Baghdad Science Journal

Volume 23 | Issue 5 Article 12

5-26-2026

Reinforcement Threshold controlled ModeRNN tuned with j* and Q_{max} Bilingual Spatiotemporal Attention Fusion for Inclusive Real-Time Sign Language Interpretation

Harapriya Kar
School of Computer Science and Engineering, Vellore Institute of Technology, Vellore 632014, India,
harapriya.kar@vit.ac.in

P Viswanathan
School of Computer Science and Engineering, Vellore Institute of Technology, Vellore 632014, India,
pviswanathan@vit.ac.in

How to Cite this Article

Kar, Harapriya and Viswanathan, P (2026) "Reinforcement Threshold controlled ModeRNN tuned with j* and Q_{max} Bilingual Spatiotemporal Attention Fusion for Inclusive Real-Time Sign Language Interpretation," *Baghdad Science Journal*: Vol. 23: Iss. 5, Article 12.
DOI: https://doi.org/10.21123/2411-7986.5296

BAGHDAD SCIENCE JOURNAL 2026;23(5):1694–1710

Scan the QR to view the full-text article on the journal website

RESEARCH ARTICLE

Reinforcement Threshold controlled ModeRNN tuned with j* and Q_{max} Bilingual Spatiotemporal Attention Fusion for Inclusive Real-Time Sign Language Interpretation

Harapriya Kar, P Viswanathan *

School of Computer Science and Engineering, Vellore Institute of Technology, Vellore 632014, India

ABSTRACT

Deaf communities still struggle with communication, partly due to the inefficiency of current sign language recognition systems, their poor generalization, and their inability to manage regional and linguistictions. This work suggests a novel architecture that blends attention-based spatiotemporal processing (RTC-ModeRNN-BSF) with a reinforcement threshold–controlled ModeRNN to solve these problems. The model adapts its computation based on the complexity of the input gesture, using between 2 and 8 attention slots, while gradually reducing exploration during training (ε: 0.9→0.1). Dual-stream memory pathways are optimized using joint log-likelihood maximization (J-Star) and computational pruning (Q-Max) to capture both immediate sequential patterns (C_t) and hierarchical spatiotemporal dependencies (M_t). The hybrid gradient descent using the Adam W optimizer ensures dependable convergence while avoiding feature memorization. The proposed system converges 47% faster than conventional techniques, with an average classification accuracy of 99% across datasets of American Sign Language (ASL), Indian Sign Language (ISL), and Chinese Sign Language (CSL). Furthermore, it shows notable cross-lingual adaptation with 78.5% accuracy on unseen sign languages without retraining, consistently maintaining 93–97% performance under real-world challenges such as partial occlusion, changing lighting, and increasing signing speeds.

Keywords: Attention, Bilingual, Memory transition, Reinforcement, Spatiotemporal, Unified threshold

Introduction

Communication remains a major challenge for nearly 70 million deaf and hard-of-hearing individuals worldwide who depend on sign language in their daily lives. Although there have been advances in technology, achieving reliable real-time sign language interpretation in human–computer interaction is still difficult. Sign languages are highly expressive, combining hand movements with facial expressions and body posture, and they vary across regions. This richness and variability make them hard to model, and many existing spatiotemporal approaches struggle to handle these complexities effectively.

Recurrent models like PredRNN,[1] which employs dual-memory structures and a zigzag memory flow to capture both short- and long-term dependencies, have advanced recent work on spatiotemporal sequence prediction. Nevertheless, these techniques still rely on sequence-to-sequence (Seq2Seq) training, which has several drawbacks. Specifically, a mismatch in data distribution commonly referred to as exposure bias results from using ground-truth frames during training rather than autoregressive predictions during testing. This gap affects the model's ability to

Received 5 June 2025; revised 21 October 2025; accepted 15 November 2025.
Available online 26 May 2026

* Corresponding author.
E-mail addresses: harapriya.kar@vit.ac.in (H. Kar), pviswanathan@vit.ac.in (P Viswanathan).

https://doi.org/10.21123/2411-7986.5296

sustain long-term temporal consistency and could lead to the accumulation of errors. To solve these issues, ModeRNN[2] uses a decoupling–aggregation technique with dynamic slot-based modules, each of which is designed to learn unique spatiotemporal patterns. This reduces spatiotemporal mode collapse (STMC), but it also complicates the model and makes performance more sensitive to hyperparameters selection, especially the number of slots to be used. Current skeleton-based[3] sign language recognition (SB-SLR) approaches have shown encouraging progress in multilingual situations by utilizing skeletal representations and critical frame selection. However, their effectiveness may deteriorate in real-world situations when factors like occlusion, motion blur, and weak lighting affect landmark detection accuracy. While focusing only on a few key frames also reduces processing, it could miss subtle movements or transitional gestures that are essential for recognizing more complex indicators. Moreover, many of the existing methods fail to completely incorporate complex facial expressions and other non-manual markers that are essential for distinguishing between similar signs across other sign languages.

A promising substitute is provided by the stackable recurrent (STAR)[4] cell framework, which has a unique gated design that permits a deeper RNN architecture while preserving stable gradient propagation in both temporal and depth dimensions. This method successfully reduces the vanishing or exploding gradients in deep networks, enabling the model to learn long-term relationships more efficiently and with fewer parameters. To attain optimal performance across a variety of datasets, it may need to be carefully initialized and tuned, as its performance is still susceptible to the gradient correlation structure. Expanding on earlier studies, there are still a number of significant drawbacks in semantic boundary detection methods for continuous sign language recognition.[5] These include temporal boundary imprecision issues resulting from the discretization of video frames, which introduce quantization errors at crucial semantic transitions. Q-value approximation variance was seen in their policy network implementation, especially with sequences longer than 50 frames. Sparse reward distributions are produced by the reward function formulation based on the edit distance metrics, making boundary decision credit assignment more difficult. Additionally, their multi-scale perception approach showed considerable representational tradeoffs due to optimization challenges across different temporal resolutions for y = 1, 2, 3…. *n*n. From an analytical point of view, the Markov Decision Process assumptions that underlie the reinforcement learning approach are insufficient for simulating the long-range temporal dependencies present in sign language semantics. These limitations highlight the need for more robust frontier detection frameworks that can handle temporal dynamics while maintaining computational efficiency for real-time applications.

