arXiv:2608.06407v1 [cs.CV] 3 Aug 2026

TransSLR: A Lightweight Transformer for Sign Language Recognition

Wanchi Lucia Yen[1] , **Samuel Johnny**[1] , **Victor Tolulope Olufemi**[1] , **Emmanuel Aaron**[1] , **Moise Busogi**[1]

[1]Carnegie Mellon University Africa

{wluciaye, sjohnny, volufemi, eaaron, mbusogi}@andrew.cmu.edu

Abstract

Automated Sign Language Recognition for underrepresented languages remains a largely unsolved problem. Central African Sign Language (CASL) exemplifies this gap: the only available benchmark, CASL-W60, has a best reported accuracy of 69.93%, and we show that the common heuristic of fine-tuning high-resource models fails to close it. This failure stems from two compounding factors: the limited scale of available CASL data and the significant lexical and visual domain gap between CASL and large-scale corpora such as WLASL, which renders pre-trained representations largely uninformative.

To address this, we propose **TransSLR**, a lightweight Temporal Transformer Encoder trained from scratch on 64-frame normalized pose sequences, with average pooling and a classification head. By operating on geometric keypoint representations rather than raw RGB, **TransSLR** achieves signer-independent generalization without relying on visual appearance. On the CASL-W60 benchmark, **TransSLR** establishes a new state-of-the-art accuracy of **80.39%**, surpassing the prior best by +10.46%. Beyond accuracy, our encoder-only design significantly reduces computational overhead, making deployment feasible in resource-constrained environments. We conduct extensive experiments on the CASL-W60 benchmark, comparing against RGB-based and multimodal baselines, and demonstrate that TransSLR achieves state-of-the-art performance.

Keywords: CASL, sign language recognition, domain gap, pose transformers, TransSLR, temporal attention, MediaPipe

1 Introduction

Sign languages are rich and fully expressive linguistic systems, yet progress in Sign Language Processing (SLP) remains highly uneven across regions. While languages such as American Sign Language (ASL), British Sign Language (BSL), and Chinese Sign Language (CSL) benefit from large-scale datasets and well-established benchmarks, many low-resource sign languages remain severely underrepresented. Central African Sign Language (CASL), used across the Central African region, is one such example, where limited data availability restricts the development of robust recognition systems and hinders accessibility to AI-driven communication tools.

Figure 1: Example of lexical divergence between ASL and CASL. Although both signs convey "sleep," they use distinct articulatory strategies, highlighting why knowledge learned from ASL does not transfer effectively to CASL.

Although ASL and CASL are both natural sign languages, they are not mutually interchangeable. They differ in vocabulary, articulation, and the linguistic strategies used to express the same concept. Figure 2 illustrates one such difference using the sign for "sleep". In ASL, the sign is produced with a downward hand-closing motion that iconically represents closing eyelids, whereas in CASL, the sign is expressed through a lateral head tilt that conveys the act of resting. This substantial difference in sign formation highlights the linguistic divergence between the two languages and explains why an ASL-pretrained recognition model achieves 0.00% zero-shot accuracy on the CASL-W60 dataset.

Sign Language Recognition (SLR) is generally divided into isolated sign language recognition (ISLR) and continuous sign language recognition(CSLR) settings. Recent advances in ISLR have been driven by large-scale visual datasets and powerful architectures leveraging RGB video, multimodal inputs, and pre-training strategies. However, these approaches rely heavily on data abundance and often

fail to generalize to low-resource settings with little corpus of data. In particular, the widely adopted assumption that models pre-trained on high-resource corpora (e.g., WLASL) can transfer effectively to unseen sign languages remains insufficiently validated.

In this work, we investigate the transferability of state-of-the-art sign language recognition (SLR) models to Central African Sign Language (CASL). Our experiments show that models pre-trained on high-resource sign language datasets fail to generalize in the zero-shot setting. We find that this failure is associated with substantial linguistic and kinematic differences between the source and target sign languages, which limit the effectiveness of direct cross-lingual transfer, even when the target language has been historically influenced by the source language.

To address this challenge, we propose TransSLR, a lightweight Transformer-based framework that emphasizes geometric representation learning over raw visual appearance. By operating on normalized skeletal pose sequences, our approach promotes invariance to signer-specific and environmental variations, making it particularly suitable for low-resource settings. Unlike RGB-dependent methods, TransSLR focuses on capturing the intrinsic motion dynamics of sign language.

