERSITI MALAYA

Model Architecture

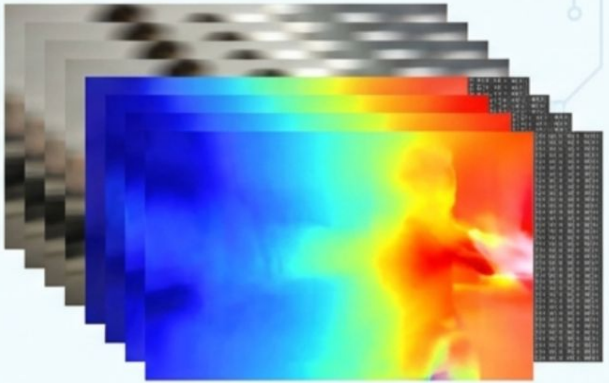
ARCHITECTURE: GEOMETRY VS. PIXELS

Approach A: Geometry-Based



- Tracks coordinates (shoulders, elbows, wrists).
- Models: Bi-LSTM (Baseline), Transformer Encoder.

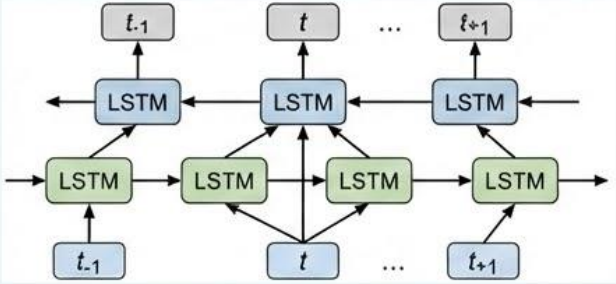
Approach B: Pixel-Based



- Analyzes raw visual features and motion dynamics.
- Models: I3D Classifier (RGB & Optical Flow).

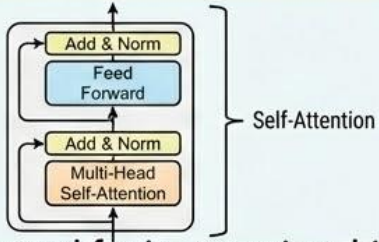
Modeling Skeletal Motion over Time

The Robust Baseline: Bi-LSTM



- **Role:** Selected as a reliable baseline for sequential data modelling.
- **Config:** 2 Layers, Bidirectional.
- **Key Mechanism:** Processes skeletal movements in both forward and reverse directions to capture immediate temporal context.

The State-of-the-Art: Transformer Encoder



- **Role:** Selected for its superior ability to model
- **Key Mechanism:** Self-Attention. Unlike RNNs that step sequentially, it can instantly relate the start of a complex sign to its end, crucial for longer, intricate movements.
- **Efficient Config:** We utilized a lightweight configuration (2 Layers, 4 Heads, d_{model}=128) proving that massive models aren't always necessary for skeletal data.

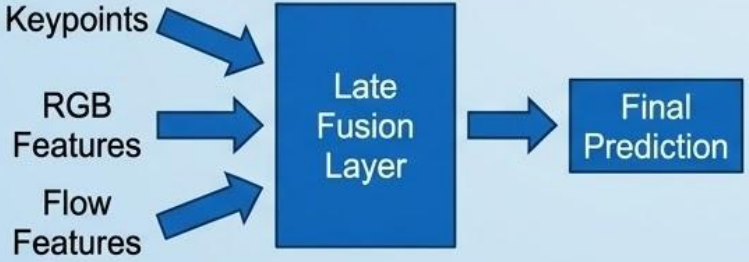
Combining Modalities for Robustness

Pixel-Based Feature Extraction (I3D)

Strategy: We utilize pre-trained I3D networks as fixed feature extractors for both RGB and Flow streams, training only a lightweight Multi-Layer Perceptron (MLP) on top.

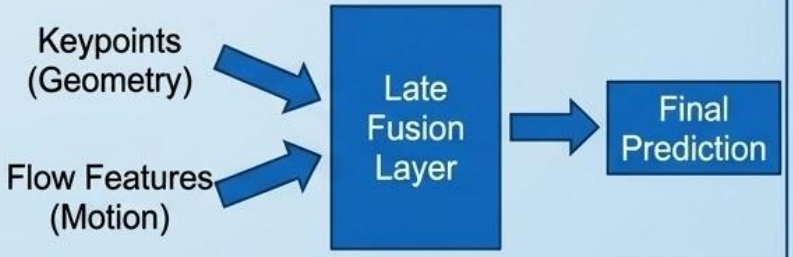
Late Fusion Architectures

Triple Stream



Description: The ultimate combination of Geometry, Appearance, and Motion dynamics.

KPS-Flow



Description: A targeted fusion focusing purely on structural movement, ignoring RGB texture to improve robustness to lighting.

Combating Overfitting on a Limited Dataset

The "Small Data" Training Recipe

Strategy 1: Label Smoothing (Calibration)

Value: 0.1

The Why: WLASL is small. Models tend to become "overconfident" quickly, memorizing training data. Label smoothing prevents the model from predicting 100% certainty, forcing it to learn softer, more generalizable patterns.

Strategy 2: Temporal Jitter Augmentation (Invariance)

Action: Randomly shifting sequence start times by ±2 frames during training.

