

A Hybrid Kinematic GCN and Elastic Transformer for Efficient Indian Sign Language Recognition

Pushp Garg · Himangshu Sarma · Mrinmoy Ghorai

Abstract Sign Language Recognition (SLR) presents complex challenges due to variable signing speeds and spatial intricacies. Specifically, Indian Sign Language (ISL) requires precise modeling of its inherently bimanual nature and is bottlenecked by the scarcity of large-vocabulary benchmarks and the inability of standard models to handle irregular signing velocities. To address this, a structurally-aware hybrid framework for efficient, gloss ISL recognition is proposed. To minimize computational redundancy, an entropy-driven SVD sampler that dynamically extracts informative keyframes is used. Because this subsampling creates irregular physical time gaps, the proposed architecture employs a Continuous Dual Positional Encoding paired with a distance-aware Elastic Transformer to preserve natural signing rhythm. Spatially, a Kinetically-Conditioned Adaptive GCN explicitly models bimanual geometry using unified bone-joint features. To validate our framework, a 120-class, 9,453-video native ISL dataset is created. Augmented by Physical Node Dropout and ArcFace optimization to mitigate subject-specific memorization, our framework achieves a competitive 89.8% accuracy, establishing a strong baseline for complex sign language modeling.

Keywords Indian Sign Language · Sign Language Recognition · Graph Convolutional Networks · Transformer · Elastic Temporal Modeling

1 Introduction

Disabling hearing loss affects over 430 million people globally, a figure projected to reach 700 million by 2050 [58]. For millions of individuals, sign language serves

as the primary mode of communication [16]. Despite the existence of over 300 sign languages used by 70 million people [41, 56], Indian Sign Language (ISL) remains severely under-resourced [23]. ISL is a highly dynamic visual language featuring rapid micro-motions, prolonged hold-phases, and complex bimanual coordination [18], necessitating architectures with precise spatial and adaptive temporal modeling capabilities.

Standard spatial-temporal models, including Graph Convolutional Networks (GCNs) and Video Transformers, face critical bottlenecks in this domain [3]. Uniform frame sampling wastes computation on static hold-phases while risking the loss of rapid kinematic transitions [11]. Conversely, dynamically dropping redundant frames creates irregular physical time gaps that disrupt standard Transformers' sequential rhythm assumptions [8].

To address these limitations, an efficient, structurally-aware framework for large-vocabulary ISL recognition is proposed. The proposed model is a hybrid architecture: a Kinetically-Conditioned Adaptive Part-GCN handles spatial skeleton structure via graph convolution, while a distance-aware Elastic Transformer performs temporal reasoning over the resulting pose tokens. This cross-architectural design bridges the complementary strengths of graph-based spatial modeling and attention-based temporal reasoning for irregular sequences. To minimize computational redundancy, our framework utilizes an entropy-driven dimensionality reduction pipeline based on Singular Value Decomposition (SVD) [51] and Kullback-Leibler (KL) divergence [30] to extract only the most kinematically innovative keyframes. To process these irregularly sampled frames without losing natural signing rhythm, a Continuous Dual Positional Encoding with a distance-aware Elastic Transformer is paired. Spatially, the Part-GCN processes a unified 8-channel

spatio-kinematic embedding to accurately model bimanual geometry.

Finally, to mitigate subject-memorization and distinguish visually similar classes, Physical Node Dropout alongside an Additive Angular Margin Loss (ArcFace) [10] is employed. To robustly validate our architecture, a novel 9,453-video, 120-class native ISL benchmark featuring diverse abstract terminology is created. Unlike existing ISL corpora that suffer from severe class sparsity and force models into few-shot paradigms, our dataset provides substantial depth (averaging ~ 78 videos per gloss) with signer-disjoint train and test partitions. This density provides the necessary foundation to effectively learn identity-invariant kinematics.

In summary, our primary contributions are:

- A new 120-class, 9,453-video native ISL dataset. On this benchmark, our model achieves an 89.8% Test Accuracy and 90% Macro F1-Score, demonstrating strong generalization with a minimal ($\sim 3.5\%$) generalization gap.
- An SVD-driven cumulative subsampling module paired with Continuous Dual Positional Encodings, enabling the Elastic Transformer to process irregular time steps while preserving physical rhythm.
- Our hybrid framework, which efficiently bridges spatial graph convolutions and temporal transformers to model isolated sign language kinematics.

2 Related Work

Broader Sign Language Processing The research community has actively pursued tools and techniques for processing sign languages, establishing an intersection between the computer vision [31] and natural language processing [62] domains. Consequently, a wide array of sign language processing (SLP) tasks have been formalized. These include sign language detection [37], identification [36], segmentation [6], generation [59, 48], and continuous translation [7, 38, 63, 39, 20]. While self-supervised pretraining [49] and phonological modeling [54] have recently gained traction, the severe unavailability of densely annotated resources remains the primary challenge for robust, data-driven SLP approaches.

ISL Datasets and Recognition Constraints Data-driven Sign Language Recognition (SLR) is strictly constrained by the availability of large-scale benchmarks. Early Indian Sign Language (ISL) research predominantly focused on static alphabet recognition using 2D object detectors [53] or limited-vocabulary gestures [35, 42, 26]. For dynamic temporal glosses, large-vocabulary corpora (**Table 1**) like CISLR, introduced by [22], and

INCLUDE [50] provided foundational baselines. However, their severe sparsity (e.g., CISLR averaging 1.5 videos per word) forces models into complex few-shot or prototypical learning paradigms to compensate for data scarcity [5]. While recent large-scale continuous benchmarks address sentence-level translation [23, 21], a gap remains for deep, isolated gloss evaluation where high sample density per class allows models to learn identity-invariant kinematics. Our proposed 120-class dataset directly addresses this by providing substantial depth (averaging ~ 78 videos per class), delivering the data density necessary to learn identity-invariant kinematics without few-shot workarounds.

Architectural Modeling for Sign Language Standard spatial temporal architectures, including 3D CNNs and standard Video Transformers, struggle to capture the complex bimanual nature and irregular rhythmic velocities of ISL, resulting in sub-optimal recognition [23, 21]. Furthermore, standard temporal graph networks implicitly assume uniform frame sampling [65]. This assumption is inherently flawed for sign language: it wastes computation on static hold-phases while risking the loss of rapid micro-motions. As demonstrated in our ablation study (Table 7), replacing entropy-driven SVD sampling with uniform temporal sampling causes a 2.3% accuracy drop, directly confirming that content-aware frame selection is critical. The proposed architecture directly addresses these established bottlenecks by replacing naive spatial processing with a Kinetically-Conditioned Adaptive GCN and bypassing the uniform-sampling assumption via a distance-aware Elastic Temporal Transformer.