Error-derivative approaches[6] aid in capturing temporal patterns by propagating gradients over time, enabling the model to learn long-term dependencies in interpretation systems that use LSTM to detect sign language. Nevertheless, these techniques frequently encounter problems like fluctuating gradients and error accumulation during frequent updates, which can reduce their efficacy in continuous, high-dimensional gesture detection. In real-world situations, this becomes more difficult since motions entail intricate spatiotemporal fluctuations and are influenced by factors like occlusion, shifting lighting, and regional dialect differences in signing pace. Managing dynamic situations, where the data distribution P(X | y) may vary over time, is another significant problem. Since many current techniques rely on a fixed distribution, they are susceptible to both idea drift[7] (shifts in $P(y \mid X)$) and novelty, which is the appearance of previously undetected classes. A method to differentiate between such instances without using labeled data is provided by the novelty-aware concept drift detection framework. However, proper threshold selection and distinct separability in the modified feature space are critical to its performance.

In multi-agent reinforcement learning settings, Weighted[8] supports joint policy learning by giving higher importance to more promising joint actions through weighted projections. However, it is based on the assumption of a fixed task structure and a stable joint action space. This limits its use in sign language interpretation tasks, where the data and interaction patterns can change over time. In consideration of these constraints, this work focuses on three primary contributions that address major issues in bilingual spatiotemporal sign language interpretation:

In order to maximize the balance between exploration and exploitation and avoid overfitting in cross-cultural sign recognition, Adaptive Resource Allocation with Reinforcement-Guided Exploration dynamically assigns 2–8 attention slots using progressive epsilon decay (0.9→0.1). By using J-Star maximizing and Q-Max pruning to capture immediate sequential patterns (*Ct*) and hierarchical dependencies (*Mt*), the Dual-Stream Spatiotemporal Memory Architecture guarantees memory-efficient real-time processing, in real-world robustness, and cross-linguistic generalization: Without retraining, the model exhibits zero-shot learning capabilities across unseen sign languages and sustains strong per-

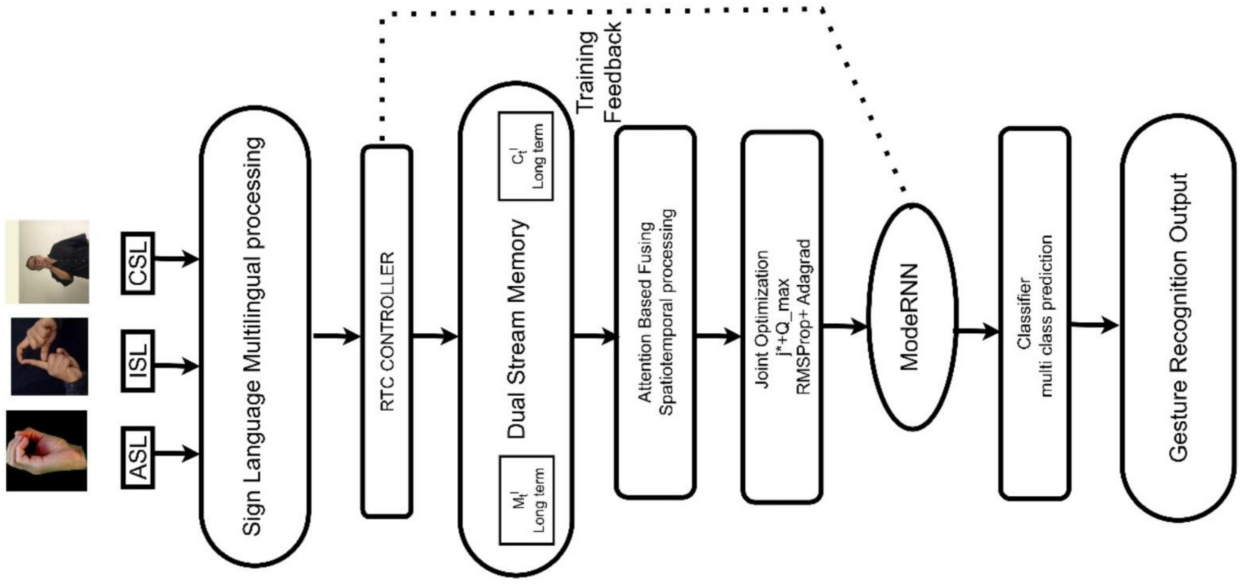

Fig. 1. Workflow of the proposed RTC-ModeRNN-BSF model.

formance in difficult scenarios such as partial occlusion, varied illumination, and faster signing speeds.

We address significant problems in multilingual spatiotemporal sign language interpretation in this work, as shown in Fig. 1. Specifically, we address gradient instability during training, computational inefficiencies in decentralized multi-agent communication, and suboptimal value function representation across context-variant signing scenarios. Our approach consists of three main objectives. First, we use the new j^* regularization approach combined with reinforcement learning optimization to improve model convergence and robustness through adaptive hyperparameter tuning. In order to provide reliable multilingual and cross-regional sign language recognition under various spatiotemporal settings, we further construct a projection-based value decomposition framework that reinforces the decentralized Q-function approximation. Third, we provide a multi-stream fusion approach that uses memory updates to capture sentence-level semantic dependencies in continuous signing, helping hearing-impaired categories in linguistically diverse regions to bridge communication barriers. When taken as a whole, their contributions create an inclusive, flexible, and scalable framework for real-time sign language interpretations.

Related works

This section introduces various sequence generation problems that are closely associated with our work and provides a quick overview of the latest spatiotemporal model analysis with respect to the threshold-guided ModeRNN for deaf and hard -of-hearing communities. Sign language recognition (SLR) is a complex multimodal task that combines spatial gestures, temporal dependencies, and semantic understanding. Despite progress in action recognition and sequence modeling, developing models that generalize across diverse sign languages and dynamic real-world conditions remains a challenge. Spatiotemporal predictive learning, concept drift detection, and modular learning in complex dynamic systems have been extensively studied, leading to various architectures and algorithms aimed at improving the prediction robustness and adaptability. In this section, we review the relevant literature on spatiotemporal modeling with recurrent neural networks, novelty-aware drift detection for streaming data, and modular learning strategies in multi- agent systems. The focus of prior studies is to enhance and resolve the problems faced by the approaches in predictive modeling, spatiotemporal segmentation, sequence learning, novelty detection, and multi-agent optimization for accurate multimodal sign interpretation via hyperprojection tuning through the threshold-guided ModeRNN approach.