The Why: Forces the model to be "temporally invariant." It shouldn't matter exactly which millisecond a sign starts; the model must recognize the motion pattern regardless of slight timing variations.

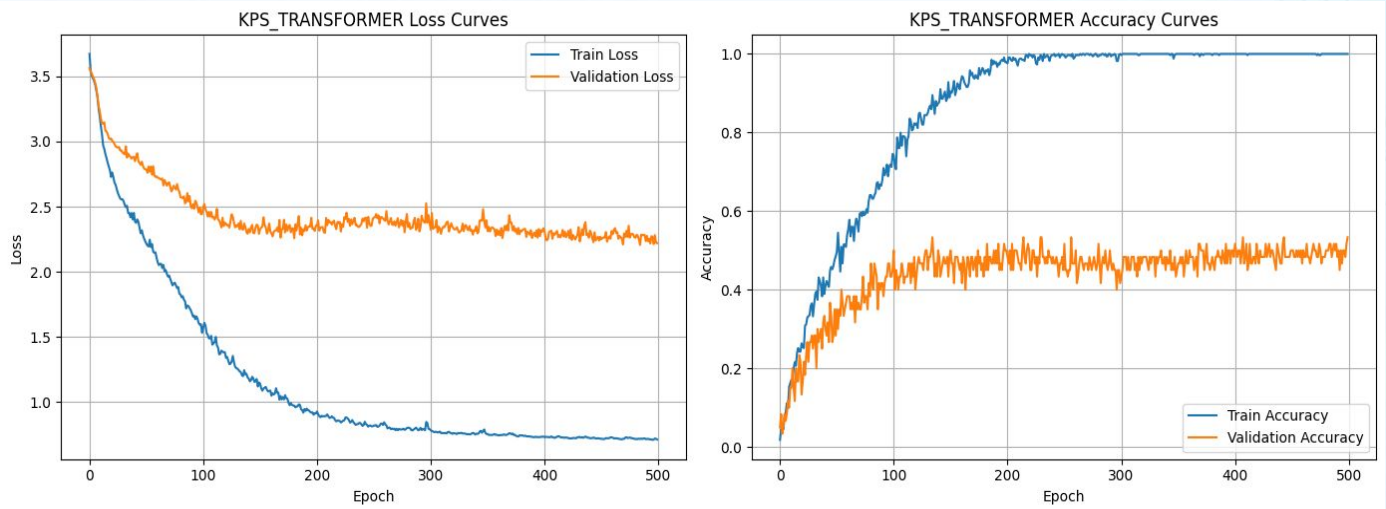
Strategy 3: Gaussian Noise Augmentation (Robustness)

Action: Adding random noise (σ=0.02) to skeletal coordinates.

The Why: Prevents the model from overfitting to exact pixel locations of joints. It ensures the model learns the general structure of the pose, not just specific coordinates in the training set.