Irregular Temporal Modeling A parallel line of work in sequence modeling addresses the challenge of non-uniform temporal spacing. Continuous-time Transformers such as ContiFormer [8] extend standard attention to handle continuous-time inputs by incorporating time-delta information into the attention kernel, while Neural ODE-based approaches model latent dynamics between irregular observations. Our Continuous Dual Positional Encoding and distance-aware attention bias are motivated by this paradigm but are specifically tailored for sign language: rather than modeling continuous-time dynamics, we encode both the discrete logical ordering of entropy-selected keyframes and their absolute physical timestamps, allowing the Elastic Transformer to reason about irregular physical gaps without sacrificing sequential coherence.

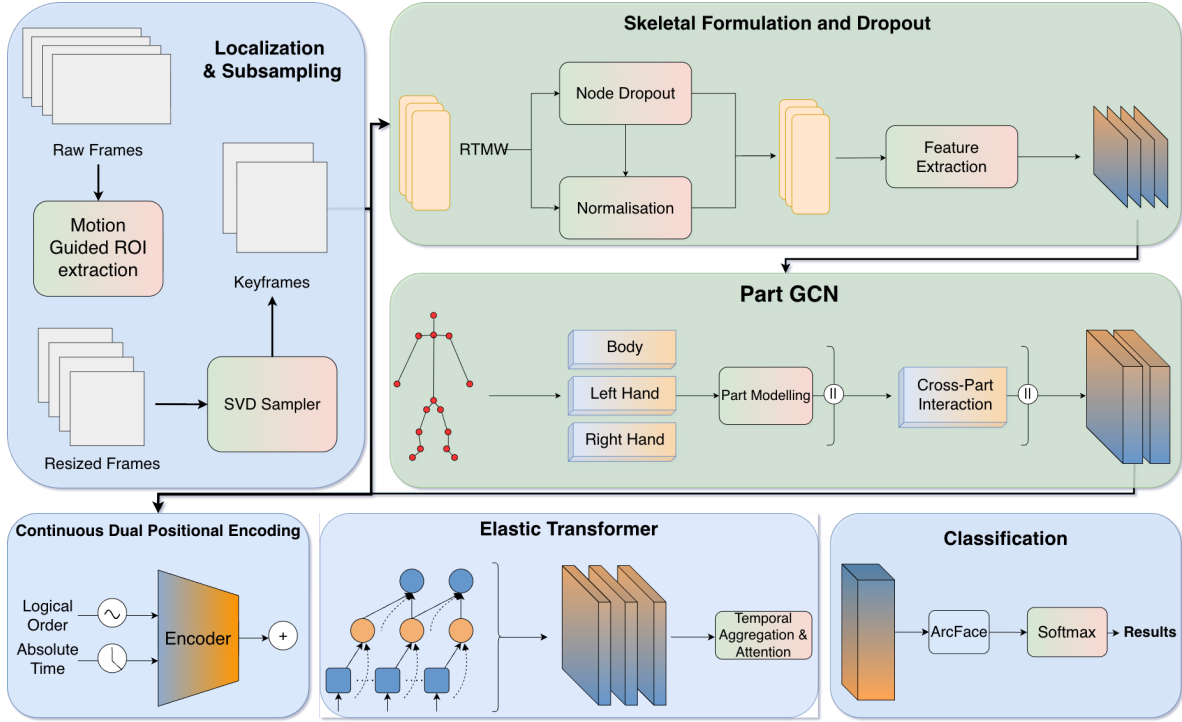


Fig. 1: Overall architecture of the proposed hybrid framework. The architecture is a hybrid: a Kinetically-Conditioned Part-GCN handles spatial skeleton structure, while the distance-aware Elastic Transformer performs temporal reasoning over the resulting pose tokens.

Table 1: Comparison of existing Indian Sign Language (ISL) datasets based on vocabulary size, total videos, and average videos per word.

Datasets	Words	Videos	Avg. Samples
[40]	22	600	27.3
[27]	80	800	10
INCLUDE [50]	263	4287	16.3
CISLR [22]	4765	7050	1.5
ISL-CSLTR [14]	186	700	3.8
Our	120	9453	78.8

3 Method

The proposed hybrid framework (Figure 1) transforms raw video into structurally-aware skeletal tokens through dynamic localization, entropy-driven subsampling, and spatio-temporal modeling.

3.1 Localization and Subsampling

To minimize background interference, a Motion-Guided ROI extraction module computes a global motion heatmap. Given a video $V = \{F_1, \dots, F_n\}$, temporal innovation between Gaussian-smoothed frames is $\Delta_t(x, y) = |(F_t * G) - (F_{t-1} * G)|$. Motion energy is accumulated into a global map A by thresholding pixel-wise differences (τ_L):

$G) - (F_{t-1} * G)|$. Motion energy is accumulated into a global map A by thresholding pixel-wise differences (τ_L):

$$A(x, y) = \sum_{t=2}^n \mathbb{I}(\Delta_t(x, y) > \tau_L) \quad (1)$$

Otsu’s thresholding [43] generates a binary mask S , followed by morphological closing $S_{closed} = (S \oplus K) \ominus K$, where $\oplus$ and $\ominus$ denote dilation and erosion by structuring element K , respectively. This operation fills interior holes and smooths contour boundaries to extract a clean, localized square bounding box.

To preserve signing dynamics while pruning redundant hold-phases, a Cumulative SVD-Innovation Sampler extracts grayscale proxies P_t , performs SVD ($P_t = U \Sigma V^T$), and normalizes the singular value spectrum:

$$s_t(i) = \frac{\sigma_i}{\sum_j \sigma_j + \epsilon} \quad (2)$$

A sequence-specific threshold Θ is defined using Kullback-Leibler (KL) Divergence modified by a sparsity factor α :

$$\Theta = \mu(D_{KL}(s_t \parallel s_{t-1})) + \alpha \cdot \sigma(D_{KL}(s_t \parallel s_{t-1})) \quad (3)$$

Crucially, a frame t is selected as a keyframe only if its innovation relative to the last selected keyframe k exceeds Θ :

$$D_{KL}(s_t \parallel s_k) > \Theta \quad (4)$$

This enforces a minimum stride ($t - k > 1$) and acts as a lightweight gatekeeper, drastically reducing the number of frames passed to the heavy pose extractor.