We evaluate our method on the CASL-W60 benchmark, the only publicly available dataset for CASL. TransSLR significantly outperforms prior approaches and establishes a new state of the art in signer-independent recognition. These results highlight the effectiveness of pose-driven modeling for low-resource sign languages and provide a practical pathway toward scalable and accessible SLR systems.

Our main contributions are:

- We identify and empirically characterize the *overconfident ignorance* phenomenon in cross-lingual transfer for sign language recognition.
- We propose TransSLR, a Transformer-based ISLR framework designed for low-resource settings, leveraging normalized skeletal representations for improved generalization.
- We demonstrate state-of-the-art performance on the CASL-W60 dataset, highlighting the advantages of geometric modeling over appearance-based approaches in data-scarce environments.

2 Related Work

2.1 Sign Language Recognition

Sign Language Recognition (SLR) aims to interpret visual signing into linguistic representations and is commonly divided into isolated sign language recognition (ISLR) and continuous sign language recognition(CLSR). Early approaches relied on RGB video inputs, where convolutional neural networks (CNNs) such as I3D and S3D were used for spatial feature extraction, followed by recurrent neural networks (RNNs), including LSTMs and GRUs, for temporal modeling [Hu *et al.*, 2021a; Hu *et al.*, 2021b; Jiang *et al.*, 2021; Vaezi Joze and Koller, 2019; Zuo *et al.*, 2022; Gan *et al.*, 2023a].

More recent work has shifted toward efficient and privacy-preserving representations using skeletal pose estimation. Frameworks such as MediaPipe [Huynh *et al.*, 2025] and OpenPose [Cao *et al.*, 2019] enable the extraction of structured keypoints, which reduce redundancy while preserving motion dynamics. These representations have been effectively modeled using Graph Convolutional Networks (GCNs) to capture spatial dependencies [Yang *et al.*, 2026; Jiang *et al.*, 2021].

Transformer-based architectures have further improved temporal modeling by capturing long-range dependencies through self-attention [Gan *et al.*, 2023b]. SPOTER [Boháček and Hrúz, 2022] demonstrated the effectiveness of Transformers with spatial normalization, while more recent methods such as SNDA [Hassan *et al.*, 2025], enhance attention mechanisms for sign recognition. Additionally, end-to-end approaches such as AutoSign [Johnny *et al.*, 2025] explore direct pose-to-text translation for continuous sign language recognition, highlighting the growing importance of skeletal representations.

Large-scale datasets such as WLASL [Li *et al.*, 2020] and MS-ASL [Vaezi Joze and Koller, 2019] have enabled the development of powerful spatiotemporal models, including VideoMAE [Kumar and others, 2025]. However, these datasets primarily focus on American Sign Language, limiting their applicability to other linguistic and regional contexts.

2.2 Isolated Sign Language Recognition

Isolated Sign Language Recognition (ISLR) focuses on classifying individual signs and serves as a fundamental component of broader SLR systems [Sarhan and Frintrop, 2023; Laines *et al.*, 2023]. Traditional ISLR approaches combine CNN-based feature extraction with RNN-based temporal modeling, which can struggle with sequence variability and temporal alignment [Tao *et al.*, 2024].

Transformer-based methods have emerged as a strong alternative due to their flexibility in handling variable-length sequences. However, their effectiveness often depends on preprocessing strategies such as temporal sampling and spatial normalization.

For low-resource sign languages, the CASL-W60 dataset [Lucky *et al.*, 2025] provides the first benchmark for isolated Central African Sign Language recognition. Existing methods have demonstrated the feasibility of pose-based recognition on CASL, yet there remains limited understanding of how to effectively exploit kinematic representations for robust recognition in data-scarce settings. This work builds on these efforts by investigating whether a Transformer-based architecture operating on normalized kinematic pose sequences can improve recognition performance while maintaining a simple and efficient design.

3 Methodology

3.1 Problem Formulation

Given an input video sequence of a signer $\mathbf{F} = \{f_i\}_{i=1}^{N}$ consisting of N frames, the goal of isolated sign language recognition (ISLR) is to learn a mapping function $\mathcal{G}$ that predicts

the corresponding sign label $y \in \mathcal{Y}$, where $\mathcal{Y}$ denotes the predefined vocabulary of sign classes.

Formally, the task can be expressed as:

$$y = \mathcal{G}(\mathbf{F})$$

where $\mathcal{G}$ represents the end-to-end model that encodes spatio-temporal information from the input video and outputs the predicted sign.