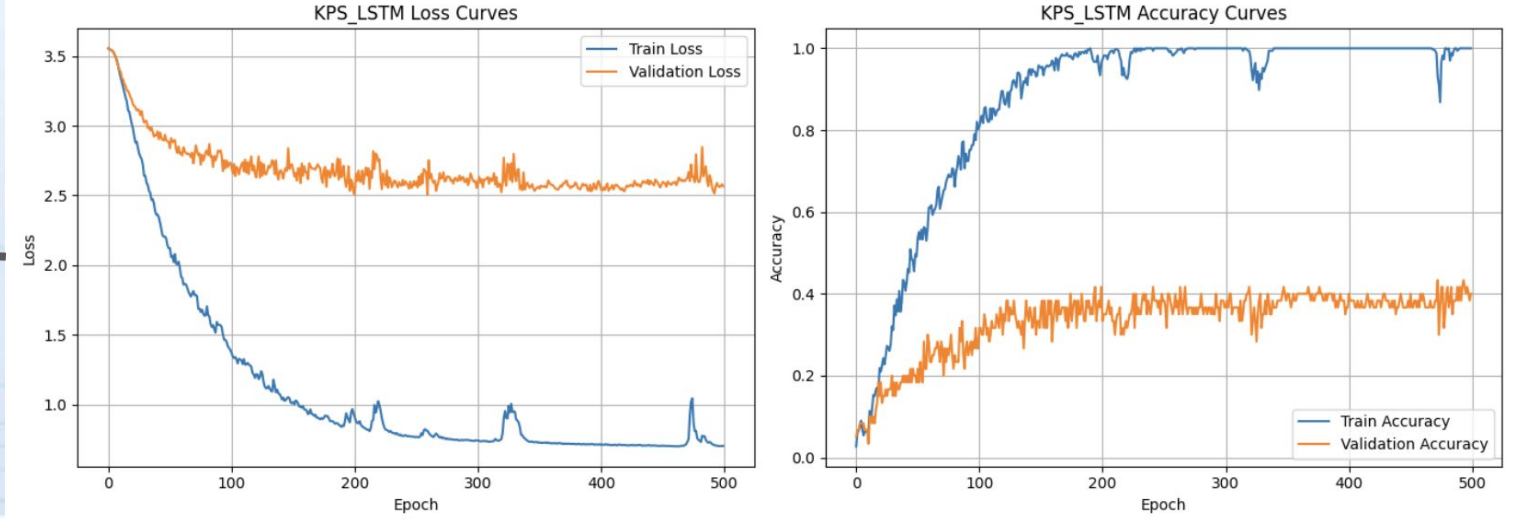
UNIVERSITI MALAYA

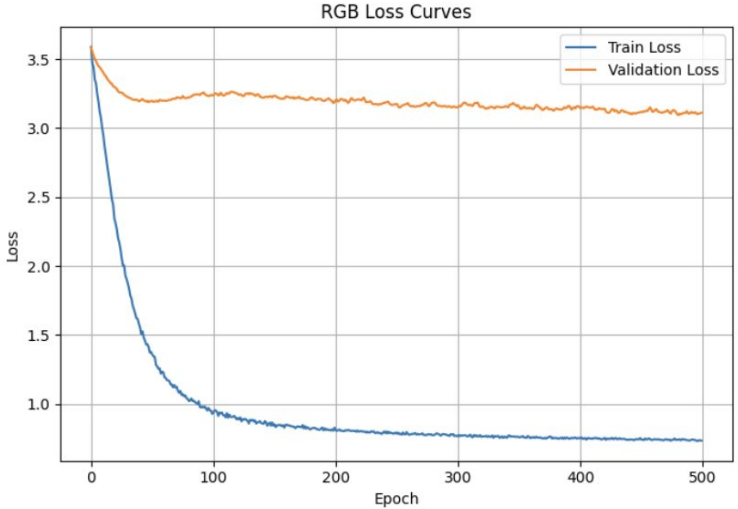
Training Plot

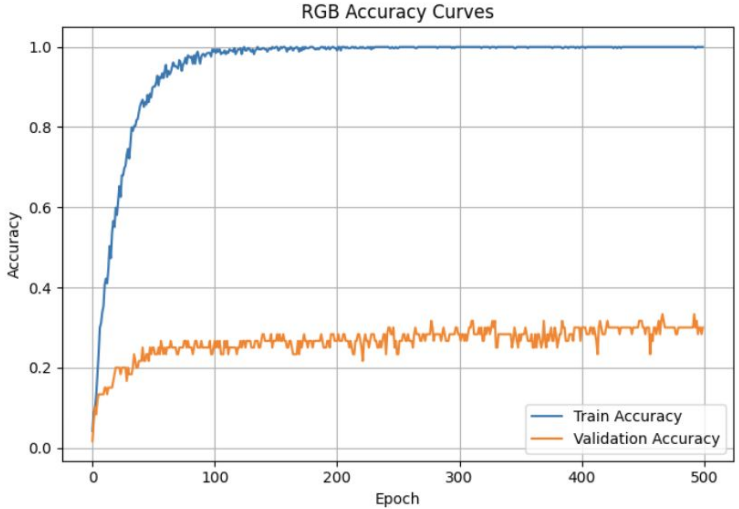


KPS Transformer

KPS LSTM