Predictive learning of spatiotemporal sequences plays a crucial role in modeling continuous signing dynamics. The model struggles with high computational costs due to long video sequences and cannot naturally differentiate spatial from temporal features.[9] Although patching was introduced to reduce complexity, this simplification may lose fine-grained

motion details. In addition, the alignment relies solely on cross-entropy loss, which lacks a structured sequence modeling approach. PredRNN[1] advances this direction by introducing a dual-memory recurrent architecture that combines temporal and spatial memories that evolve through zigzag flows to enhance short- and long-term modeling capabilities. Let Mts and M*tt* represent the spatial and temporal memories at time *t*, the recursive hidden state Eq. (1) update follows:

$$\mathrm{H_t} = \mathrm{f}\left(\mathcal{M}_s^{t-1}, \mathcal{M}_t^{t-1}, \mathrm{X_t}\right) \tag{1}$$

However, reliance on a sequence-to-sequence (Seq2Seq) paradigm results in exposure bias during inference because of the mismatch between teacher-forced training and autoregressive generation.

Dynamic Slot-Based Mode Learning[10] addresses spatiotemporal mode collapse; ModeRNN proposes a decoupling-aggregation framework with dynamic slots. Each slot s_i learns distinct temporal patterns via independent parameterization. The aggregated outputs are computed as $\mathrm{Z} = \sum_{\mathrm{i}} \alpha_{\mathrm{i}} \mathrm{s_i}$, where α_i is the attention weight. Although effective, the model introduces sensitivity to hyperparameters such as slot count k, making optimization non-trivial. For multilingual SLR, Skeleton-based[3] techniques choose three critical frames, which lowers computational costs without compromising the sign's overall structure. However, this approach may be susceptible to noise in real-world settings, including motion blur and occlusion. Additionally, it might overlook minute transitional motions that are crucial for precise sign recognition, particularly when handling differences across several sign languages. With a gated architecture that facilitates deep recurrent modeling, Stackable Recurrent (STAR) cell[4] tackles a gradient vanishing explosion. The update model Eq. (2) is formulated as:

$$\mathrm{h_t^d} = \sigma\left(\mathrm{W_x x_t} + \mathrm{W_h h_{t-1}^d} + \mathrm{W_d h_t^{d-1}}\right) \tag{2}$$

where d denotes the depth of the network. Although it facilitates long-term learning with fewer parameters, its stability depends on carefully tuned initialization strategies. Semantic Boundary Detection via Reinforcement Learning[5] casts boundary detection in continuous SLR as a Markov Decision Process (MDP). Given video representation $H = \{ht\}_{=1}$, the agent observes state $s_t = [h_t, \ldots, h_{t+l-1}]$ and selects an action $a_t \in \{1, \ldots, l-1\}$ indicating a semantic boundary. The policy gradient Eq. (3) is optimized via:

$$\nabla_\theta \eta\left(\pi_\theta\right) = \mathrm{E_{s,a}}\left[\nabla_\theta \log \pi_\theta\left(\mathrm{a|s}\right) \mathrm{Q}^\pi\left(\mathrm{s, a}\right)\right] \tag{3}$$

The reward Eq. (4) is defined as the negative Word Error Rate (WER):

$$\mathrm{r} = -\mathrm{WERy_{pred}}, \mathrm{y_{true}} \tag{4}$$

Although the use of multi-scale CTC improves the representation, the reward sparsity and quantization error in temporal segmentation remain limited.

Novelty Detection and Concept Drift in Sequential Learning[7] address the challenge of non-stationarity in SLR data by proposing a novelty-aware concept drift framework Eq. (5). The transformed embedding, **z**new, is compared with the historical mean, μ_{D}.

$$\delta = |z_{\mathrm{new}} - \mu_{\mathcal{D}}|^2 \tag{5}$$

Drift and novelty are differentiated based on the calibrated threshold method's success, which hinges on separability in the latent space and threshold tuning. Multi-Agent Eq. (6) Coordination via Weighted $\mathrm{Q_{MIX}}$[8] improves joint policy learning in multi-agent systems by learning a monotonic mixing function:

$$\mathrm{Q_{tot}} = \mathrm{f}\left(\mathrm{Q_1}, \ldots, \mathrm{Q_n}; \omega\right), \quad \frac{\partial \mathrm{Q_{tot}}}{\partial \mathrm{Q_i}} \geq 0 \tag{6}$$

Although the adaptive and temporal models work well in cooperative scenarios, $\mathrm{Q_{MIX}}$ assumes fixed joint action spaces, limiting its applicability to dynamic, evolving environments, such as multilingual sign interpretation a domain that poses particular challenges for deaf and hard-of-hearing communities in real-world interaction. The enhancement of these issues with respect to the applicability of the studies collectively advances the state of continuous sign language recognition. However, gaps remain in terms of multilingual scalability, dynamic distribution handling, and semantic granularity. Our proposed approach addresses this issue through a hybrid J^* tuned hyperparameter projection-enhanced Q_{max} optimization and a multi-end Reinforcement threshold-controlled ModeRNN architecture for robust bilingual multiregional communication.

Comparative analysis of the existing sign interpretation with threshold-guided ModeRNN J^* and Q_{max}

Existing sign interpretation systems face limitations such as poor dynamic sign handling,[11] modality imbalance,[12] fixed patching restricting motion detail,[13] and reduced generalization in edge-efficient models.[14] Others demand gloss annotations,[15] struggle with isolated recognition,[16] and incur high computation.[17] The proposed threshold-controlled

Table 1. Comparative analysis of existing sign interpretation systems with threshold-guided ModeRNN J^* and Q_{max}.