3.2 Data Preprocessing

To ensure spatiotemporal consistency across the video samples, we developed an automated pipeline that transforms raw RGB footage into standardized kinematic tensors. We addressed temporal variance by implementing a uniform sampling strategy, compressing or expanding each video into a fixed 64-frame window via linear interpolation. Spatial features were extracted using the MediaPipe Holistic framework, yielding a $(64, 225)$ matrix representing 75 skeletal keypoints. To neutralize signer-positioning bias and biometric variance, we applied a dual-normalization protocol. First, we enforced spatial centering by setting the nasal landmark $P_{nose}(x, y, z)$ as the coordinate origin for all subsequent points:

$$P'_i = P_i - P_{nose} \tag{1}$$

Secondly, we implemented a Gap-Filling logic to resolve occlusion-based data loss. Missing manual landmarks (estimated at 25% of the raw extraction) were restored using temporal linear interpolation, governed by forward and backward fill heuristics:

$$f(t) = f(t_{prev}) + (t - t_{prev})\frac{f(t_{next}) - f(t_{prev})}{t_{next} - t_{prev}} \tag{2}$$

This process yielded smooth, continuous "motion ribbons", effectively isolating the linguistic intent from raw visual noise and providing a scale-invariant geometric foundation for the TransSLR encoder.

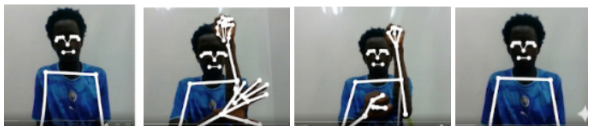


Figure 2: The RGB and 64-Frame Motion Ribbon. A horizontal time-strip showing the centralized progression of a sign from onset to offset.

3.3 TransSLR Architecture

Overview

TransSLR is a lightweight, Transformer-based encoder designed for Isolated Sign Language Recognition (ISLR) under low-resource constraints. Given a normalized pose sequence as input, TransSLR learns temporal dependencies across keypoint trajectories and produces a class probability distribution over 60 CASL signs where the model outputs the top 1 as the result.

As illustrated in Figure 3, the pipeline consists of four sequential components: a linear input projection, sinusoidal positional encoding, a Transformer encoder stack, and a classification head.

Unlike sequence-to-sequence tasks such as Continuous Sign Language Recognition (CSLR) or Sign Language Translation (SLT), ISLR is a fixed-input, single-label classification problem: given a bounded gesture sequence, the model produces exactly one class prediction. This eliminates the need for an autoregressive decoder, showing that TransSLR, a Transformer encoder-only model paired with a Global Average Pooling classification head, is both necessary and sufficient for the task.

Input Representation

Each input sample of poses is represented as a kinematic tensor $\mathbf{X} \in \mathbb{R}^{T \times D}$, where $T = 64$ denotes the number of temporal frames and $D = 225$ is the dimensionality of the flattened pose keypoints per frame. To ensure spatial invariance across signers, all sequences are normalized via mid-hip centering and shoulder-distance scaling prior to being fed into the model. This geometric normalization is a deliberate design choice: by discarding appearance information, thereby Improving generalization, as well as effective training and data utilization while retaining only skeletal structure, TransSLR is encouraged to learn signer-independent motion patterns rather than overfit to subject-specific visual cues.

Linear Projection.

Since the raw keypoint dimension $D = 225$ does not directly align with the Transformer latent dimension d_{model}, a learned linear layer projects each frame-level feature vector into the model embedding space, effectively treating each frame as a positional token for the subsequent self-attention computation.

Positional Encoding

Since Transformer encoders are inherently permutation-invariant with respect to input tokens, self-attention computes pairwise interactions without an explicit notion of temporal order. In isolated sign language recognition (ISLR), however, the temporal structure of hand articulation is semantically important, as different phases of a gesture contribute differently to class identity. Therefore, explicit positional information is required to model the temporal evolution of each sign.

We incorporate fixed sinusoidal positional encodings, which are added to frame-level embeddings prior to the first encoder layer. These encodings provide a deterministic and continuous representation of temporal position, enabling the model to distinguish between frames at different time steps while preserving smooth interpolation across sequence positions. Unlike learned positional embeddings, the sinusoidal formulation does not introduce additional parameters and generalizes more robustly to variable-length sequences. This is important in ISLR, where sign durations vary significantly across instances and signers, requiring the model to maintain consistent temporal representations under variable sequence lengths.

