A LARGE LANGUAGE MODEL–ASSISTED VIETNAMESE SPEECH-TO-SIGN LANGUAGE TRANSLATION SYSTEM

Chi Hoang[1,2]

[1] Faculty of Information Science and Engineering,
University of Information Technology, Ho Chi Minh City, Vietnam
[2] Vietnam National University Ho Chi Minh City,
Ho Chi Minh City, Vietnam

WHY?

- Deaf communities in Vietnam have **limited access to spoken information** in education and daily services.
- Existing VSL systems often rely on **dictionary lookup** or **handcrafted animations**, leading to **unnatural and discontinuous motions**.
- Progress is limited by **scarce aligned Vietnamese Speech–Text–Sign data** and challenges in modeling **prosody and non-manual cues**.

WHAT?

We propose **Vi-Sign**, an **end-to-end** system that translates Vietnamese speech/text into **continuous VSL motions**:

- **Gloss-free** (does not require gloss supervision).
- **LLM-assisted**: uses an LLM to infer sentence-level structure, emphasis, and prosodic cues.
- **Prosody-aware control signals** to improve rhythm and continuity.
- A new aligned Speech–Text–Sign dataset (~**6,000 sentences**).

RESEARCH QUESTION

How can Vietnamese speech and text be translated into **continuous and expressive** VSL motions **without gloss supervision**?

EXPECTED RESULTS

- **Smoother**, **more stable** continuous sign motions.
- Better **prosody–gesture alignment**.
- Richer **non-manual expressions**.

CONTRIBUTIONS

- **Prosody-aware, gloss-free** VSL generation system
- A **new aligned** Vietnamese Speech–Text–Sign **dataset**
- A **reusable pipeline** for low-resource sign languages

OVERVIEW

Vi-Sign translates Vietnamese speech/text into continuous VSL motions by combining **LLM-based sentence-level reasoning** with **prosody-aware motion control**. The inferred control signals are refined to generate **smooth**, **stable**, and **expressive** sign movements.

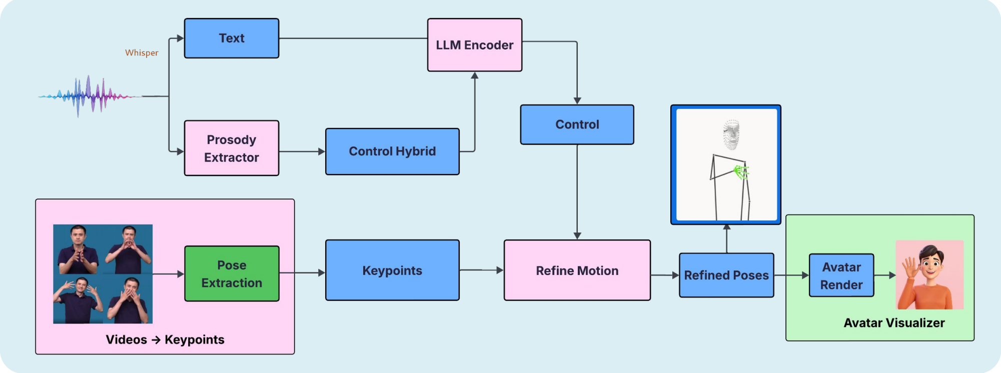

Fig. 1. Pipeline for Vietnamese Speech-to-Sign Translation.

DATA CONSTRUCTION

- **Aligned** Speech–Text–Sign **dataset.**
- **High-quality audio** for prosody modeling.
- **Keypoint-based** sign representation.
- **Multiple signers** to improve generalization.

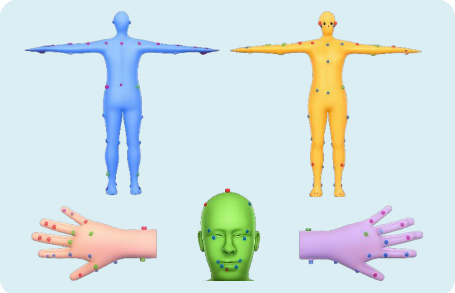

Fig. 2. Keypoint-based Sign Representation

LLM-ASSISTED CONTROL

- LLM for **sentence-level context.**
- Predicts **expressive control signals.**
- Adapts Vietnamese syntax to VSL.
- **No gloss supervision** required.

MOTION GENERATION

- Generates **continuous sign motion.**
- **Temporal refinement** for smoothness.
- **Geometric consistency** constraints.
- **Avatar-based visualization.**

CONCLUSION

- **Vi-Sign** is a prosody-aware, gloss-free Vietnamese Speech-to-Sign system.
- Using **LLM-based sentence-level reasoning**, Vi-Sign generates **smooth**, **stable**, and **expressive** VSL motions.
- The pipeline is reusable for **low-resource sign languages**.

Chi Hoang - University of Information Technology, Ho Chi Minh City, Vietnam
Email: chihqqc.20@grad.uit.edu.vn