3.2 Skeletal Formulation and Dropout

Selected keyframes are processed via RTMPose-Wholebody [19] to extract a 65-node topology ($V = 65$), yielding coordinates $C \in \mathbb{R}^{T \times V \times d}$. To mitigate subject-memorization, Physical Node Dropout randomly occludes an entire subgraph (Left Hand, Right Hand, or Body) during training with probability p :

$$\hat{C}_{t,i} = C_{t,i} \odot M_i \ ; \quad M_i = 0 \ \forall i \in \mathcal{V}_{subgraph} \quad (5)$$

Surviving coordinates are translation-normalized via the nose centroid. An 8-channel spatio-kinematic tensor $X_{in} \in \mathbb{R}^{T \times V \times 8}$ is then formulated from five biomechanically-motivated feature groups: spatial coordinates (X, Y) provide absolute joint position; velocity $v_t = (\hat{C}_t - \hat{C}_{t-1}) / \max(\Delta t, \epsilon)$ encodes dynamic motion; spatial bone vectors $\hat{b}_i = \hat{C}_i - \hat{C}_{anchor}$ capture limb orientation; joint angles $\cos(\theta_{i,j}) = (\hat{b}_i \cdot \hat{b}_j) / (\|\hat{b}_i\| \|\hat{b}_j\| + \epsilon)$ represent relative joint configuration invariant to subject scale; and pose confidence scores down-weight unreliable landmarks during propagation.

3.3 Kinematically-Conditioned Adaptive Part-GCN

To ground the network in human biomechanics, spatial processing is initialized from a binary physical adjacency matrix prior $A_{prior} \in \{0, 1\}^{V \times V}$, where $A_{prior,ij} = 1$ if joints i and j share a direct anatomical connection and 0 otherwise. The learnable adjacency matrix A_{learn} is initialized as $A_{prior} + \mathcal{N}(0, 0.01)$. The graph convolution operation employs a row-wise Softmax to dynamically normalize edges and allow asymmetric directed information flow:

$$Z = \text{Softmax}(A_{learn}, \text{dim} = -1)(X_{in}W) + b \quad (6)$$

Instead of uniform processing, the architecture partitions the 65-node spatio-kinematic tensor into three independent subgraphs Body (23 nodes), Left Hand (21 nodes), and Right Hand (21 nodes) each processed by a dedicated Part-GCN block to extract robust, localized kinematic representations.

To capture the bimanual coordination crucial to sign language, the network computes Euclidean distances between physical anchors (e.g., wrist-to-wrist, index-to-index, hands-to-face). These distances are embedded via a Multi-Layer Perceptron (MLP) to form a cross-part interaction token. Finally, localized outputs from the three Part-GCNs are concatenated with this interaction token, constructing a holistic, structurally-aware pose token.

3.4 Continuous Dual PE and Elastic Transformer

To capture localized micro-motions prior to self-attention, the sequentially embedded tokens $X_{pe} \in \mathbb{R}^{B \times D_{pose} \times T}$ pass through a 1D Temporal Convolutional Network (TCN) within a residual block:

$$X_{local} = \mathcal{F}_{TCN}(X_{pe}) + X_{pe} \quad (7)$$

Because entropy-driven subsampling creates irregular time gaps, standard discrete positional encodings fail. Input sequences are therefore mapped to a Continuous Dual Positional Encoding. A learnable embedding matrix $E_{rel} \in \mathbb{R}^{T_{max} \times (D/2)}$ captures the discrete, logical sequence order of the surviving keyframes, while absolute physical timestamps $t_{abs} \in [0, 1]$ drive sinusoidal encodings [57]:

$$PE_{abs}(t_{abs}, 2i) = \sin\left(\frac{t_{abs}}{10000^{2i/(D/2)}}\right) \quad (8)$$

$$PE_{abs}(t_{abs}, 2i + 1) = \cos\left(\frac{t_{abs}}{10000^{2i/(D/2)}}\right) \quad (9)$$

Global reasoning is performed by Elastic Transformer Blocks. A Distance-Aware Attention Bias dynamically modulates the attention matrix based on the physical temporal gap $\delta_{ij} = |t_{abs,i} - t_{abs,j}|$ between tokens i and j , using a multi-layer perceptron:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}} + \text{MLP}(\delta_{ij}) + \mathcal{M}\right)V \quad (10)$$

A Masked Temporal Aggregator condenses the variable-length sequence into a video-level vector v by learning scalar attention weights α_t over all T time steps:

$$\alpha_t = \text{Softmax}_t(\mathbf{w}^T \tanh(W_a X_{local,t} + b_a)) \quad (11)$$

$$v = \sum_{t=1}^T \alpha_t X_{local,t} \quad (12)$$

where Softmax_t denotes normalization over the time dimension, and $X_{local,t}$ is the post-TCN, post-Transformer token at step t .

3.5 Angular Margin Optimization

Standard linear classification is replaced with an Additive Angular Margin Loss (ArcFace) [10]. ArcFace enforces a fixed angular margin m between classes on a hypersphere, producing more compact intra-class and more discriminative inter-class feature distributions than standard cross-entropy a property particularly beneficial for sign language, where many glosses share overlapping sub-movements. Given L2-normalized representations v and classifier weights W_j :

$$\mathcal{L}_{ArcFace} = -\log \frac{e^{s \cdot \cos(\theta_{y_i} + m)}}{e^{s \cdot \cos(\theta_{y_i} + m)} + \sum_{j \neq y_i} e^{s \cdot \cos \theta_j}} \quad (13)$$

To equitably propagate gradients across the dataset’s inherent class imbalance, this objective is weighted by the inverse class frequency $w_{y_i} \propto 1/\text{freq}(y_i)$. The final training loss is therefore:

$$\mathcal{L}_{Total} = -\sum_{i=1}^N w_{y_i} \log \left(p_{y_i}^{(ArcFace)} \right) \quad (14)$$

where $p_{y_i}^{(ArcFace)}$ is the scaled softmax probability derived from Eq. (13).

4 Experiments

This section comprehensively evaluates the efficacy and efficiency of the proposed framework. A 120-class native ISL benchmark is introduced. Subsequently, the architectural contributions are evaluated through rigorous comparisons against established baselines and public datasets, concluding with detailed ablation studies and computational profiling.

4.1 Dataset Description

The scarcity of large-vocabulary skeletal benchmarks severely bottlenecks dynamic ISL recognition. To provide a robust testbed, a native 120-class ISL repository comprising 9,453 videos (averaging 42.2 frames per video) is collected. The vocabulary spans 5 semantic categories (**Table 2**). Class frequencies range from 33 to 97 videos per gloss; this imbalance necessitates the dynamic class-weighted loss formulation from Section 3.5.

Acquisition and Partitioning: Data was recorded using a Kinect V2 at 15 FPS. To maximize real-world webcam generalizability, the RGB stream is deliberately utilized, explicitly ignoring depth and infrared modalities. The dataset is partitioned via a class-stratified

Table 2: Dataset distribution across semantic categories including word counts, weight ranges, and signer averages.