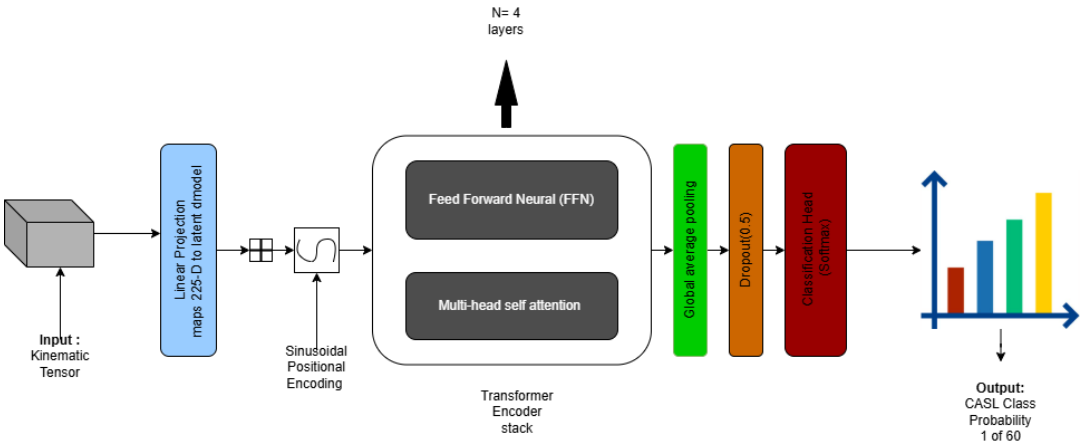


Figure 3: TransSLR Architectural Pipeline; The model processes (64, 225) kinematic tensors through Linear Projection and Positional Encoding, followed by a Multi-Head Self-Attention stack. Final classification is achieved via Global Average Pooling and a 60-class Softmax head, optimized for low-resource regional SLR.

Transformer Encoder Stack

The projected sequence is processed through $N = 4$ stacked Transformer encoder layers. Each layer applies Multi-Head Self-Attention (MHSA) followed by a position-wise Feed-Forward Network (FFN), with residual connections and layer normalization at each sublayer. By attending over all 64 temporal positions simultaneously, the encoder captures both local motion dynamics and long-range temporal dependencies across the signing sequence without recurrence.

Classification Head

Following the encoder stack, Global Average Pooling (GAP) aggregates the temporal token representations into a single fixed-size feature vector. A Dropout layer ($p = 0.5$) is applied for regularization, after which a linear layer with Softmax activation projects the feature vector to a 60-class output label. The full model is trained end-to-end using cross-entropy loss.

Training Objective

We optimize TransSLR using the standard Cross-Entropy loss, which is widely adopted for supervised multi-class classification. Given the predicted class probabilities $\hat{\mathbf{y}} \in \mathbb{R}^D$, where $D = 60$, and the corresponding one-hot ground-truth label $\mathbf{y}$, the loss is defined as

$$\mathcal{L}_{CE} = -\sum_{c=1}^{D} \mathbf{y}_c \log(\hat{\mathbf{y}}_c) \tag{3}$$

Minimizing this objective encourages discriminative feature learning by maximizing the likelihood of the correct sign class for each input sequence.

4 Experimental Setup

4.1 Dataset

Our experiments are conducted on the CASL-W60 dataset [Lucky *et al.*, 2025], the first and only publicly available benchmark for isolated Central African Sign Language recognition as of the time of this research. CASL-W60 contains 5,889 samples spanning 60 isolated signs from 19 unique signers, partitioned into 3,667 training, 679 validation, and 1,543 test samples. All splits are signer-independent, meaning no signer identity appears across more than one partition, making this a strict test of cross-signer generalization.

4.2 Implementation Details

All experiments are conducted on a single NVIDIA A100 GPU using PyTorch. TransSLR is trained for 60 epochs with a batch size of 16. Input pose sequences are represented as 225-dimensional keypoint vectors across 64 frames. We optimize with AdamW [Loshchilov and Hutter, 2019] with a learning rate of 1×10^{-4} and weight decay of 0.01, and apply a cosine annealing schedule ($T_{\text{max}} = 60$) to ensure smooth convergence in the later stages of training.

4.3 Evaluation Protocol

We evaluate model performance using three metrics: top-1 accuracy, top-5 accuracy, and Classification Error Rate (CER).

Top-1 accuracy This measures the proportion of test samples for which the highest-confidence prediction matches the ground-truth sign class.

Top-5 accuracy measures whether the correct class appears among the five highest-confidence predictions.

