VI-SIGN
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

WHAT?

We propose Vi-Sign, a system converting Vietnamese speech/text to continuous VSL motions.

- The system operates without gloss supervision.
- It integrates prosody-aware control signals assisted by a Large Language Model.
- A new Speech–Text–Sign dataset (~6,000 sentences) is constructed.

WHY?

- Deaf people in Vietnam face limited access to spoken information.
- Existing VSL systems rely on dictionary lookup or handcrafted animations.
- Current approaches lack rhythm, continuity, and expressive non-manual features.
- Aligned multimodal VSL data is scarce.

OVERVIEW

Vi-Sign is an end-to-end system that translates Vietnamese speech or text into continuous Vietnamese Sign Language (VSL) motions.

It employs an LLM as a high-level sentence-level contextual reasoning module to infer sentence structure, emphasis, and prosodic cues, which are converted into expressive control signals and refined into smooth, stable sign movements.

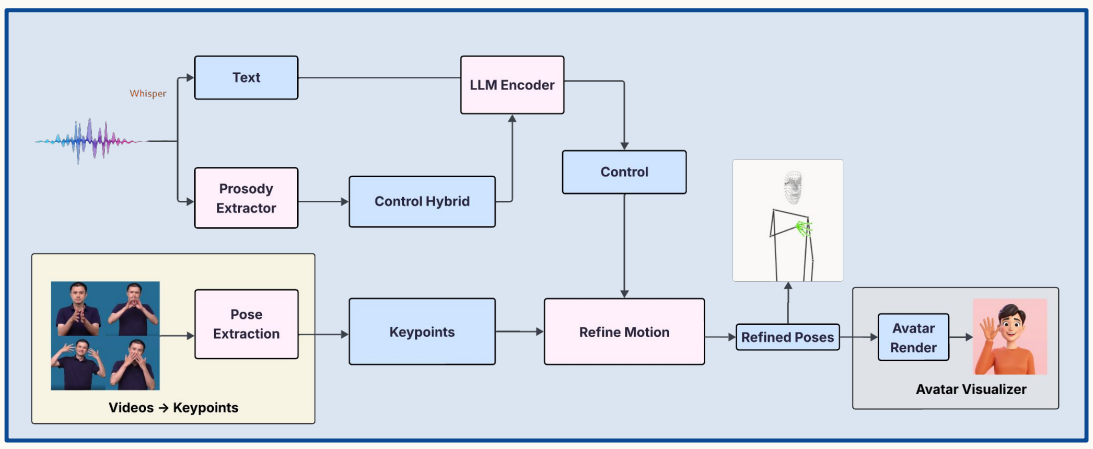

Fig. 1. Pipeline for Vietnamese Speech-to-Sign Translation.

DESCRIPTION

1. Data Construction

- Aligned Speech–Text–Sign dataset (~6,000 sentences)
- High-quality audio for prosody modeling
- Keypoint-based sign representation (hands, body, face)
- Multiple signers to improve generalization

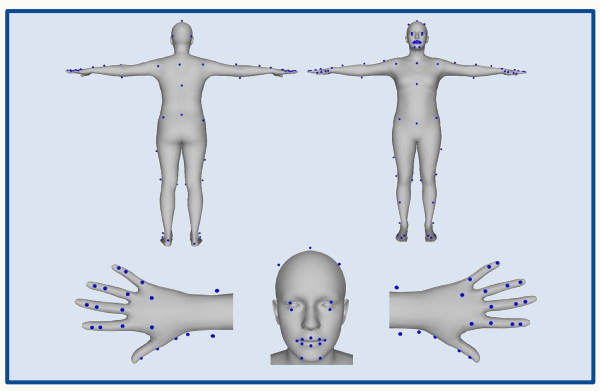

Fig. 2. Keypoint-based Sign Representation

2. LLM-assisted Control Prediction

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

3. Motion Generation & Refinement

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

EXPECTED RESULTS

- Smoother and more stable sign motions
- Improved prosody–gesture alignment
- Improved non-manual expressions (face, head, gaze)

CONTRIBUTIONS

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

Chi Hoang - University of Information Technology, Ho Chi Minh City, Vietnam
TEL : 0986098697 Email : chihqqc.20@grad.uit.edu.vn