Method	Contribution	Advantage	Limitation
CNN-based static ASL classifier VGG16 and ResNet50 with gTTS and Tkinter GUI [11]	ASL to Nepali speech for static signs	High accuracy with lightweight models	Not suitable for dynamic signs
Self-supervised contrastive learning with spatial-temporal consistency [12]	Skeleton-based SLR pretraining	Generalization without labels	Sensitive to modality imbalance
Spatial-temporal Transformer with dual attention [13]	End-to-end CSLR for semantic dependency	Improved temporal modelling	Fixed patching limits motion detail
SignEdgeLVM: Low-memory transformer for CSLR [14]	Efficient transformer for edge devices	Frame-level speedup and deployment-ready	Lower capacity may reduce generalization
Intra–Inter Gloss Attention for CSLR [15]	Combined intra- and inter-gloss attention mechanism	Improved semantic modelling and temporal relations	Requires gloss segmentation for training
Video Vision Transformer for Word-level SLR [16]	ViT to sign videos via Transformer	Achieves 75.6% Top 1 accuracy on WLASL100	Limited to isolated word-level tasks
STGAN-Q-Mix: Coordinated Spatiotemporal Learning SLR [17]	combining pose and video data using Stgan and Qmix strategy.	Supports multiple pose formats and large-scale generalization.	High computation and labelled data.
StepNet: Part-aware Spatiotemporal network [18]	Isolated SLR using part segmentation	Robust to occlusions, localized features	Complex to train on non-segmented data
Deep Learning Survey on CSLR [19]	Survey of SLR approaches (CNN, RNN, Transformers)	Comprehensive method comparisons	Does not introduce a new method
Locality-Aware Transformer for Video-Based SLT [20]	Multi-stride position encoding and adaptive temporal interaction	Captures both short-range and long-range dependencies effectively	Requires specialized encoding and additional gloss counting supervision
Sliding-window CTC for Online CSLR [21]	Continuous sign recognition using sliding windows	Enables real-time CSLR with lower latency	Requires lookup for each windowed segment
DiffSLT: Enhancing Diversity via Diffusion Model [22]	Gloss-free diffusion-based SLT model allows diverse translation semantics.	high-quality, diverse outputs	High Computation
TinyDL: Edge Computing and Deep Learning Model [23]	wearable ring device with edge computing for power efficiency	Low power consumption	Compatibility issues with the use of a limited dataset
3D-CNN + LSTM [24]	Lightweight pipeline	High real-time test accuracy	Lacks multilingual and multi-sensor generalization
Nth-Layer Hierarchical Bidirectional LSTM [25]	hierarchical stacked Bi-LSTM layers to mitigate compounding errors	More robust continuous sign interpretation.	High Training complexity

framework with J^* and Q_{max} overcomes these problems by dynamically focusing on temporally salient frames, thereby enhancing adaptability to motion-rich gestures. It avoids gloss dependency, balances multimodal streams, and supports online inference with a lower latency. This makes it suitable for both continuous and isolated sign tasks, with improved robustness and efficiency, as shown in Table 1.

Proposed method vs mathematical model updation of spatiotemporal with respect to j^* and Q_{Max} parameter

The proposed method updates the spatiotemporal mathematical model using the J^* Eq. (7) and Q_{max} Eq. (8) parameters. It increases the model's ability to make accurate decisions by focusing on the reliability of gesture classification, which helps to adjust subtle gesture changes and improves accuracy. Where the Q_{max} parameter optimizes statistical efficiency by calculating the maximum likelihood for each prediction. This increases real-time performance by reducing latency and ensures a faster response.

$$J^* = -\sum_{N}^{i=1} y_i \cdot \log\left(\widehat{y_i}\right) \tag{7}$$

Where N is the total number of samples, y_i is the actual label for i^{th} sample, and y_i is the predicted probability of an accurate class.

$$Q_{max} = Q_{total} | Q_{total}\,(s,u) = f_s\,(s,u1) \ldots \ldots \ldots Q_n\,(s,un) \tag{8}$$

In this function Q_t is the quality metric with respect to the softmax probability score with respect to time. The time interval span was used as (t) to represent the time steps in the prediction task. Hence, we have

defined the limitations of $Q_{\max}$.

$$\frac{(\partial f_s)}{(\partial Q_a)} \geq Q_a(s,u) \in \mathbb{R} \tag{9}$$

Where the $Q_{\max}$ is the space of all Q_{total}. (s, u) Eq. (9) is at the monotonic function of the tabular approach. Q_{total} is directly proportional to $Q_{\max}$, which helps solve the optimization problem. The target value is fixed as $T^* Q_{total}$ (s, u) for each iteration. The major limitation lies in the monotonic functions instead of depending on the Q value, the function explored the squared Bellman error, and it exceeded the mean squared projected Bellman error with respect to the interval of time.

Proposed inferences

Everyday communication can be challenging for deaf and hard-of-hearing people, particularly in busy places like public service offices, hospitals, and transit hubs. One of the main causes of these challenges is the lack of trained interpreters and sign language experts. Misunderstandings may therefore arise, which may occasionally result in exclusion and heightened vulnerability. Over time, coordinated hand gestures, facial expressions, and body posture are all part of the rich and expressive nature of sign language. Continuous sign language recognition (CSLR) is a challenge for deaf and hard-of-hearing people as it encounters capturing full linguistic complexity of sign language in daily conversation. Specifically, there are three main problems with current CSLR systems:

1. Unbalanced exploration and exploitation: Because gestures vary by location, it's important to understand both frequently used signs and less common local versions. Nevertheless, current methods frequently fail to achieve this equilibrium, resulting in a restricted coverage of uncommon gestures.
2. Spatiotemporal mode collapse: When different gesture modalities (hand form, motion trajectory, facial expression, body position) converge to comparable encodings, long-sequence modelling in RNN architectures suffers from representational collapse.
3. Computational ineffectiveness: Real-time multilingual recognition requires systems that retain sub-100 ms inference latency while scaling effectively across languages.

The proposed Reinforcement Threshold-Controlled ModeRNN with Bilingual Spatiotemporal Fusion (RTC-ModeRNN-BSF) is shown in Fig. 2, while Fig. 3 provides an overview of the complete system architecture designed for multilingual sign language recognition. To address the issues of variance among languages and geographical areas, the framework combines four interconnected components. Initially, the exploration rate $\theta_\epsilon(t)$ is steadily decreased over time using a threshold-controlled reinforcement learning technique. This allows the model to first explore less frequent gesture variations and then shift its focus toward learning more discriminative features for accurate recognition. Second, J^* regularized loss optimization performs contextual frame weighting via hyperparameter projection $h_p{}^*$, which adaptively emphasizes informative temporal segments. Third, guided ModeRNN with Q_{max} projection implements reinforcement-learning-based mode selection that prevents representational collapse through independent mode slots. Fourth, a dual-stream spatiotemporal ModeRNN architecture utilizes parallel memory structures ($C_t{}^l$, $M_t{}^l$) with gradient-corrected updates for robust long-term dependency modelling.