Classification Error Rate We report CER alongside top-1 accuracy to express model error in absolute terms and to facilitate direct comparison with the CASL-W60 baseline [Lucky *et al.*, 2025], which frames performance as an error rate. For an isolated recognition benchmark where each sample corresponds to a single lexical token, CER is defined as:

$$\text{CER} = 1 - \text{Acc}_{\text{top-1}}, \tag{4}$$

The direct complement of top-1 accuracy under closed-set classification. Unlike Word Error Rate, a sequence-level metric defined via Levenshtein edit distance over multi-token utterances, CER is the appropriate error measure for isolated

sign recognition, where no insertions, deletions, or substitutions are possible.

All metrics are computed on the held-out test set of CASL-W60, ensuring fair and reproducible comparison across methods.

5 Results

5.1 Fine-Tuning Evaluation and Domain Gap

Table 1 presents the zero-shot and fine-tuned performance of VideoMAE on CASL-W60. Under zero-shot evaluation, VideoMAE achieves 0.00% accuracy, producing a CER of 100% a complete failure that directly confirms the absence of any transferable visual representation between WLASL and CASL.

Fine-tuning on the CASL-W60 training split recovers performance to 57.55%, yet this remains approximately 12 points below the prior state-of-the-art [Lucky *et al.*, 2025] and over 22 points below our proposed TransSLR. We attribute this persistent gap to a representational mismatch: VideoMAE encodes spatiotemporal appearance features that do not transfer to CASL-W60, where lexical distinctions are carried primarily by subtle kinematic differences, handshape transitions, movement velocity, and spatial trajectory rather than by signer appearance. At the scale of CASL-W60, the available target-domain signal is insufficient to re-orient the model's inductive priors toward the kinematic structure that isolated sign discrimination requires.

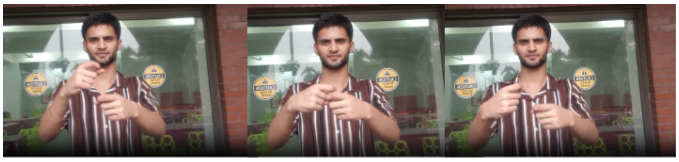
5.2 Comparison with State-of-the-Art Models

The TransSLR architecture achieves state-of-the-art performance on the CASL-W60 benchmark (Table 1) with a test accuracy of 80.39% and a classification error rate of 19.61% on the signer-independent test split (862 samples). This corresponds to a 10.86% improvement in Top-1 accuracy over the strongest prior baseline. As shown in Table 1, TransSLR consistently outperforms all prior pose-based approaches, including recurrent, convolutional, and transformer-based models. In particular, it surpasses the previous best pose-based transformer baseline (76.13%) by 4.26% in Top-1 accuracy.

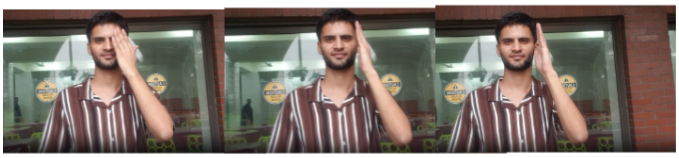
In addition to Top-1 performance, TransSLR achieves a Top-5 accuracy of 91.07%, indicating strong class-level ranking consistency under ambiguous predictions. The model also reduces classification error rate to 19.61%, improving over all prior methods across modalities, including RGB-based and hybrid RGB+pose architectures. Overall, the results demonstrate that the proposed encoder-only pose-based transformer provides a strong trade-off between accuracy and computational efficiency compared to more complex multimodal or RGB-heavy architectures.

5.3 Modality Superiority: Geometry vs. Visual Texture

The results in table 2 provide a compelling, albeit counter-intuitive, scientific insight: geometric pose-only modeling significantly outperforms multimodal RGB-pose fusion for signer-independent CASL recognition. While traditional paradigms suggest that adding RGB context should improve accuracy by providing visual texture, our results show a catastrophic representation collapse in the multimodal MViT V2 + Bi-LSTM baseline, which achieved a Recall@1 of only 29.90%.



(a) Correctly classified as G004('Hard')



(b) Incorrectly classified 'Blue' as G018("Don't know")

Figure 4: Qualitative results of the **TransSLR** model

This performance delta is primarily attributed to biometric noise overfitting. Raw RGB pixels allow the model to "cheat" during the optimization phase by memorizing subject-specific artifacts such as clothing patterns, skin tone, and localized lighting conditions found in the training set. When the architecture is subjected to the novel variance of unseen signers, these visual features become mathematical liabilities rather than assets.