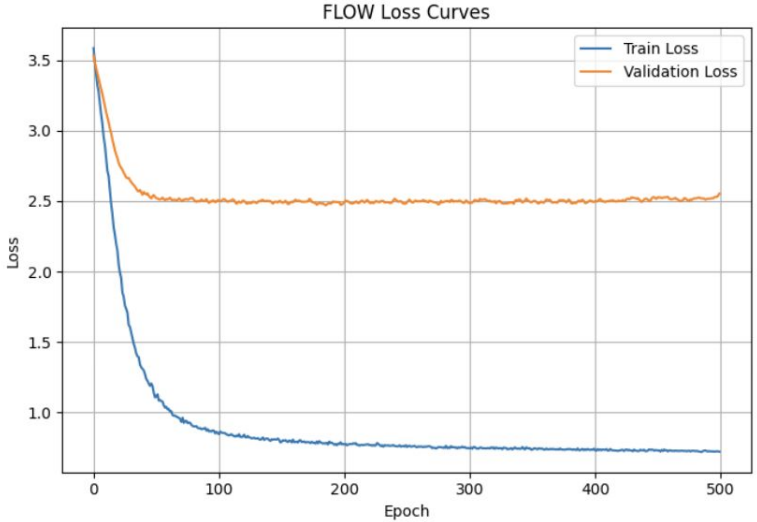


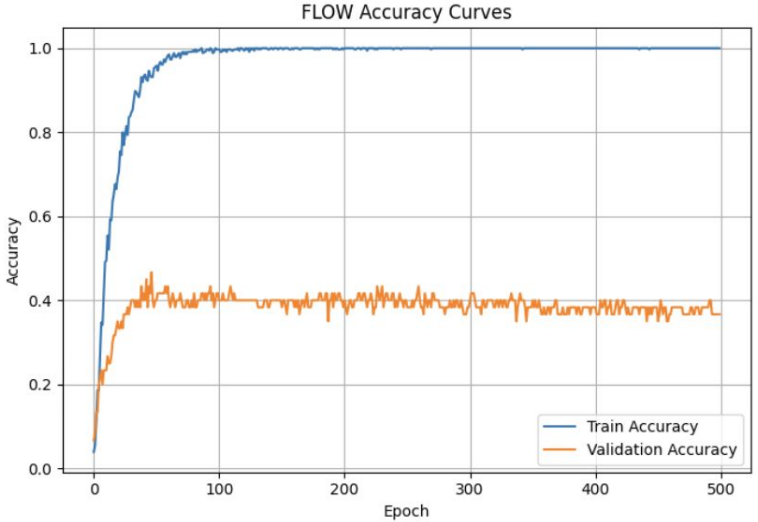


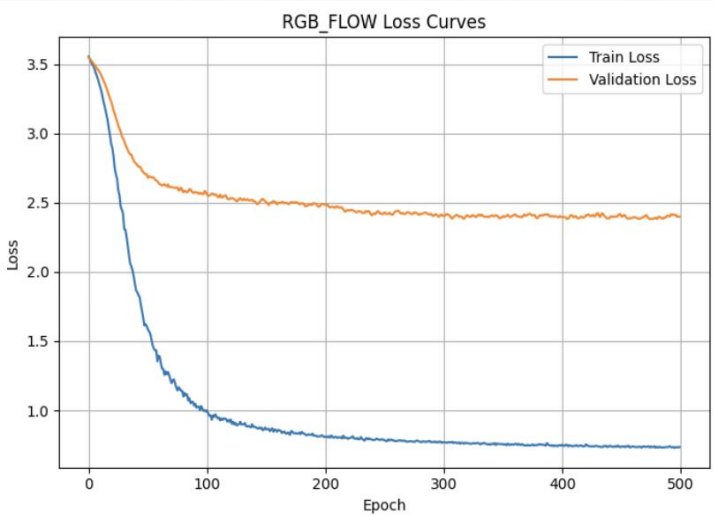


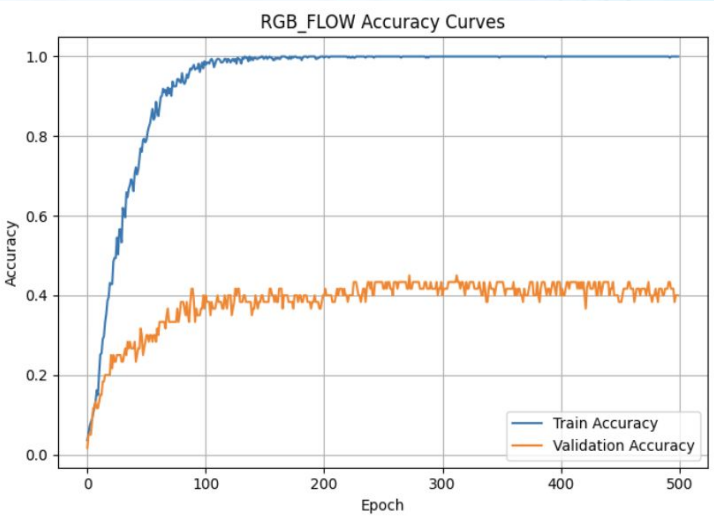
RGB I3D

Flow I3D



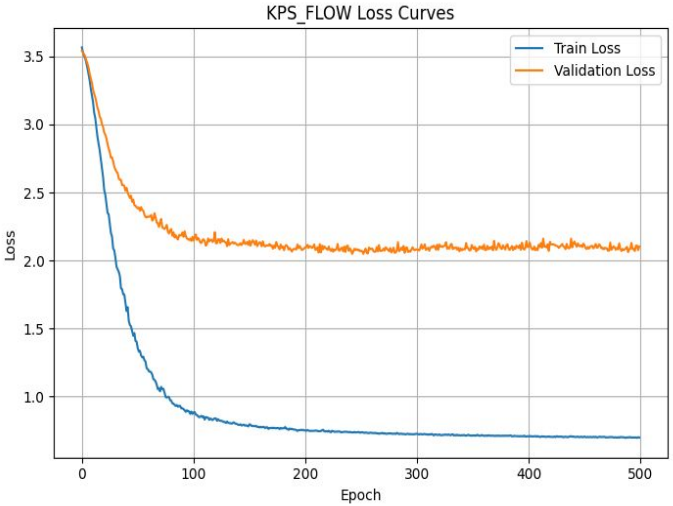