Category	Words	Class Weight	Avg. samples
Medical	27	[0.64, 1.26]	61.30
Geographical	23	[0.90, 2.66]	86.90
Technical	18	[0.90, 2.66]	87.11
Action Verbs	20	[0.86, 2.52]	86.25
Daily Life	32	[0.85, 2.47]	78.30
Full Dataset	120	[0.81, 2.41]	78.80

split into a training repository (7,527 samples) and a held-out test repository (1,926 samples) used for all final evaluations. To support reproducibility and future research, the dataset and model code will be made publicly available upon acceptance of this work.

Signer Distribution and Identity Protocol: Data was collected over several years through multiple acquisition drives at academic institutions and community schools, yielding hundreds of distinct volunteer signers. Crucially, each volunteer performed any given gloss at most once. Consequently, for every gloss class, the set of signers present in the training partition is strictly disjoint from the set present in the test partition guaranteeing a zero-shot identity scenario for each word. While the total count of unique signers across all classes was not formally tracked due to the distributed, multi-year nature of collection, the scale and diversity of acquisition sites make cross-partition identity overlap across classes negligibly unlikely. This protocol forces the architecture to learn invariant kinematic semantics rather than memorizing subject-specific anatomy.

4.2 Data Preprocessing

To maximize computational efficiency, the raw dataset is dynamically processed (**as illustrated in Figure 2**) before entering the main architecture.

Visual Localization: Raw videos are first localized via Motion-Guided ROI extraction using a temporal innovation threshold τ_L and padding factor λ . The active signing region is resized to a 512×512 grayscale proxy to mitigate background and lighting variances.

Dynamic SVD Subsampling: To eliminate temporal redundancy and variable signing velocities, the SVD-Innovation Sampler extracts high-entropy keyframes based on a sparsity multiplier α and sequence constraint T_{max} . This pipeline successfully compresses the raw input from an

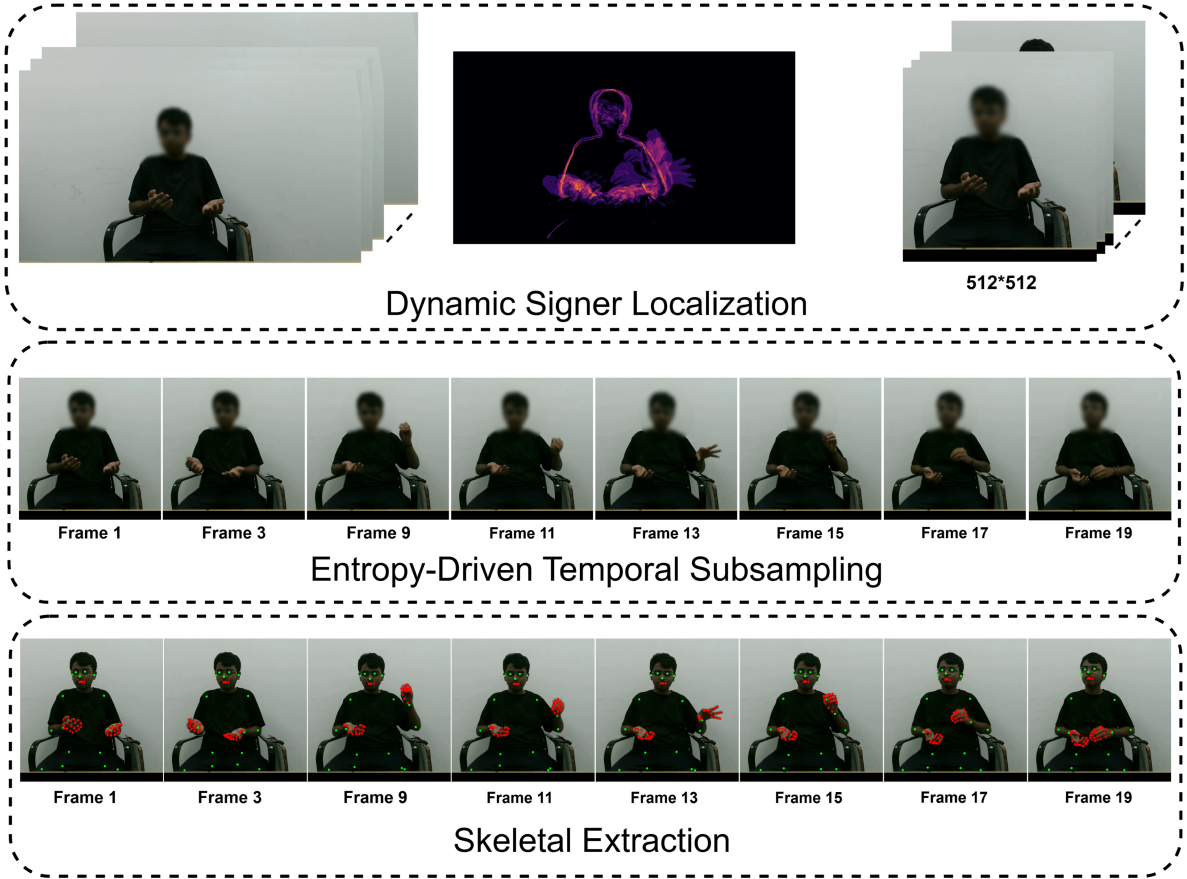


Fig. 2: The dynamic preprocessing pipeline of our framework. Raw sequences undergo Motion-Guided Region of Interest (ROI) extraction, followed by entropy-driven SVD temporal subsampling, and finally RTMW skeletal extraction to form the unified input representation.

average of 42.2 frames down to a sparse 13.2 keyframes per video a 68.6% reduction in temporal dimensionality that significantly minimizes computational overhead without sacrificing critical gestural semantics.

Input Feature Composition: The extracted keyframes undergo skeletal pruning via RTMW to isolate a 65-node topology tailored for ISL: 23 upper-body/facial nodes ($\mathcal{V}_{body}$) and 21 nodes per hand ($\mathcal{V}_{lh}, \mathcal{V}_{rh}$). The coordinates are normalized relative to the nose centroid and shoulder width. Finally, they are expanded into the unified 8-channel spatio-kinematic representation, forming the final input tensor $X_{in} \in \mathbb{R}^{T \times 65 \times 8}$.

4.3 Training Details and Reproducibility

Experiments utilized an Intel i9-13900K CPU and a 24GB RTX 4090 GPU. While temporal subsampling used grayscale proxies, RTMPose skeletal extraction utilized full-resolution RGB for maximum landmark

precision. The network trained for 100 epochs using AdamW (with high weight decay to regularize the transformer) and a ReduceLROnPlateau scheduler. Strict deterministic behavior (*seed* = 42) across splits and augmentations guaranteed exact reproducibility.