In contrast, TransSLR leverages purely kinematic trajectories that have been subjected to nasal-centering and shoulder-scaling. This geometric abstraction effectively strips away the signer's identity, forcing the Transformer Encoder to learn scale-invariant motion primitives. The superiority of this approach is most evident in the retrieval metrics; TransSLR achieved a Recall@10 of 93.62% and a perfect Median Rank of 1.0, compared to a Median Rank of 4.0 for the baseline.

5.4 Linguistic Failure Analysis and Feature Space Collapse

Despite high global accuracy, a granular evaluation of the precision-recall matrix reveals localized "Linguistic Dead Zones." Specifically, five classes (30, 35, 51, 57, and 59) returned a 0.00% F1-score. Rather than indicating random noise, this failure pattern highlights a phenomenon of "High-Frequency Motion Overlap." [Ueda *et al.*, 2012] This is evidenced by examining the corresponding false-positive sinks in the model's latent space. For example, Class 53 achieved a perfect recall of 1.0000 but a severely degraded precision of 0.3333. This indicates an over-prediction bias, where the model maps subtle or ambiguous regional gestures into these dominant feature clusters. When faced with minimal pairs, signs that share identical manual trajectories and differ only in non-manual markers, the pose-only spatial encoder defaults to the statistically dominant cluster. These findings suggest that while skeletal normalization resolves signer-positioning bias, future iterations will necessitate fa-

Method	Modality	Top-1 (%) ↑	Top-5 (%) ↑	CER (%) ↓
VideoMAE [Kumar and others, 2025] (Zero-Shot)	RGB	0.00	–	100.00
VideoMAE [Kumar and others, 2025] (Fine-Tuned)	RGB	57.55	–	42.45
SLT [Lucky *et al.*, 2025]	Pose	69.93	–	30.07
Bi-GRU + Attention	Pose	70.34	–	29.66
Transformer (with MHSA + Positional Encoding)	Pose	70.24	–	29.76
Enhanced Transformer (temporal + pooling)	Pose	73.14	–	26.86
Transformer + MLP Head	Pose	75.82	–	24.18
Transformer Encoder (Final Baseline)	Pose	**76.13**	–	**23.87**
MViT V2 + Bi-LSTM	RGB+Pose	74.23	–	25.77
TransSLR (Ours)	Pose	**80.39**	**91.07**	**19.61**

Table 1: Results on CASL-W60. Top-1 denotes standard classification accuracy, while Top-5 measures whether the correct label appears within the top five predictions. ↑ indicates higher is better and ↓ lower is better. Top-5 is not reported for prior work.

Mod.	Model	R@1	R@5	R@10	MedR
Mult.	MViT-BiLSTM	29.90%	57.00%	73.78%	4.0
Pose	**TransSLR**	**80.39%**	**91.07%**	**93.62%**	**1.0**

Table 2: Evaluation of retrieval performance on unseen signers using Recall@K and Median Rank. Results show that TransSLR consistently improves over the multimodal baseline across all evaluation metrics.

cial meshes to disambiguate overlapping signs.

5.5 Qualitative Analysis of TransSLR

Figure 4 illustrates an example of correct and incorrect predictions from TransSLR on the CASL-W60 test set. In Figure 4(a), the model correctly classifies the sign (*Hard*), demonstrating its ability to capture discriminative hand shape and motion trajectories from normalized pose sequences. Figure 4(b) shows a failure case where *Blue* is misclassified as (*Don't know*), suggesting that visually similar kinematic trajectories between certain sign pairs remain a challenge under the limited training data available in CASL-W60. This highlights the need for expanded vocabulary annotation as a priority for future data curation efforts.

5.6 Latent Manifold Integrity and Signer Invariance

The t-SNE projection of the TransSLR latent representations (Figure 5) shows partially separable structure across a subset of sign classes, with 15% of the vocabulary achieving near-perfect classification performance (F1 = 1.00). This suggests that the Temporal Transformer Encoder learns embeddings that capture discriminative motion patterns for certain sign categories. However, other classes exhibit overlapping regions in the embedding space, indicating residual ambiguity.

We observe failure modes in a small number of classes that are visually similar in pose trajectory but differ in non-manual components such as facial expressions. Since the current model operates on pose-only inputs, these results suggest that incorporating facial or non-manual features may improve discrimination in these cases.