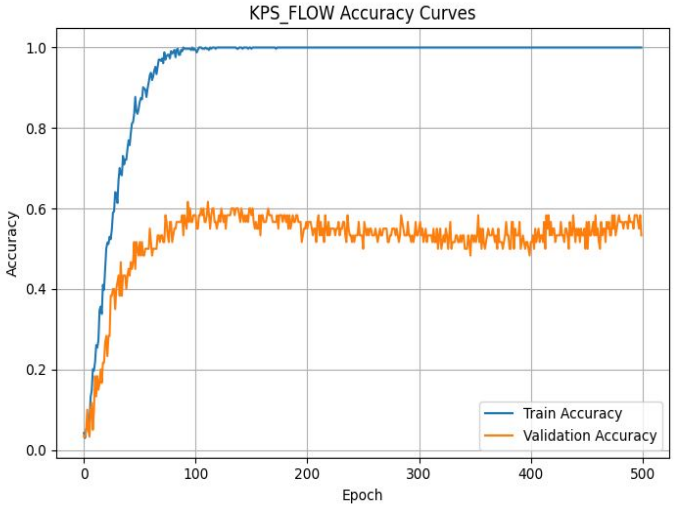


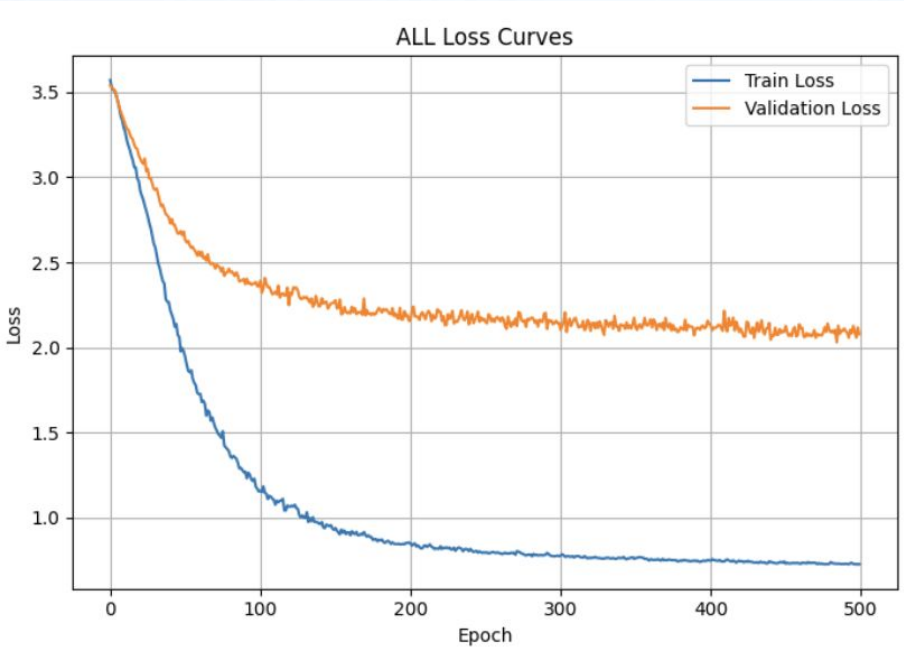


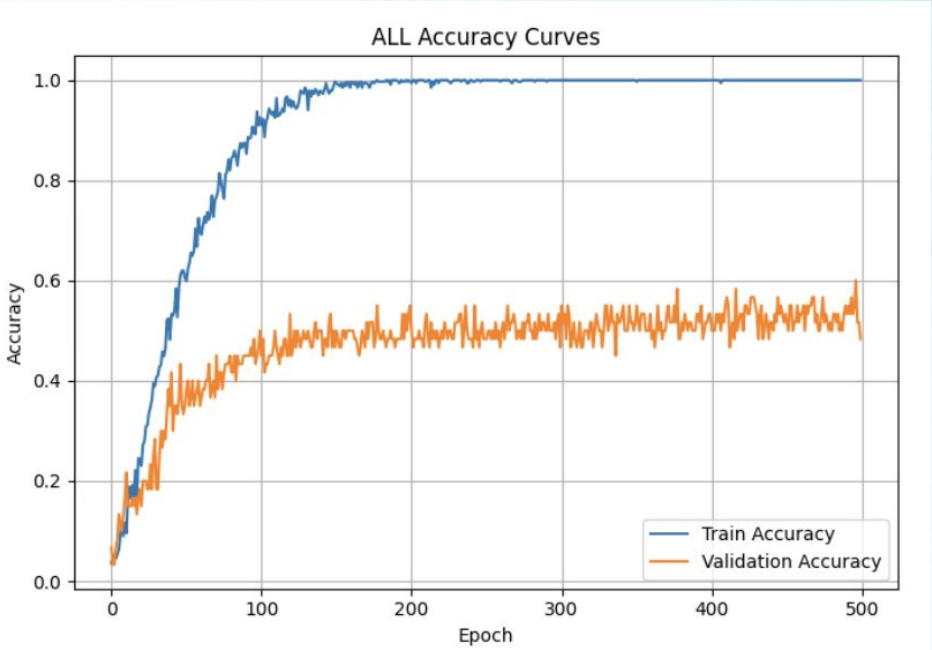
Dual Fusion (RGB,Flow)

Dual Fusion (KPS,Flow)