From the 7,527 training samples, an internal class-stratified 80/20 split yields approximately 6,022 samples for gradient updates and 1,505 samples for an internal validation subset, used exclusively for checkpoint selection via minimum validation loss monitoring. All reported performance metrics are derived strictly from the separate, held-out test repository of 1,926 samples, which is never used during training or model selection. All exhaustive configuration parameters including spatial dimensions, ArcFace margins, and learning rates are consolidated in **Appendix A**.

Quantitative Performance Our framework is evaluated on our native 120-class dataset, deriving all metrics strictly from the held-out test partition using the checkpoint with the minimum validation loss. As detailed in

Table 3, the model achieves an 89.8% Test Accuracy and a 90.06% Macro F1-Score. This tight alignment indicates that our dynamic class-weighting and ArcFace angular margin successfully mitigate inherent dataset imbalance, ensuring equitable feature representation across both frequent and rare classes.

Table 3: Category-level performance breakdown of the proposed model.

Category	Train Acc	Train F1	Test Acc	Test F1
Medical	99.2%	99.2%	98.2%	98.2%
Geographical	97.0%	96.8%	91.0%	90.1%
Technical	98.0%	98.0%	90.7%	90.5%
Action Verbs	97.5%	97.5%	88.3%	88.3%
Daily Life	96.2%	96.3%	89.1%	89.2%
Full	93.3%	93.5%	89.8%	90.0%

Training Dynamics and Generalization Over a 100-epoch schedule, the network demonstrated stable convergence. Optimal generalized weights were recovered at Epoch 80, achieving a minimal test loss of 2.088 alongside a Train Accuracy of 93.84% and Train F1-Score of 94.14%. The narrow $\sim 3.5\%$ generalization gap validates our structural regularization strategy. Consistent with our ablation results (Table 7), this resistance to overfitting is primarily attributable to Physical Node Dropout ($p = 0.20$) and the Entropy-Driven Subsampling, which together force the model to learn invariant kinematic semantics rather than redundant, easily overfit spatial trajectories.

4.4 Evaluation on Public Benchmarks

To validate cross-domain and signer-independent generalization, our framework is evaluated on the Argentinian Sign Language [47] LSA64, INCLUDE50, and INCLUDE ISL benchmarks. For LSA64, a strict zero-identity-leakage protocol (Subjects 1–8 for training, 9–10 for testing) is applied.

Our SVD sampler demonstrated robust efficiency across framerates: on LSA64 (60 FPS), it condensed 127.8-frame sequences to 38.8 keyframes (69.6% reduction); on INCLUDE configurations (25 FPS), 64.25-frame sequences were compressed to 26.7 keyframes (58.4% reduction). This yields substantial computational savings during pose extraction while preserving critical kinematics. As detailed in **Table 4**, our approach achieves competitive predictive performance on both datasets, proving our architecture successfully transfers generalized representations to unseen signers and diverse recording environments.

4.5 Comparison with Baseline Architectures

To isolate our architectural contributions, standard skeleton baselines (ST-GCN, CTR-GCN) are evaluated on our 120-class benchmark using the exact same SVD-extracted spatio-kinematic features.

As detailed in **Table 5**, our model significantly outperforms both baselines. The sharp performance drop in the standard networks (e.g., CTR-GCN dropping to 64.8% Test Acc) highlights a fundamental architectural limitation: standard temporal graph networks implicitly assume uniform time steps and fail to interpret irregular temporal gaps. By replacing standard convolutions with a distance-aware Elastic Transformer and continuous positional encodings, the proposed architecture natively preserves physical signing rhythm, maximizing the utility of subsampled data.

Furthermore, **Table 6** presents a targeted comparison against highly parameterized vision architectures (VideoMAE, MViTv2). Even against these resource-intensive models, our lightweight framework demonstrates highly competitive performance.

Ablation Study To quantify the impact of our architectural contributions, an exhaustive ablation study (**Table 7**) is conducted. Removing the Continuous Positional Encoding (PE) and Kinematic features causes a severe performance collapse (dropping to 80.3% Test Acc), proving that standard sequence modeling fails on irregular temporal gaps without explicit spatial-temporal priors. The ablation also directly confirms the benefit of entropy-driven selection over uniform sampling (87.5% vs. 89.8%), validating that content-aware keyframe extraction is essential. Furthermore, bypassing the SVD Sampler to process “All Frames” yields only a marginal 1.2% accuracy gain, validating that our dimensionality reduction pipeline successfully discards redundancy while fully preserving critical gestural semantics.

4.6 Computational Efficiency and Hardware Deployment

As demonstrated in our ablation study (Table 7), processing the full raw sequence (“All Frames”) yields a marginal accuracy improvement (91.0% vs 89.8%). However, this requires processing an average of 42.2 frames per video. By utilizing the proposed entropy-driven SVD sampler, our framework reduces the temporal dimensionality by 68.6% (averaging 13.2 keyframes), trading a 1.2% accuracy drop for substantial computational savings.

Profiling the full proposed architecture (inclusive of the TCN and Transformer blocks) with a maximum

Table 4: Performance of the proposed framework on established public sign language benchmarks. Training accuracies are reported for our model only to demonstrate the absence of underfitting; they were not available in the original baseline publications (denoted –).

Model	Dataset	Train Acc	Test Acc
MobileNetV2 BiLSTM [50]	INCLUDE50	–	94.5%
DCDW-LSTM [12]	INCLUDE50	–	91.3%
BERT Transformer [44]	INCLUDE50	–	89.5%
C3D-BiLSTM MHAttention [13]	INCLUDE50	–	92.46%
Meta-Learning + LSTM RNN [24]	INCLUDE50	–	84.52%
MediaPipe CNN BiLSTM [15]	INCLUDE50	–	97.44%
Mediapipe Attention-based BiLSTM [52]	INCLUDE50	–	88.63%
MediaPipe LSTM [25]	INCLUDE50	–	94.8%
EfficientNet MultiFeature Attention GRU [61]	INCLUDE50	–	98.41%
MobileNetV2 BiLSTM [50]	INCLUDE	–	85.6%
Transfer Learning Deep Learning [33]	INCLUDE	–	96.0%
InceptionV3-GRU [34]	LSA64(Leave Two Subject Out)	–	74.22%
Resnet-18 with RReLU [1]	LSA64(Leave Two Subject Out)	–	94.68%
BoW+SubCls [46]	LSA64(Leave One Subject Out)	–	91.7%
Cumulative SD-VLAD with SVM [45]	LSA64(Leave One Subject Out)	–	85%
Transformer [2]	LSA64(Leave One Subject Out)	–	91.09%
LSTM [29]	LSA64(Random Split)	–	98.09%
VGG-16, LSTM [28]	LSA64(Random Split)	–	99.84%
MEMP network [64]	LSA64(Random Split)	–	99.063%
CNN, kernel-based extreme learning machine [17]	LSA64(Random Split)	–	97.81%
Proposed Framework (Ours)	INCLUDE50	99.3%	97.1%
Proposed Framework (Ours)	INCLUDE	98.2%	95.8%
Proposed Framework (Ours)	LSA64(Leave Two Subject Out)	98.5%	94.4%
Proposed Framework (Ours)	LSA64(Stratified Sampling)	98.6%	99.1%

Table 5: Comparison of our proposed model against established skeleton baselines on the 120-class ISL dataset. All models utilize the proposed SVD preprocessing pipeline.