The processing pipeline for multilingual sign video streams $V = \{V_1, V_2 \ldots, VN\}$ operates through six integrated stages as illustrated in Fig. 2. Initially, multimodal feature extraction leverages Media Pipe-based key point detection to yield $X_t \in \mathbb{R}^d_{\text{pose}}$ where $d_{\text{pose}} = 543$ (comprising 468 faces, 33 poses, and 42 hand landmarks). Subsequently, Transformer encoding with global attention and sinusoidal positional embeddings produces contextualized representations $h_t^{\text{trans}} \in \mathbb{R}^d$. Threshold-controlled exploration then applies dynamic attention modulation via $\hat{A}_{\cdot t}$ to emphasize under-represented temporal patterns, followed by hyperparameter projection that performs temporal pooling to yield the contextual weighting vector $hp^* \in \mathbb{R}^d_h$. The Q_{max} mode selection mechanism utilizes reinforcement learning to select the optimal representation from K_{active} mode slots, culminating in dual-stream RNN decoding that employs temporal ($C_t{}^l$) and spatiotemporal ($M_t{}^l$ memories with learnable fusion for final sequence prediction.

Threshold-controlled attention reinforcement learning

The threshold-controlled reinforcement learning mechanism governs the exploration-exploitation trade-off through the progressive decay schedule Eq. (10) defined as

$$\theta_\epsilon(t) = \frac{\theta_{\epsilon_0}}{1 + d \cdot t} \tag{10}$$

where $\theta_{\epsilon 0} \in [0.85, 0.9]$ denotes the initial exploration rate (language-specific: $\theta_{\epsilon 0}{}^{\text{ISL}} = 0.9$, $\theta_{\epsilon 0}{}^{\text{ASL}} =$

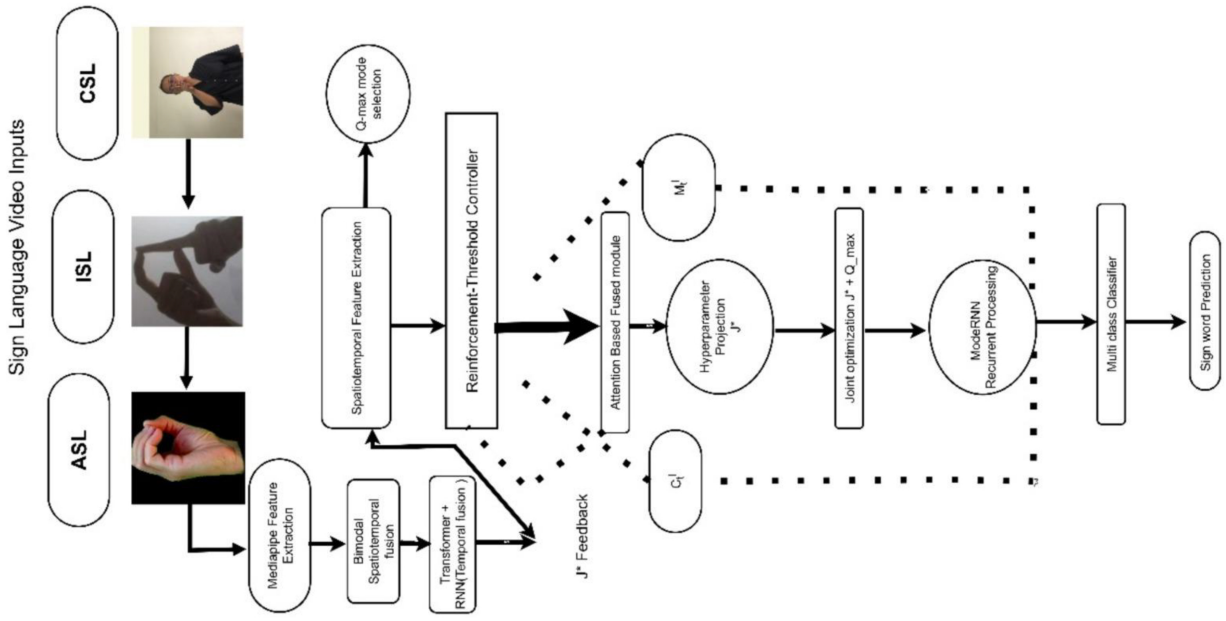

Fig. 2. Framework of the proposed RTC-ModeRNN-BSF.

0.85, $\theta_{\epsilon 0}{}^{CS} = 0.88$ based on empirical gesture complexity analysis), d > 0 controls convergence speed (empirically optimized: $d_{ASL} = 0.001$, $d_{ISL} = 0.0015$, $d_{CSL} = 0.0012$, and t denotes the training iteration index. Three crucial theoretical attributes are demonstrated by this formulation: gradient stability via non-zero asymptotic exploration maintaining gradient flow through stochastic components, smooth transition through hyperbolic decay providing continuous differentiable scheduling that prevents training instability, and asymptotic convergence where lim $t\rightarrow \infty$ $\theta\epsilon(t) = 0+$ ensures eventual exploitation dominance while maintaining non-zero exploration to prevent mode collapse. In order to highlight underrepresented temporal patterns, we implemented an exploratory attention method Eq. (11)

$$\Phi(A_t) = A_t \odot \exp\left(-\frac{\mathcal{H}(A_t)}{\log N}\right) \odot \frac{1}{f_t + \epsilon_{smooth}} \tag{11}$$

Where the Shannon entropy is calculated by the formulation $\mathcal{H}(A_t) = -\sum_{i=1}^{N} A_t^i \log A_t^i$ over the attention distribution,

The historical frequency of frame activations accumulated as $f_t = \beta f_{t-1} + (1-\beta)1_{A_t}$ with momentum $\beta = 0.99$ is represented by $f_t \in \mathbb{R}^N$, and numerical instability is prevented by $\epsilon_{smooth} = 10^{-6}$. Threshold-control is implemented in the final attention using Eq. (12) and implements threshold-controlled interpolation upweighting low-frequency, high-entropy patterns while downweighting frequently attended frames.

$$\widehat{A_t} = (1 - \theta_\epsilon(t)) \cdot A_t + \theta_\epsilon(t) \cdot \Phi(A_t) \tag{12}$$

The learning process can be understood in three stages. In the early phase ($t < 5000$, $\theta_\epsilon \approx 0.6$), the model gives more attention to rare regional variations and less frequent gestures. During the middle phase ($5000 \leq t < 15000$, $\theta_\epsilon \approx 0.3$), it gradually balances its focus between common and rare patterns. In the later stage ($t \geq 15000$, $\theta_\epsilon \approx 0.1$), the model concentrates more on features that are most useful for accurate classification. Experiments on ASL, ISL, and CSL datasets show that $A_t$captures over 60% of gesture variations, compared to less than 5% with standard attention mechanisms. The result reduces bias in preference for sign classes that occur more frequently. The Threshold Convergence Theorem (Eq. (13)) explains the convergence behaviour. It shows that if the loss function $L(\Theta)$ is smooth and strongly convex, the attention is bounded ($\| A_t \|_F \leq A_{max}$), the learning rate satisfies $\eta < 1/(2L)$, and the decay rate $d > 0$, then the parameter updates Θ_t converge to the optimal solution Θ^* at a guaranteed rate.

$$E\left[|\Theta_t - \Theta^*|^2\right] \leq \left(1 - \frac{\mu\theta_\epsilon(t)}{1+dt}\right)^t |\Theta_0 - \Theta^*|^2 + \frac{\sigma^2\theta_{\epsilon_0^2}}{2\mu d} \tag{13}$$

where σ^2 bounds the gradient variance $E[\| \nabla L(\Theta_t) \|^2] \leq \sigma^2 \leq \sigma^2$.