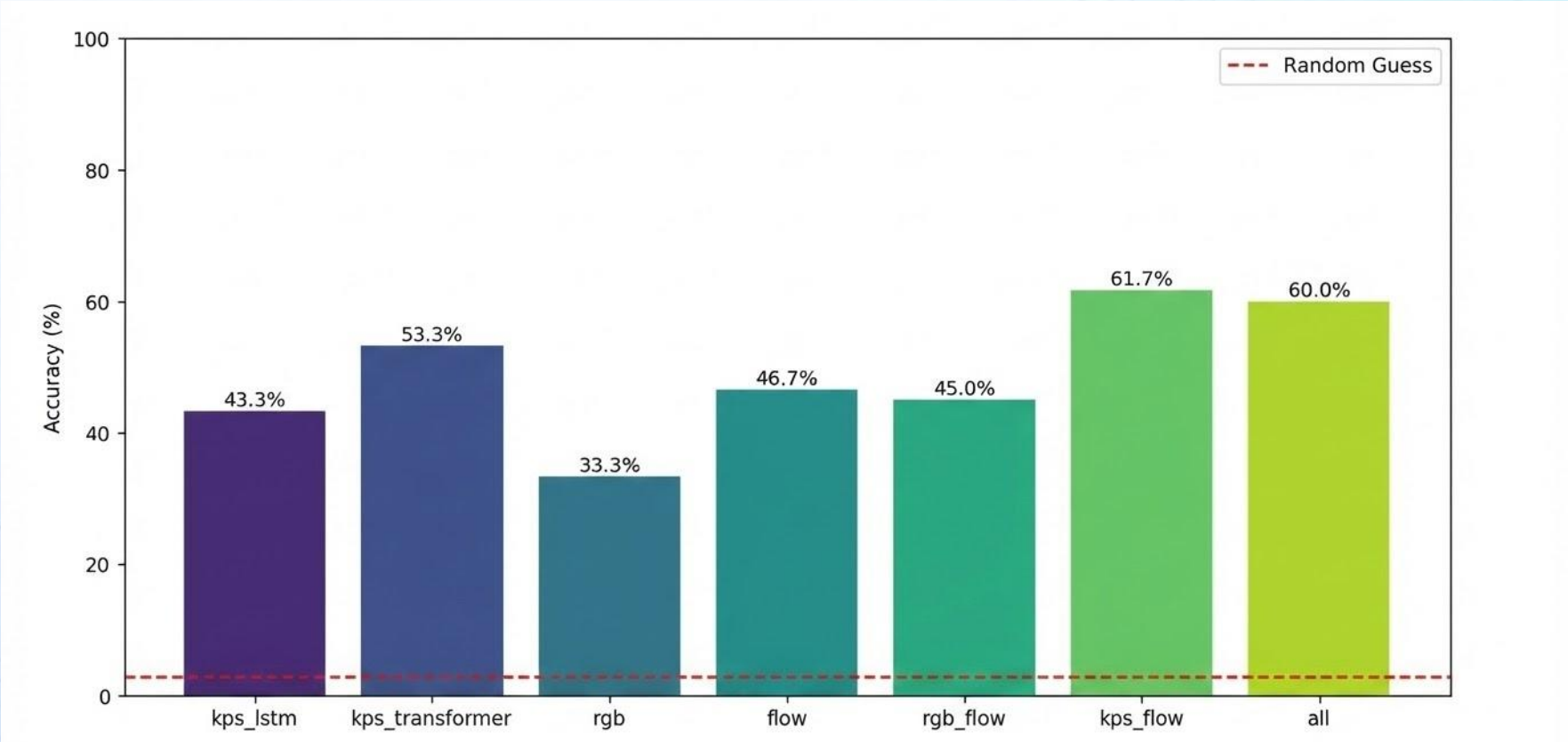


Triple Fusion (Key points, RGB, Flow)