Architecture	Train Acc	Test Acc
CTR-GCN [9]	69.9%	64.8%
ST-GCN [60]	85.1%	83.0%
Proposed (Ours)	93.3%	89.8%

sequence length of $T = 25$ and $D_{pose} = 512$ yields the highly efficient footprint detailed in **Table 8**.

This optimized profile is a direct consequence of the SVD sampler: because the self-attention mechanism’s complexity scales quadratically with sequence length ($O(T^2)$), reducing the sequence to a maximum bound of $T = 25$ (averaging 13.2) collapses the attention matrix footprint, allowing the 6.1M parameters to operate with high efficiency.

5 Conclusion

This work presented an efficient hybrid framework for large-vocabulary Indian Sign Language recognition. By

unifying spatial graph convolutions with elastic temporal attention, the architecture effectively isolates invariant kinematic structures from raw video sequences. A cornerstone of this success is the dimensionality reduction pipeline; the entropy-driven temporal subsampling dynamically filters out redundant frames, significantly reducing computational overhead while preserving crucial gestural semantics. Combined with ArcFace optimization to enforce robust decision boundaries, the model achieved 89.8% Test Accuracy and a 90.06% Macro F1-score on a novel 120-class ISL benchmark, demonstrating its capability in handling complex, abstract gloss terminology.

Limitations

While our approach establishes a strong computational baseline, its reliance on pose extractors means severe occlusions or extreme motion blur can corrupt initial skeletal tokens and propagate errors through the network. Additionally, the primary dataset’s acquisition at 15 FPS may restrict the recognition of rapid finger-spelling, despite the temporal convolution network’s ability to capture local micro-motions. Furthermore, the current structural embedding focuses strictly on hand and body posture, lacking explicit tracking of non-

Table 6: Performance comparison of alternative vision architectures. These models were evaluated exclusively on the Geographical word category subset due to their extreme computational overhead.

Approach	Pre-trained Prior	Train Acc (%)	Test Acc (%)
SlowFast_R50	Kinetics-400	63.40	30.62
MViTv2 [32]	Kinetics-400	33.84	33.63
VideoMAE	Kinetics-400	99.00	42.60
VideoMAE [55]	SSv2	99.00	74.62
VideoMAE (Staged Fine-Tuning)	SSv2	100.00	81.20
VideoMAE (LoRA Only)	SSv2	80.00	60.00
VideoMAE (LoRA + Unfreeze)	SSv2	99.00	76.60
V-JEPA-2 [4]	SSv2	97.46	88.32

Table 7: Ablation study of the proposed framework on the 120-class ISL dataset.

Modification	Train Acc	Train F1	Test Acc	Test F1
Proposed Full Framework (Ours)	93.3%	93.5%	89.8%	90.0%
<i>Loss Functions</i>				
w/o Weighted Loss	93.1%	93.3%	89.4%	89.5%
w/o ArcFace	94.2%	94.4%	88.0%	88.2%
w/o ArcFace + w/o Weighted Loss	93.9%	94.2%	87.4%	87.8%
<i>Architecture Modules</i>				
w/o Absolute Positional Encoding (PE)	93.9%	94.0%	88.5%	88.7%
w/o Physical Node Dropout	96.3%	96.5%	88.1%	88.3%
w/o Temporal Convolution (TCN)	91.2%	91.5%	88.0%	88.1%
<i>Data & Preprocessing</i>				
w/o SVD Sampler (All Frames)	96.0%	96.2%	91.0%	91.0%
Uniform Temporal Sampling	91.5%	91.7%	87.5%	87.7%
w/o Kinematic Features (X,Y only)	82.3%	83.0%	83.2%	83.0%
w/o PE + w/o Kinematic Features	81.3%	82.0%	80.3%	80.4%

Table 8: Computational footprint of the proposed architecture.

Metric	Measurement
Trainable Parameters	6.13 Million
Theoretical Compute	252.57 MFLOPs
Model Weights (float32)	23.39 MB
Peak VRAM ($B = 1$)	34.91 MB
Inference Latency	2.87 ± 0.19 ms
Throughput (RTX 4090)	348.7 samples/sec

manual markers such as fine-grained facial expressions or mouthing that are crucial to natural sign language linguistics. Future work will explore lightweight multi-modal fusion to integrate these non-manual markers and expand the dimensionality reduction pipeline to handle dynamic frame-rate inputs. Finally, we plan to scale the dataset by collecting more isolated and continuous data, facilitating the evaluation of the framework on larger corpora and extending the architecture to sentence-level continuous translation tasks.

References

1. Akdağ, A., Baykan, Ö.K.: Isolated sign language recognition through integrating pose data and motion history images. *PeerJ Computer Science* **10**, e2054 (2024)
2. Alyami, S., Luqman, H., Hammoudeh, M.: Isolated arabic sign language recognition using a transformer-based model and landmark keypoints. *ACM Transactions on Asian and Low-Resource Language Information Processing* **23**(1), 1–19 (2024)
3. Arib, S.H., Akter, R., Rahman, S., Rahman, S.: Signformer-gcn: Continuous sign language translation using spatio-temporal graph convolutional networks. *PloS one* **20**(2), e0316298 (2025)
4. Assran, M., Bardes, A., Fan, D., Garrido, Q., Howes, R., Muckley, M., Rizvi, A., Roberts, C., Sinha, K., Zholus, A., et al.: V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985* (2025)
5. Aziz, S.: Attention-enhanced few-shot prototypical network for indian sign language recognition. In: 2025 IEEE 7th International Conference on Computing, Communication and Automation (ICCCA) (2025)
6. Bull, H., Gouiffès, M., Braffort, A.: Automatic segmentation of sign language into subtitle-units. In: *European Conference on Computer Vision*, pp. 186–198. Springer (2020)
7. Camgoz, N.C., Hadfield, S., Koller, O., Ney, H., Bowden, R.: Neural sign language translation. In: *Proceedings*