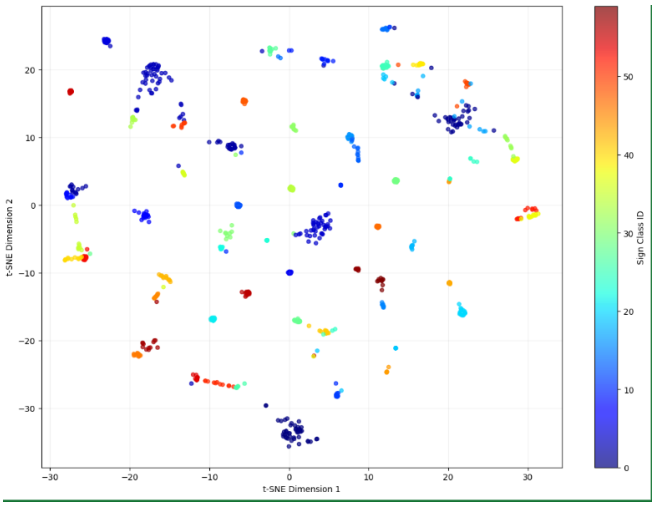


Figure 5: t-Distributed Stochastic Neighbor Embedding (t-SNE) visualization of the TransSLR latent manifold. The plot represents 862 unseen signer samples projected from the penultimate layer. The high inter-class separability validates the model's ability to map regional motion primitives into a structured, signer-invariant semantic space.

Method	Params (M) ↓	FLOPs (G) ↓
VideoMAE (Fine-Tuned)	65.021	101.85
Bi-GRU + Attention	9.95	1.40
MViT V2 + Bi-LSTM	47.47	7.17
TransSLR (Ours)	**8.67**	**0.28**

Table 3: Comparison of computational complexity across evaluated models. Parameter count is computed over all trainable parameters, while FLOPs are measured for processing a single 64-frame pose sequence. ↓ indicates lower is better.

5.7 Efficiency Analysis

We evaluate the computational efficiency of TransSLR by comparing its model size and inference complexity against representative RGB-based, pose-based, and multimodal baselines. As shown in Table 3, TransSLR has the smallest model size, requiring only 8.67M trainable parameters, while also exhibiting the lowest computational cost at 0.277 GFLOPs per 64-frame pose sequence. In comparison, VideoMAE and MViT V2 + Bi-LSTM require substantially larger models (65.02M and 47.47M parameters) and higher computational cost (101.85 and 7.17 GFLOPs, respectively). These results show that the proposed pose-based Transformer achieves superior retrieval performance while maintaining both a compact model size and low inference complexity.

Conclusion and Future Work

We presented TransSLR, a lightweight Transformer encoder for Isolated Sign Language Recognition (ISLR) that operates exclusively on normalized pose sequences. Through extensive experiments on the CASL-W60 benchmark, we demonstrated that a geometry-driven, encoder-only architecture trained from scratch outperforms fine-tuned high-resource RGB models, establishing a new state-of-the-art accuracy of 80.39% under signer-independent evaluation a 10.46 percentage point improvement over the prior best.

Central to our approach is the insight that spatial normalization via mid-hip centering and shoulder-distance scaling effectively decouples linguistic motion from signer-specific biometric variance, providing a scale-invariant geometric foundation that generalizes across unseen signers without requiring large-scale data. Collectively, our results confirm that domain-specific regional models are both linguistically necessary and computationally practical for low-resource African sign languages and that pose-based geometric modeling offers a viable and deployable path forward for underrepresented signing communities.

This work opens several avenues for future research. A key priority is the development of larger annotated corpora for Central African Sign Language. While CASL-W60 establishes an important benchmark for isolated sign recognition, its limited vocabulary constrains progress toward practical sign language understanding. Developing large-scale Continuous Sign Language Recognition (CSLR) datasets with richer vocabularies, longer signing sequences, and greater signer diversity will be essential for advancing research in this direction.

Given sufficiently large continuous sign language datasets, extending TransSLR from isolated to continuous sign language recognition represents a natural next step. Such an extension will require modeling unsegmented sign streams, learning long-range temporal dependencies, and handling variable-length output sequences, potentially through techniques such as Connectionist Temporal Classification (CTC) or attention-based sequence decoding. Another promising direction is the integration of self-supervised representation learning and large-scale pre-training to reduce reliance on labeled data. Learning transferable kinematic representations in this manner could improve data efficiency and facilitate the development of recognition systems for other low-resource African sign languages beyond CASL.

References

[Boháček and Hrúz, 2022] Matyáš Boháček and Marek Hrúz. Sign pose-based transformer for word-level sign language recognition. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 182–191, 2022.