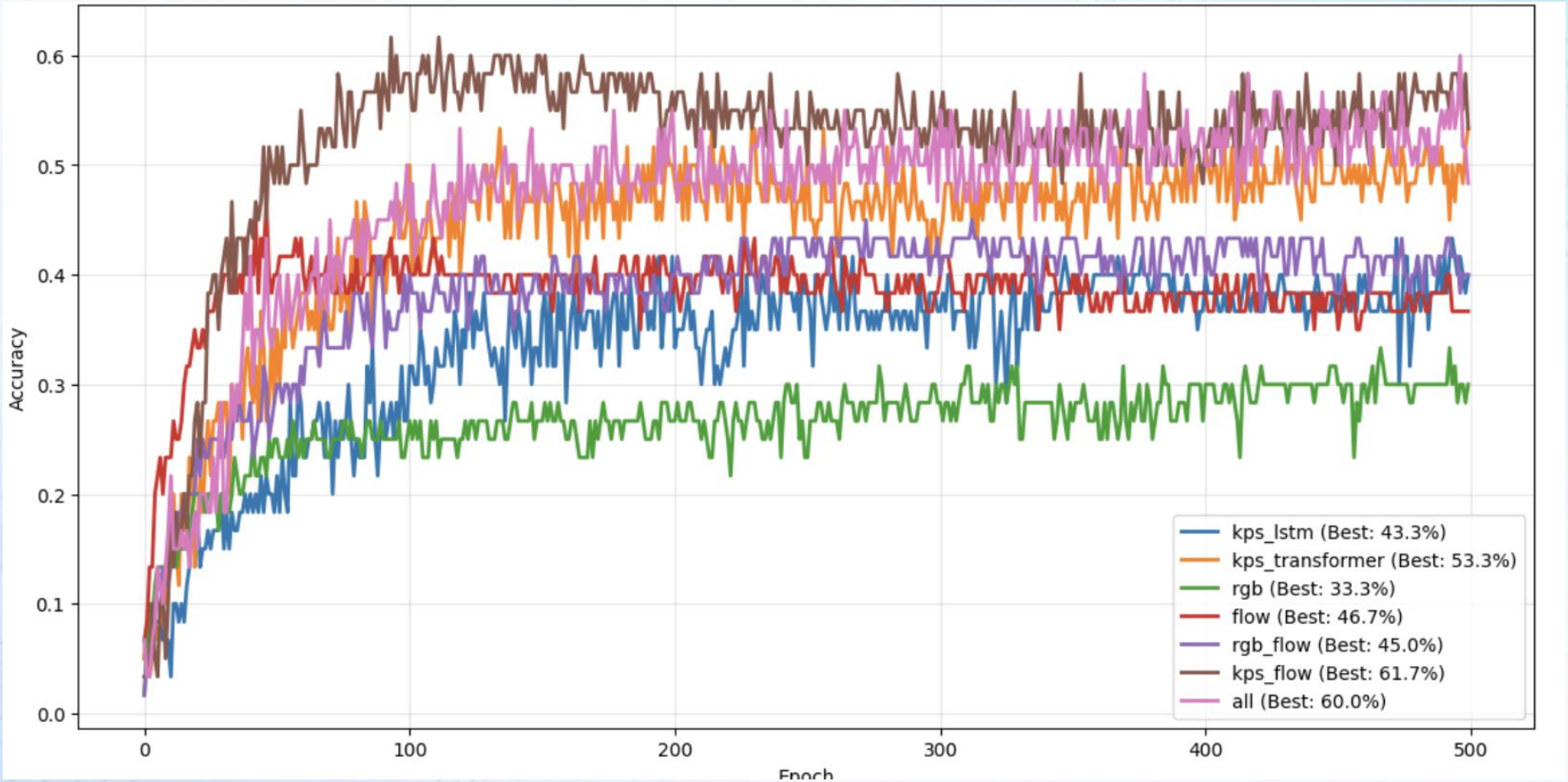
UNIVERSITI MALAYA

Evaluation

Evaluation Results - (Model Comparison - WLASL 35)