- of the IEEE conference on computer vision and pattern recognition, pp. 7784–7793 (2018)
8. Chen, Y., Ren, K., Wang, Y., Fang, Y., Sun, W., Li, D.: Contiformer: Continuous-time transformer for irregular time series modeling. *Advances in Neural Information Processing Systems* **36**, 47,143–47,175 (2023)
 9. Chen, Y., Zhang, Z., Yuan, C., Li, B., Deng, Y., Hu, W.: Channel-wise topology refinement graph convolution for skeleton-based action recognition. In: *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 13,359–13,368 (2021)
 10. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4690–4699 (2019)
 11. Deng, Z., Min, W., Han, Q., Liu, M., Li, L.: Vtan: A novel video transformer attention-based network for dynamic sign language recognition. *Computers, Materials, & Continua* **82**(2), 2793 (2025)
 12. Dey, A., Biswas, S.: Gesture recognition for isl-question signs in videos using dcdw-lstm attention. In: *2023 IEEE 5th PhD Colloquium on Emerging Domain Innovation and Technology for Society (PhD EDITS)*, pp. 1–2. IEEE (2023)
 13. Dey, A., Biswas, S., Le, D.N.: Recognition of wh-question sign gestures in video streams using an attention driven c3d-bilstm network. *Procedia Computer Science* **235**, 2920–2931 (2024)
 14. Elakkiya, R., Natarajan, B.: Isl-csltr: Indian sign language dataset for continuous sign language translation and recognition (2021). DOI 10.17632/kcmpdxky7p.1. URL <https://doi.org/10.17632/kcmpdxky7p.1>
 15. Ghanimi, H.M., Sengan, S., Sadu, V.B., Kaur, P., Kaushik, M., Alrooba, R., Baqasah, A.M., Alsafyani, M., Dadheech, P.: An open-source mp+ cnn+ bilstm model-based hybrid model for recognizing sign language on smartphones. *International Journal of System Assurance Engineering and Management* **15**(8), 3794–3806 (2024)
 16. Holmer, E., Schönström, K., Andin, J.: Associations between sign language skills and resting-state functional connectivity in deaf early signers. *Frontiers in psychology* **13**, 738,866 (2022)
 17. Imran, J., Raman, B.: Deep motion templates and extreme learning machine for sign language recognition. *The Visual Computer* **36**(6), 1233–1246 (2020)
 18. Jain, S., Raja, K.S., Mukerjee, M.P.A.: Indian sign language character recognition. Indian Institute of Technology, Kanpur Course Project-CS365A (2016)
 19. Jiang, T., Lu, P., Zhang, L., Ma, N., Han, R., Lyu, C., Li, Y., Chen, K.: RtmPose: Real-time multi-person pose estimation based on mmpose. *arXiv preprint arXiv:2303.07399* (2023)
 20. Jiang, Z., Moryossef, A., Müller, M., Ebling, S.: Machine translation between spoken languages and signed languages represented in signwriting. In: *Findings of the Association for Computational Linguistics: EACL 2023*, pp. 1706–1724 (2023)
 21. Joshi, A., Agrawal, S., Modi, A.: Isltranslate: Dataset for translating indian sign language. In: *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 10,466–10,475 (2023)
 22. Joshi, A., Bhat, A., Gole, P., Gupta, S., Agarwal, S., Modi, A., et al.: Cislr: Corpus for indian sign language recognition. In: *Proceedings of the 2022 conference on empirical methods in natural language processing*, pp. 10,357–10,366 (2022)
 23. Joshi, A., Mohanty, R., Kanakanti, M., Mangla, A., Choudhary, S., Barbate, M., Modi, A.: isign: A benchmark for indian sign language processing. In: *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 10,827–10,844 (2024)
 24. Katti, R.K., Sujatha, C., Desai, P., Shankar, G.: Character and word level gesture recognition of indian sign language. In: *2023 IEEE 8th International Conference for Convergence in Technology (I2CT)*, pp. 1–6. IEEE (2023)
 25. Khartheesvar, G., Kumar, M., Yadav, A.K., Yadav, D.: Automatic indian sign language recognition using mediapipe holistic and lstm network. *Multimedia Tools and Applications* **83**(20), 58,329–58,348 (2024)
 26. Kirandevraj, R., Kurmi, V.K., Nambodiri, V.P., Jawahar, C.: Continuous fingerspelling dataset for indian sign language. In: *Proceedings of the Workshop on Sign Language Processing (WSLP)*, pp. 33–38 (2025)
 27. Kishore, P.V.V., Kumar, P.R.: A video based indian sign language recognition system (inslr) using wavelet transform and fuzzy logic. *International Journal of Engineering and Technology* **4**(5), 537 (2012)
 28. Konstantinidis, D., Dimitropoulos, K., Daras, P.: A deep learning approach for analyzing video and skeletal features in sign language recognition. In: *2018 IEEE international conference on imaging systems and techniques (IST)*, pp. 1–6. IEEE (2018)
 29. Konstantinidis, D., Dimitropoulos, K., Daras, P.: Sign language recognition based on hand and body skeletal data. In: *2018-3DTV-Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON)*, pp. 1–4. IEEE (2018)
 30. Kullback, S., Leibler, R.A.: On information and sufficiency. *The annals of mathematical statistics* **22**(1), 79–86 (1951)
 31. Li, D., Rodriguez, C., Yu, X., Li, H.: Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In: *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 1459–1469 (2020)
 32. Li, Y., Wu, C.Y., Fan, H., Mangalam, K., Xiong, B., Malik, J., Feichtenhofer, C.: Mvitv2: Improved multiscale vision transformers for classification and detection. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4804–4814 (2022)
 33. Mahyoub, M., Natalia, F., Sudirman, S., Mustafina, J.: Sign language recognition using deep learning. In: *2023 15th International Conference on Developments in eSystems Engineering (DeSE)*, pp. 184–189. IEEE (2023)
 34. Marais, M., Brown, D., Connan, J., Boby, A., Kuhlman, L.L.: Investigating signer-independent sign language recognition on the lsa64 dataset. In: *Southern Africa Telecommunication Networks and Applications Conference (SA TNAC)*. Rhodes University Grahamstown, South Africa (2022)
 35. Mistree, K., Thakor, D., Bhatt, B.: Towards performance improvement in indian sign language recognition. In: *Proceedings of the 17th international conference on natural language processing (ICON)*, pp. 349–354 (2020)
 36. Monteiro, C.D., Mathew, C.M., Gutierrez-Osuna, R., Shipman, F.: Detecting and identifying sign languages through visual features. In: *2016 IEEE international symposium on multimedia (ISM)*, pp. 287–290. IEEE (2016)
 37. Moryossef, A., Tsochantaridis, I., Aharoni, R., Ebling, S., Narayanan, S.: Real-time sign language detection using human pose estimation. In: *European Conference on Computer Vision*, pp. 237–248. Springer (2020)
 38. Müller, M., Ebling, S., Avramidis, E., Battisti, A., Berger, M., Bowden, R., Braffort, A., Camgöz, N.C., España-Bonet, C., Grundkiewicz, R., et al.: Findings of the first