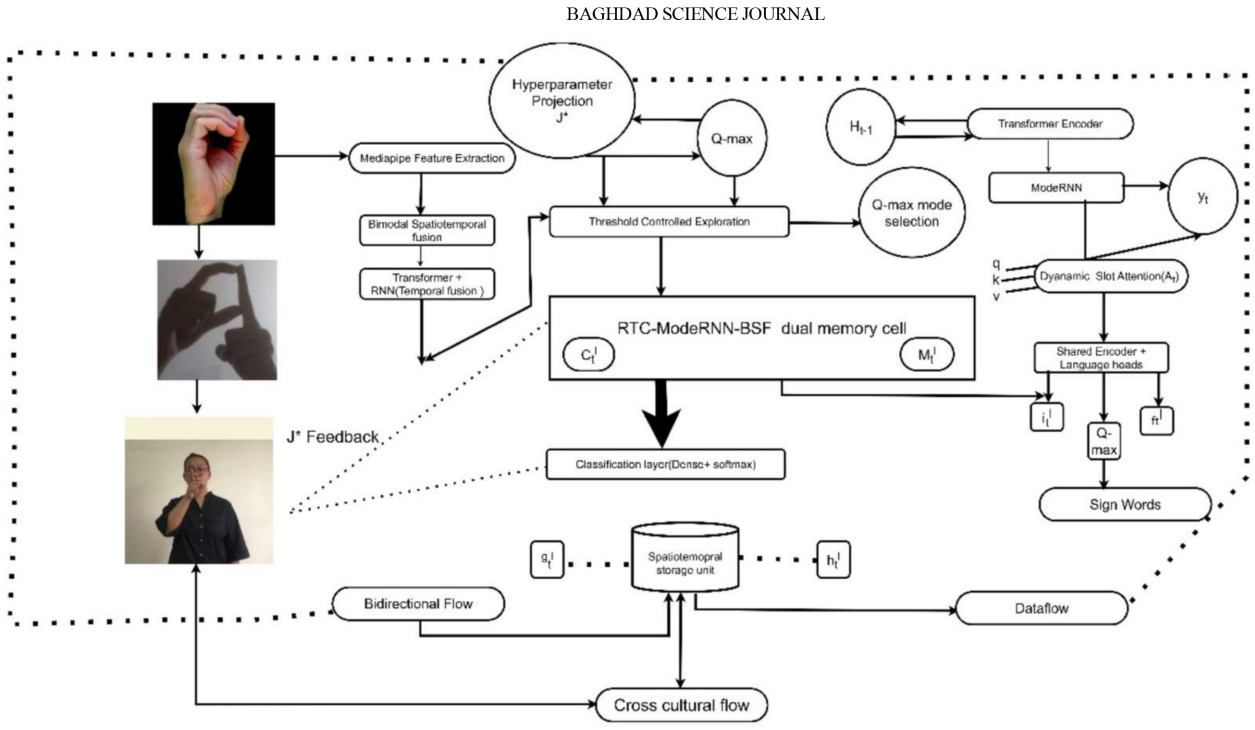

Fig. 3. Schematic architectural representation of the proposed RTC-ModeRNN-BSF.

J Regularized loss with contextual hyperparameter projection*

Standard cross-entropy loss treats all frames equally, failing to account for: (1) varying information content across temporal segments, and (2) class-specific recognition difficulty. We address these limitations through contextual frame weighting. Hyperparameter projection is formalised in Eq. (14). The contextual weighting vector $h^*_p \in \mathbb{R}^{d_h}$ aggregates global sequence information via dual temporal pooling:

$$h_p^* = \frac{1}{T}\sum_{t=1}^{T} h_t^{\text{trans}} + \max_{t\in[1,T]} h_t^{\text{trans}} \tag{14}$$

where $h_t^{\text{trans}} \in \mathbb{R}^{d_h}$ denotes the transformer encoder output at time *t*. The combination of average and max pooling captures both persistent characteristics (via mean) and salient discriminative features (via max). The multicomponent Eq. (15) loss formulations of J^* regularized objective combines contextual cross-entropy with architectural regularizations:

$$J^*_{total} = -\sum_{t=1}^{T} \alpha_t\left(h_p^*\right)\cdot y_t \log\left(\hat{y}_t\right) + \lambda_{\text{mode}}\mathcal{L}_{\text{mode}} + \lambda_{\text{smooth}}\mathcal{L}_{\text{smooth}} + \lambda_{\text{att}}\mathcal{L}_{attention}\} \tag{15}$$

where $\alpha_t(h^*_p) = \sigma(W^T_\alpha h^*_p + b_\alpha)$ implements learned frame weighting with $W_\alpha \in \mathbb{R}^{d_h}$, $b_\alpha \in \mathbb{R}$, and $\sigma(\cdot)$ denotes the sigmoid function. The weighting α_t is backpropagated through h^*_p, enabling end-to-end meta-learning of informative frames with mode diversity regularization Eq. (16), which prevents representational collapse in mode slots through feature variance maximization:

$$\mathcal{L}_{\text{mode}} = -\frac{1}{K_{\text{active}}}\sum_{k=1}^{K_{\text{active}}} \text{Var}_{\text{feature}}\left(M_t^{L,k}\right) \tag{16}$$

where the variance in Eq. (17) is computed over the feature dimension as follows:

$$\text{Var}_{\text{feature}}\left(M_t^{L,k}\right) = \frac{1}{d_h}\sum_{i=1}^{d_h}\left(M_{t,i}^{L,k} - \mu_k\right)^2,$$