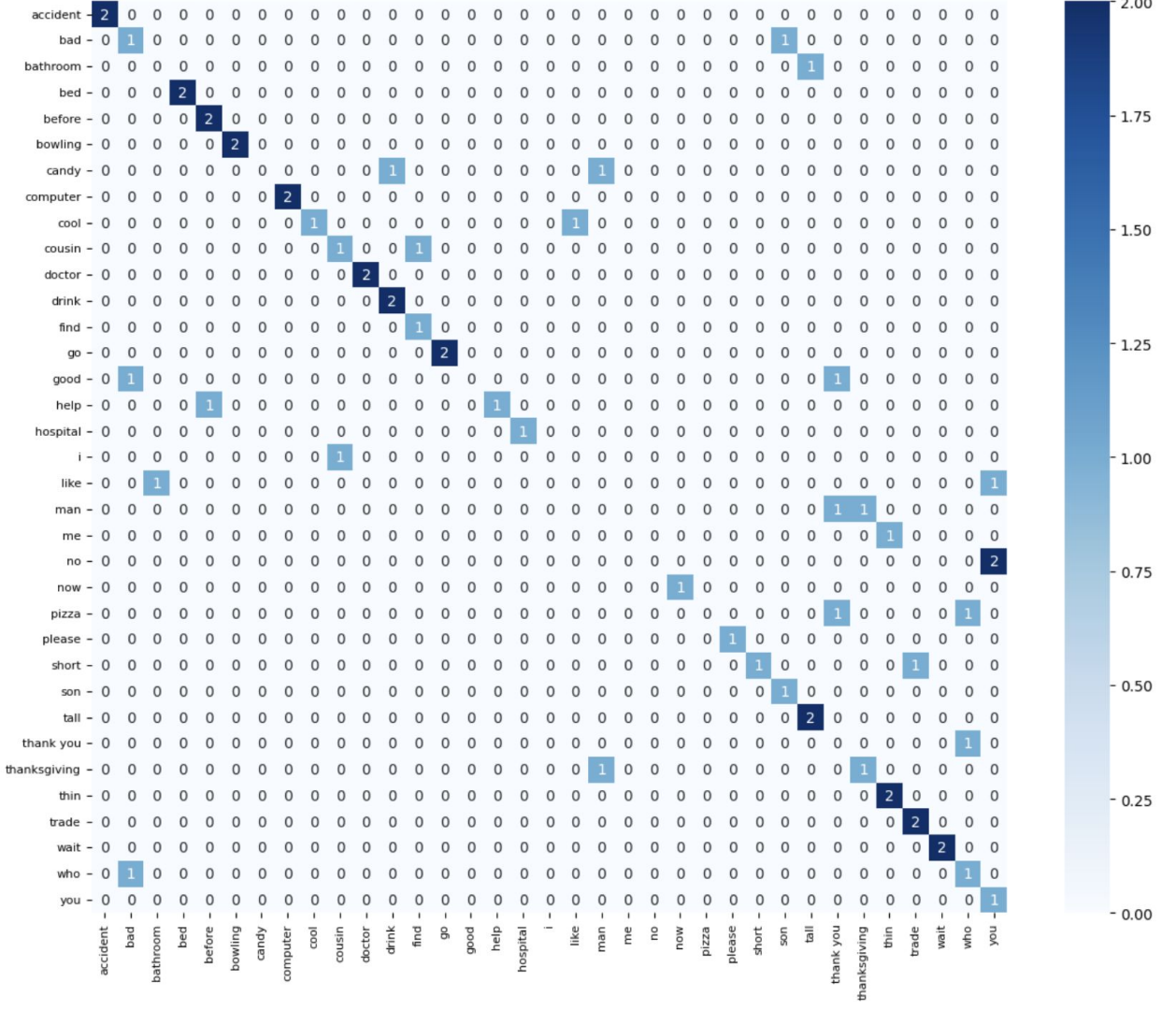
Validation Accuracy Curve



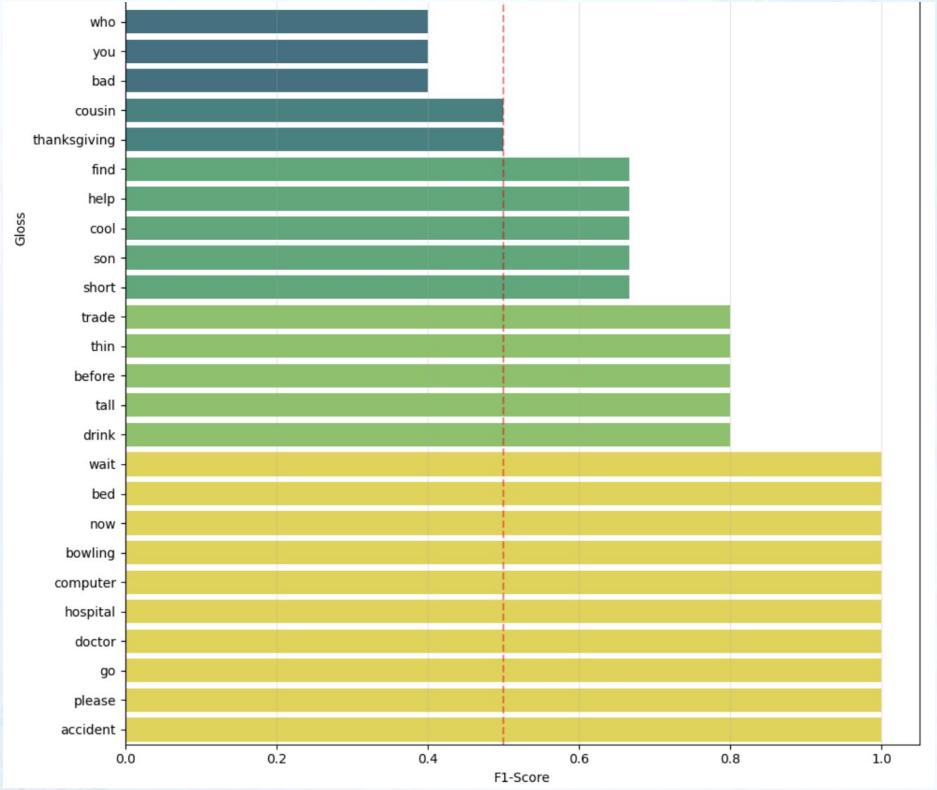
UNIVERSITI MALAYA

Best Model

KPS Flow Confusion Matrix



KPS Flow F1-Score



Model Performance Summary

Model	Accuracy	Precision	Recall	F1-Score
KPS LSTM	43.33%	35.97%	43.33%	37.44%
KPS Transformer	53.33%	49.72%	53.33%	48.83%
RGB	33.33%	31.50%	33.33%	30.17%
Flow	46.67%	41.06%	46.67%	40.68%
RGB Flow	45.00%	39.44%	45.00%	40.39%
KPS Flow	61.67%	57.08%	61.67%	57.22%
All	60.00%	51.94%	60.00%	54.56%

UNIVERSITI MALAYA

User Demo

Ethical AI: Bias & Explainability

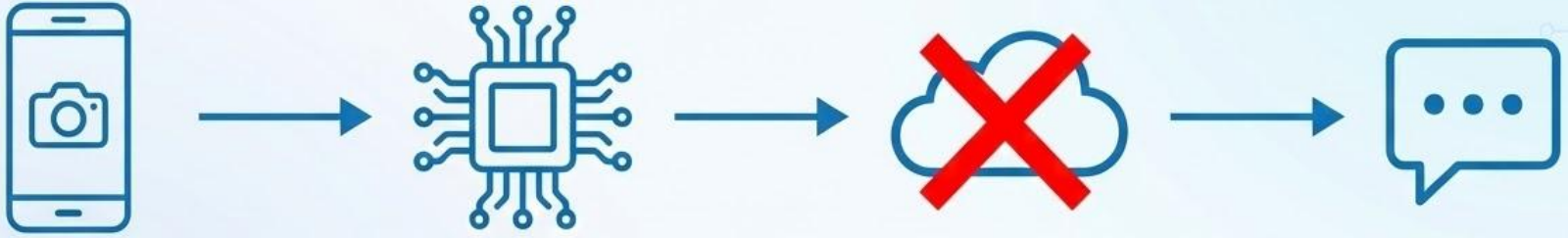
DATA BIAS & Representation Risk

The model is only as good as the WLASL dataset. If training data lacks diversity in signers (race, gender, lighting), the app may fail for underrepresented groups.

EXPLAINABILITY & The Trust Gap

Deep Learning models are 'Black Boxes'. In healthcare, users need to trust why a word was predicted. Lack of transparency increases risk during errors.

Privacy by Design: The Edge Paradigm



Real-time Video Capture

EDGE COMPUTING (On-Device Processing)

The AI model runs locally on the phone's hardware.

NO CLOUD TRANSMISSION

Video frames never leave the room.

Ephemeral Translation

Data is discarded immediately.

Long-Term Vision

Current Focus (Today)

Emergency Access & Basic Healthcare.

Future Stage 1

Education Integration: Real-time translation in mainstream classrooms, allowing Deaf students to participate fully without relying solely on scheduled interpreters.

Future Stage 2

Economic Empowerment: Breaking down employment barriers during interviews and daily workplace interactions, fostering independence.

UNIVERSITI MALAYA

Thank You

Full code available at: https://github.com/tyoppar01/sign-language-model
I3D feature extractor: https://github.com/hongjiaherng/video_features
Dataset: https://huggingface.co/datasets/jherng/wlasl_reduced