- wmt shared task on sign language translation (wmt-slt22). In: Proceedings of the Seventh Conference on Machine Translation (WMT), pp. 744–772 (2022)
39. Müller, M., Jiang, Z., Moryossef, A., Gonzales, A.R., Ebling, S.: Considerations for meaningful sign language machine translation based on glosses. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 682–693 (2023)
 40. Nandy, A., Prasad, J.S., Mondal, S., Chakraborty, P., Nandi, G.C.: Recognition of isolated indian sign language gesture in real time. In: International conference on business administration and information processing, pp. 102–107. Springer (2010)
 41. National Geographic: Sign language (2023). URL <https://education.nationalgeographic.org/resource/sign-language/>. Updated October 19, 2023; Cited October 31, 2023
 42. Nayak, A.S., Subudhi, N., Rana, T., Sahu, M., Balabantaray, R.C.: Indain sign language recognition and translation into odia. In: Proceedings of the Workshop on Sign Language Processing (WSLP), pp. 6–13 (2025)
 43. Otsu, N., et al.: A threshold selection method from gray-level histograms. *Automatica* **11**(285-296) (1979)
 44. Prathap, K.B., Swaroop, G.D., Kumar, B.P., Kamble, V., Parate, M.: Islr: Indian sign language recognition. In: 2023 2nd International Conference on Paradigm Shifts in Communications Embedded Systems, Machine Learning and Signal Processing (PCEMS), pp. 1–6. IEEE (2023)
 45. Rodríguez, J., Martínez, F.: Towards on-line sign language recognition using cumulative sd-vlad descriptors. In: Colombian Conference on Computing, pp. 371–385. Springer (2018)
 46. Ronchetti, F., Quiroga, F., Estrebou, C., Lanzarini, L., Rosete, A.: Sign language recognition without frame-sequencing constraints: A proof of concept on the argentinian sign language. In: Ibero-American Conference on Artificial Intelligence, pp. 338–349. Springer (2016)
 47. Ronchetti, F., Quiroga, F.M., Estrebou, C., Lanzarini, L., Rosete, A.: Lsa64: An argentinian sign language dataset. arXiv preprint arXiv:2310.17429 (2023)
 48. Saunders, B., Camgoz, N.C., Bowden, R.: Progressive transformers for end-to-end sign language production. In: European Conference on Computer Vision, pp. 687–705. Springer (2020)
 49. Selvaraj, P., Nc, G., Kumar, P., Khapra, M.M.: Openhands: Making sign language recognition accessible with pose-based pretrained models across languages. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 2114–2133 (2022)
 50. Sridhar, A., Ganesan, R.G., Kumar, P., Khapra, M.M.: Include: A large scale dataset for indian sign language recognition. In: Proceedings of the 28th ACM International Conference on Multimedia, pp. 1366–1375 (2020)
 51. Stewart, G.W.: On the early history of the singular value decomposition. *SIAM review* **35**(4), 551–566 (1993)
 52. Sujatha, C., Jadi, P., Shubham, N., Habib, S.N., Chaitanya, U., Desail, P.: Improved indian regional sign language recognition with extended irksl dataset. In: 2024 6th International Conference on Computational Intelligence and Networks (CINE), pp. 1–6. IEEE (2024)
 53. Swamy, B.N., Lakshmi, T.S., Sowjanya, K., Himabindu, M., Lakshmi, K.V.V.: Indian sign language recognition using opencv and the mediapipe: A step towards inclusion. In: 2025 Second International Conference on Networks and Soft Computing (ICNSOC). IEEE (2025)
 54. Tavella, F., Schlegel, V., Romeo, M., Galata, A., Cangelosi, A.: Wlasl-lex: a dataset for recognising phonological properties in american sign language. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 453–463 (2022)
 55. Tong, Z., Song, Y., Wang, J., Wang, L.: Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems* **35**, 10,078–10,093 (2022)
 56. United Nations: International day of sign languages (2023). URL <https://www.un.org/en/observances/sign-languages-day>. Cited October 31, 2023
 57. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. *Advances in neural information processing systems* **30** (2017)
 58. World Health Organization: Deafness and hearing loss (2024). <https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss>
 59. Xiao, Q., Qin, M., Yin, Y.: Skeleton-based chinese sign language recognition and generation for bidirectional communication between deaf and hearing people. *Neural networks* **125**, 41–55 (2020)
 60. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Proceedings of the AAAI conference on artificial intelligence, vol. 32 (2018)
 61. Yenisari, E., Yavuz, S.: Deep learning based sign language recognition using efficient multi-feature attention mechanism. *IEEE Access* (2025)
 62. Yin, K., Moryossef, A., Hochgesang, J., Goldberg, Y., Alikhani, M.: Including signed languages in natural language processing. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 7347–7360 (2021)
 63. Yin, K., Read, J.: Attention is all you sign: sign language translation with transformers. In: Sign Language Recognition, Translation and Production (SLRTP) Workshop-Extended Abstracts, vol. 4. sn (2020)
 64. Zhang, X., Li, X.: Dynamic gesture recognition based on mmp network. *future internet* (2019)
 65. Zhou, H., Zheng, D., Nisa, I., Ioannidis, V., Song, X., Karypis, G.: Tgl: A general framework for temporal gnn training on billion-scale graphs. arXiv preprint arXiv:2203.14883 (2022)

A Hyperparameter Details

Table 9: Comprehensive summary of model implementation configuration and hyperparameters.

Parameter	Value
<i>Hardware</i>	
GPU Model	RTX 4090 (24GB)
CPU Model	i9-13900K
<i>Preprocessing & Split</i>	
Motion Threshold (τ_L)	25
ROI Padding Factor (λ)	1.2
SVD Sparsity Factor (α)	0.1
Max Sequence Length (T_{max})	25
Train / Val Split Ratio	80% / 20%
Random Seed	42
<i>Skeletal Augmentation</i>	
Unified Input Channels	8
Node Dropout Probability (p)	0.20
<i>Model Architecture</i>	
Spatial Subgraphs	Body, LH, RH
Transformer Layers (N)	2
Attention Heads (H)	8
TCN Kernel Size (K)	3
Pose Embedding Dim. (D_{pose})	512
<i>ArcFace Optimization</i>	
Angular Margin (m)	0.30
Hypersphere Radius (s)	30.0
<i>Training (AdamW)</i>	
Base Learning Rate (LR)	5×10^{-4}
Batch Size	32
LR Scheduler Patience	5
Weight Decay	0.5
Total Epochs	100