# A Signer-Invariant Conformer and Multi-Scale Fusion Transformer for Continuous Sign Language Recognition

Md Rezwanul Haque[1 *], Md. Milon Islam[1 *], S M Taslim Uddin Raju[1], Fakhri Karray[1,2]
[1]University of Waterloo, [2]Mohammad bin Zayed University of Artificial Intelligence

ICCV OCT 19-23, 2025 HONOLULU HAWAII

## Introduction & Challenges

**Introduction**

- **Sign language**: a rich visual-gestural language for deaf and hard-of-hearing individuals.
- **Continuous Sign Language Recognition (CSLR)**: Converts sequences of sign gestures into textual representations.

**Key Challenges in CSLR**

- **Inter-Signer Variability**: Different people sign the same sentence in unique ways (speed, style, etc.).
- **Co-articulation**: The appearance of a sign changes based on the signs that come before and after it.
- **Generalization to New Sentences**: Models struggle to understand novel grammatical structures they have not seen during training.

## Our Solution: A Dual-Architecture Framework

**Signer-Invariant Conformer (for SI Task)**

- **Focus**: Learning robust, signer-agnostic representations.
- **Input**: Pose-based skeletal keypoints.

**Multi-Scale Fusion Transformer (for US Task)**

- **Focus**: Enhancing linguistic generalization to novel sentence structures.
- **Input**: Pose-based skeletal keypoints with multi-scale feature fusion.

## Signer-Invariant Conformer

- **Pose Estimator:** Extracts 2D landmarks (pose, hands, face).
- **Temporal Encoder:** Initial 1D convolutional layers capture short-range local correlations.
- **Conformer Blocks:** Stacked blocks integrating:
  - ✓ Multi-Head Self-Attention (MHSA) for global dependencies.
  - ✓ Convolution Module for local, fine-grained patterns.
- **Positional Encodings:** Provides sequence order information.
- **Classifier Head:** Generates sign gloss predictions.

## Multi-Scale Fusion Transformer

- **Pose Estimator:** Retrieves keypoint data.
- **Joint Attention Mechanism:** Dynamically weighs feature importance at the frame level to focus on salient gestural information.
- **Dual-Path Temporal Encoder:**
  - ✓ **Main Block:** Uses 1D convolutions for fine-grained, frame-level temporal dynamics.
  - ✓ **Auxiliary Block:** Uses max-pooling for downsampled representations, learning temporal features.
- **Transformer Encoder:** Models long-range dependencies and complex grammatical relationships from fused multi-scale features.
- **Classifier Head:** Generates US gloss predictions using CTC loss.

## System Architecture

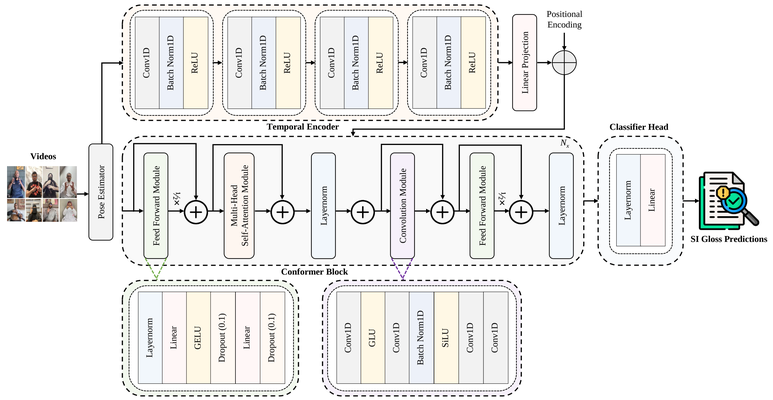


Figure 1. Signer-Invariant Conformer.