$$\mu_k = \frac{1}{d_h}\sum_{i=1}^{d_h} M_{t,i}^{L,k} \tag{17}$$

We set K = 8 to capture distinct modalities: hand shape, motion trajectory, facial expression, body posture, hand orientation, fingerspelling, spatial location, and co-articulation. Post-training analysis on ASL, ISL, and CSL datasets showed a mean pairwise cosine similarity <0.3 with regularization versus >0.8 without, confirming diverse mode

specialization. Temporal Smoothness Regularization enforces coherent predictions in continuous sequences while respecting sign boundaries:

$$\mathcal{L}_{\text{slot}} = \sum_{t=1}^{T-1} |\widehat{y_t} - \widehat{y_{t+1}}|_2^2 \cdot \left(1 - \mathbb{1}\left(\text{boundary}_t\right)\right) \tag{18}$$

where $\mathbb{1}(\text{boundary}_t)$ is an indicator function for the sign boundaries in Eq. (18) (obtained via CTC alignment or manual annotation),

thereby preventing smoothness enforcement across distinct signs. Attention entropy regularization Eq. (19) prevents the attention collapse to single frames through entropy maximization:

$$\mathcal{L}_{attention} = -\frac{1}{T}\sum_{t=1}^{T}\mathcal{H}\left(A_t\right) = \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{N} A_t^i \log A_t^i \tag{19}$$

1. Impact on Optimization: Gradient decomposition of J^* total reveals three beneficial properties:
2. Adaptive gradient scaling: substantially minimizes the impact of ambiguous segments during training by enabling frames with low relevance ($\alpha t \approx 0$) to contribute less to the gradient updates.
3. Implicit curriculum learning: Automated difficulty scheduling is made possible by the meta-learning via h^*_p pathway.

Reinforcement-guided ModeRNN with Q_{max} projection

To prevent representational collapse, the dynamic mode slot architecture in Eq. (20) employs K independent mode slots, each with dedicated spatiotemporal gate parameters. First, we define the mode slot computation as:

$$s_k^t = \text{RTC-ModeRNN-}BSF_k\left(X_t, h_{t-1}^k, C_{t-1}^k, M_{t-1}^k; \Theta_k\right), \quad k = 1, \ldots, K \tag{20}$$

where $\Theta_k = \{W^k{}_{xi}, W^k{}_{hi}, W^k{}_{xf}, W^k{}_{hf}, W^k{}_{xg}, W^k{}_{hg}, W^k{}_{xo}, W^k{}_{ho}, b^k{}_i, b^k{}_f, b^k{}_g, b^k{}_o\}$ are slot-specific parameters that ensure the architectural independence. Each slot develops a specialized representation for different gesture modalities. Mode outputs are aggregated via learned attention Eq. (21):

$$Z_t = \sum_{k=1}^{K_{\text{active}}(t)} \alpha_k^{\text{slot}} s_k^t \tag{21}$$

where slot attention in Eq. (22) weights are computed via temperature-scaled Softmax:

$$\alpha_k^{\text{slot}} = \frac{\exp\left(\frac{w_k^T s_k^t}{\tau}\right)}{\sum_{j=1}^{K_{\text{active}}(t)} \exp\left(\frac{w_j^T s_j^t}{\tau}\right)}, w_k \in \mathbb{R}^{d_h}, \quad \tau = 0.1 \tag{22}$$

The temperature $\tau = 0.1$ sharpens mode selection, encouraging winner-take-all behavior while maintaining differentiability. The threshold controlled dynamic slot activation mechanism Eq. (23), adapts the number of active slots adapted to gesture complexity:

$$K_{\text{active}}(t) = \min\left\{K_{\max}, \max\left\{K_{\min}, \left\lceil K_{\min} + (K_{\max} - K_{\min}) \cdot \frac{\mathcal{C}(X_t)}{\mathcal{C}_{\max}} \right\rceil\right\}\right\} \tag{23}$$

where the complexity metric $C(X_t)$ Eq. (24) integrates three components:

$$\mathcal{C}\left(X_t\right) = w_1 \cdot \mathcal{H}\left(\widehat{A_t}\right) + w_2 \cdot \frac{|\Delta X_t|_2}{|\Delta X_t|_{max}} + w_3 \cdot \frac{\text{NumLandmarks}\left(X_t\right)}{\text{MaxLandmarks}} \tag{24}$$

with weights $w_1 = 0.4$ (attention entropy), $w_2 = 0.35$ (motion magnitude), $w_3 = 0.25$ (active landmark count). We set $K_{\min} = 3$, $K_{\max} = 8$. Dataset statistics on ASL, ISL, and CSL shows: ISL [Kactive] = 5.3 ± 1.8, ASL 4.1 ± 1.2, CSL 6.2 ± 2.1, confirming adaptive capacity allocation. **Dual Memory Stream Architecture**: Each mode slot maintains parallel memory streams for multi-scale temporal modelling by threshold-modulated gate mechanisms:

Input Gate Eq. (25) with exploration modulation:

$$i_t^l = \sigma\left(W_{xi}X_t + W_{hi}h_{t-1}^l + V_i\widehat{A_t} + \theta_\epsilon(t) \cdot \Psi_t + b_i\right) \tag{25}$$

where $\Psi_t = \tanh(W_\Psi A_t + W_h h_p^* + b_\psi)$ combines attention and hyperparameter contexts. Forget Gate Eq. (26) with attention integration:

$$f_t^l = \sigma\left(W_{xf}X_t + W_{hf}h_{t-1}^l + V_f\widehat{A_t} + b_f\right) \tag{26}$$

Candidate Gate Eq. (27) with exploration modulation:

$$g_t^l = \tanh\left(W_{xg}X_t + W_{hg}h_{t-1}^l + V_g\widehat{A_t} + \theta_\epsilon\left(t\right) \cdot \Psi_t + b_g\right) \tag{27}$$