[Cao *et al.*, 2019] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[Gan *et al.*, 2023a] Shiwei Gan, Yafeng Yin, Zhiwei Jiang, Kang Xia, Lei Xie, and Sanglu Lu. Contrastive learning for sign language recognition and translation. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, IJCAI '23, 2023.

[Gan *et al.*, 2023b] Shiwei Gan, Yafeng Yin, Zhiwei Jiang, Kang Xia, Lei Xie, and Sanglu Lu. Contrastive learning for sign language recognition and translation. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI)*, pages –, 2023.

[Hassan *et al.*, 2025] Esraa Hassan, Mahmoud Y Shams, Tarek Abd El-Hafeez, and Marwa Elseddik. A novel model for expanding horizons in sign language recognition. *Scientific Reports*, 15(1):24358, 2025.

[Hu *et al.*, 2021a] Hezhen Hu, Weichao Zhao, Wengang Zhou, Yuechen Wang, and Houqiang Li. Signbert: Pre-training of hand-model-aware representation for sign language recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11087–11096, 2021.

[Hu *et al.*, 2021b] Hezhen Hu, Wengang Zhou, and Houqiang Li. Hand-model-aware sign language recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 35, pages 1558–1566, 2021.

[Huynh *et al.*, 2025] Thai Anh Ngoc Huynh, Hoang Nguyen Pham, and Ngoc Tho Thi Nguyen. Motion recognition control by mediapipe holistic. In *Proceedings of the 1st International Conference on Sustainability and Emerging Technologies for Smart Manufacturing: SETSM 2024, 27–28 April, Hanoi, Vietnam*, page 291. Springer Nature, 2025.

[Jiang *et al.*, 2021] Songyao Jiang, Bin Sun, Lichen Wang, Yue Bai, Kunpeng Li, and Yun Fu. Skeleton-aware multi-modal sign language recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3413–3423, 2021.

[Johnny *et al.*, 2025] Samuel Ebimobowei Johnny, Blessed Guda, Andrew Blayama Stephen, and Assane Gueye. Autosign: Direct pose-to-text translation for continuous sign language recognition. In *Proceedings of the IEEE/CVF*

International Conference on Computer Vision Workshops (ICCVW), 2025.

[Kumar and others, 2025] Ashish Kumar et al. Video vision transformers for word-level sign language recognition. *arXiv preprint arXiv:2504.07792*, 2025.

[Laines *et al.*, 2023] David Laines, Gissella Bejarano, Miguel Gonzalez-Mendoza, and Gilberto Ochoa-Ruiz. Isolated sign language recognition based on tree structure skeleton images, 2023.

[Li *et al.*, 2020] Dongxu Li, Cristian Rodriguez, Xin Li, Yiqun Li, and Nigel H Jago. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. *Winter Conference on Applications of Computer Vision*, pages 1451–1460, 2020.

[Loshchilov and Hutter, 2019] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

[Lucky *et al.*, 2025] Mwaka Lucky, Njayou Youssouf, Hasan Mahmud, and Md Kamrul Hasan. Casl-w60: A word-level dataset for central african sign language recognition. *Data in Brief*, 61:111790, jun 2025.

[Sarhan and Frintrop, 2023] Noha A. Sarhan and Simone Frintrop. Unraveling a decade: A comprehensive survey on isolated sign language recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 3202–3211, 2023.

[Tao *et al.*, 2024] Tangfei Tao, Yizhe Zhao, Tianyu Liu, and Jieli Zhu. Sign language recognition: A comprehensive review of traditional and deep learning approaches, datasets, and challenges. *IEEE Access*, 12:75034–75060, 2024.

[Ueda *et al.*, 2012] Takaaki Ueda, Kenta Fujii, Shigeki Hirobayashi, Toshio Yoshizawa, and Tadanobu Misawa. Motion analysis using 3d high-resolution frequency analysis. *IEEE Transactions on Image Processing*, 22(8):2946–2959, 2012.

[Vaezi Joze and Koller, 2019] Hamid Reza Vaezi Joze and Oscar Koller. Ms-asl: A large-scale data set and benchmark for understanding american sign language. *British Machine Vision Conference (BMVC)*, sep 2019.

[Yang *et al.*, 2026] Xi Yang, Shaoyi Li, Saisai Niu, and Xiaokui Yue. Graph network learning for human skeleton modeling: a survey. *Artificial Intelligence Review*, 59(1):31, 2026.

[Zuo *et al.*, 2022] Ronglai Zuo, Fangyun Wei, and Brian Mak. Natural language-assisted sign language recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.