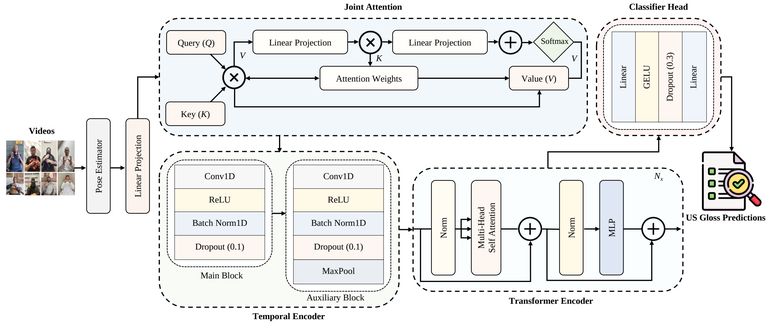


Figure 2. Multi-Scale Fusion Transformer.

## Dataset and Evaluation

- Large-scale, multi-scene Continuous Saudi Sign Language (CSSL) dataset.
- **Signer-Independent (SI) Split:** Evaluates generalization to new signers.
- Our primary metric is Word Error Rate (WER), where lower is better.

## Results and Analysis

| Methods | Dev | Test |
|---|---|---|
| VAC* [31] | 18.9 | 31.9 |
| SMKD* [19] | 18.5 | 35.1 |
| TLP* [21] | 19.0 | 32.0 |
| SEN* [23] | 19.1 | 36.4 |
| CorrNet* [22] | 18.8 | 31.9 |
| Swin-MSTP* [5] | 17.9 | 26.6 |
| SlowFastSign* [3] | 19.0 | 32.1 |
| LLM-SlowFast | 43.90 | 72.24 |
| LLaMA-Former | 21.83 | 51.21 |
| LLaMA-SlowFast | 30.13 | 46.98 |
| Mamba-Sign | 29.31 | 37.28 |
| Multi-Scale Fusion Transformer | 27.54 | 33.91 |
| BiLSTM | 17.02 | 26.08 |
| Sign-Conformer | 16.25 | 26.63 |
| CNN-BiLSTM | 14.54 | 22.62 |
| **Signer-Invariant Conformer** | **7.31** | **13.07** |

*results come from [7].

Table 1. Performance on the Isharah-1000 SI benchmark.

| Methods | Dev | Test |
|---|---|---|
| VAC* [31] | 57.0 | 49.6 |
| SMKD* [19] | 56.6 | 48.0 |
| TLP* [21] | 70.8 | 63.3 |
| SEN* [23] | 66.2 | 57.3 |
| CorrNet* [22] | 63.7 | 55.0 |
| Swin-MSTP* [5] | 73.5 | 66.1 |
| SlowFastSign* [3] | 65.5 | 56.2 |
| LLM-SlowFast | 93.07 | - |
| ST-GCN-Conformer | 91.80 | - |
| LLaMA-Former | 86.90 | - |
| DistilBERT-Former | 81.70 | - |
| BiLSTM | 79.93 | - |
| Sign-Conformer | 77.50 | - |
| CNN-BiLSTM | 74.96 | - |
| Signer-Invariant Conformer | 64.48 | - |
| Mamba-Sign | 59.51 | - |
| **Multi-Scale Fusion Transformer** | **55.08** | **47.78** |

*results come from [7].

Table 2. Performance on the Isharah-1000 US benchmark.

## Conclusion & Future Work

**Conclusion**

- Introduced a dual-architecture framework addressing SI and US CSLR distinct issues.
- **Signer-Invariant Conformer**: Achieves new SOTA for signer-agnostic feature learning (SI task).
- **Multi-Scale Fusion Transformer**: Achieves new SOTA for linguistic generalization (US task).
- Our models demonstrate the effectiveness of task-specific networks for CSLR, establishing new baselines.

**Future Directions:**

- Apply advanced encoders to Sign Language Translation (SLT).
- Investigate multi-modal fusion (RGB features like hand shape, face expressions) to enhance robustness against pose estimation errors.
- Develop a unified, multi-task architecture for both SI and US recognition within a single, efficient framework.

## Models and codes are publicly available

**Link**: https://github.com/rezwanh001/MSLRPose86K-CSLR-Isharah.

[1] Alyami et al., "Isharah: A large-scale multi-scene dataset for continuous sign language recognition," arXiv:2506.03615, 2025.
[2] Alyami & Luqman, "Swin-MSTP: Swin transformer with multi-scale temporal perception for continuous sign language recognition," Neurocomputing, 617:129015, 2025.