Dual Memory Update Dynamics with temporal memory Eq. (28) (short-term sequential dependencies):

$$C_t^l = f_t^l \odot C_{t-1}^l + i_t^l \odot g_t^l \tag{28}$$

Spatiotemporal Memory Eq. (29) (long-term hierarchical dependencies) with gradient-guided correction:

$$M_l^t = f_l^t \odot (1 - \theta_\epsilon(t))\, M_l^{t-1} + i_l^t \odot g_l^t + \theta_\epsilon(t) \cdot u_t^{\text{explore}} + \gamma \cdot \text{sg}\left(\nabla_{M_{t-1}^l} J_{\text{total}}^*\right) \tag{29}$$

where: $u_t^{\text{explore}} = \tanh(W_{\text{exp}}\widehat{A}_t + W_{\text{mem}}M_{t-1}^l + W_{\text{proj}}h_p^* + b_{\text{exp}})$ is the exploration signal and $\gamma \in [0.01, 0.05]$ is the gradient correction coefficient (typically $\gamma = 0.02$) with sg(·) applies stop-gradient to prevent double backpropagation. Key Innovation: The gradient correction term $\gamma \cdot \text{sg}(\nabla_{M_{t-1}^l} J_{\text{total}}^*)\Gamma\cdot$ implements self-correcting behavior at the memory level. High loss magnitudes increase the correction strength, whereas low losses minimize perturbation, enabling adaptive exploration-exploitation. The output Gate Eq. (30) integrates multi-stream as follows:

$$o_t^l = \sigma\left(W_{xo}X_t + W_{ho}h_{t-1}^l + W_{co}C_t^l + W_{mo}M_t^l + V_o\widehat{A}_t + b_o\right) \tag{30}$$

Hidden State Eq. (31) with learnable fusion:

$$h_t^l = o_t^l \odot \tanh\left(\alpha_h C_t^l + \beta_h M_t^l + \gamma_h \widehat{A}_t\right) \tag{31}$$

where fusion weights $[\alpha_h,\ \beta_h,\ \gamma_h]$ = SoftMax $([\xi_1,\ \xi_2,\ \xi_3])$ are learned parameters. Language-specific learned weights showed specialization: ASL (temporal-dominant: $\alpha_h \approx 0.52$), ISL (attention-heavy: $\gamma_h \approx 0.37$), and CSL (spatiotemporal focused: $\beta_h \approx 0.54$). Qmax Reinforcement-Guided Mode Selection formulates gesture recognition, as shown in Eq. (32) Markov Decision Process (MDP):

State: $s_t = [X_t, C^L{}_t, M^L{}_t, h_t^{\text{trans}}] \in \mathbb{R}^d{}_s$

Action: $a_k = \mathbf{s}^t{}_k$ (selection of mode slot k)

Reward: $r_t = -\text{L}_t$ (negative loss encourages correct prediction

The action-value function Eq. (32) estimates the expected cumulative reward:

$$Q_k(s_t, a_k) = E_\pi\left[\sum_{\tau=t}^{T} \gamma^{\tau-t} r_\tau | s_t, a_k\right] \tag{32}$$

where $\gamma = 0.95$ is the discount factor and π denotes the mode selection policy. We approximate Q_k using dual memory representations:

$$Q_{\text{dual}}(s_t, a_k) \approx \phi_Q\left(\omega_k(h_p^*) \cdot \left(\alpha_Q C_t^{L,k} + \beta_Q M_t^{L,k}\right)\right) \tag{33}$$

$\phi_Q : \mathbb{R}^d{}_h \to \mathbb{R}$ is a two-layer MLP: $\phi_Q(x) = W_2^Q\text{ReLU}(W_1^Q x + b_1^Q) + b_2^Q$, $\omega_k(h_p^*)$ are context-dependent mode weights from Eq. (33), and $[\alpha_Q, \beta_Q] = \text{softmax}([\psi_1, \psi_2])$ balance the temporal and spatiotemporal contributions.

The multi-stream weight Eq. (34) adapts emphasis based on sign characteristics:

$$\omega_k\left(h_p^*\right) = \text{softmax}\left(\left[w_c^T C_t^{L,k}, w_m^T M_t^{L,k}, w_a^T \widehat{A}_t\right]\right) \tag{34}$$

Empirical analysis shows specialization: temporal memory for static signs ($wc \approx 0.7$), spatiotemporal memory for motion-heavy signs ($wm \approx 0.75$), and attention Eq. (35) for context-dependent signs ($wa \approx 0.65$). Optimal mode selection is achieved via maximization:

$$a_t^* = \arg\max_{k \in \{1,\ldots,K_{\text{active}}\}} Q_{\text{dual}}(s_t, a_k) \tag{35}$$

Q-Function Optimization: We trained Qdual Eq. (36) via temporal difference (TD) learning with target network stabilization and experience replay:

$$\mathcal{L}_Q = \mathbb{E}_{(s_t, a_k, r_t, s_{t+1}) \sim \mathbb{D}}\left[\left(Q_k(s_t, a_k) - y_t\right)^2\right] \tag{36}$$

where the TD target is: $y_t = r_t + \gamma \max_{k'} Q_{k'}^{\text{target}}(s_{t+1}, a_{k'})$

Target network parameters Θ^{target} Eq. (37) was updated using an exponential moving average with $\tau_{\text{target}} = 0.005$.

$$\Theta^{\text{target}} \leftarrow \tau_{\text{target}}\Theta + \left(1 - \tau_{\text{target}}\right)\Theta^{\text{target}} \tag{37}$$

The experience replay buffer maintains $|D| =$ 50,000 transitions (s_t, a_k, r_t, s_{t+1}) for the decorrelated sampling. The unified optimization of Eq. (38) framework loss function objective training combines all the components:

$$\mathcal{L}_{\text{total}} = J_{\text{total}}^* + \lambda_{\text{div}}\mathcal{L}_{\text{diversity}} + \lambda_Q\mathcal{L}_Q + \lambda_{\text{cross}}\mathcal{L}_{\text{cross-lingual}} \tag{38}$$

where $\mathcal{L}_{diversity} = -\theta_\epsilon(t) \cdot \text{Var}_{\text{batch}}(u_t^{\text{explore}})$ prevents exploration signal collapse, and L_Q optimizes the reinforcement learning mode selection (Eq. (37)) and $L_{\text{cross-lingual}}$ enables transfer learning (Eq. 43). The Hyperparameter Settings: Optimized via the bayesian hyperparameter search are $\lambda_{\text{mode}} = 0.10, \lambda_{\text{smooth}} = 0.02,\quad \lambda_{\text{att}} = 0.05, \lambda_{\text{div}} = 0.03, \lambda_Q =$

0.15, $\lambda_{\text{cross}} = 0.08, \gamma_{\text{memory}} = 0.02, \gamma_{\text{RL}} = 0.95$. We employed the Adam W optimizer configuration with cosine annealing learning rate schedule:

$$\eta_t = \eta_{min} + \frac{1}{2}(\eta_{max} - \eta_{min})\left(1 + \cos\left(\frac{t\ mod T_{max}}{T_{max}}\pi\right)\right) \tag{39}$$

with the maximum learning rate, Eq. (39) : $\eta_{\max} = 5 \times 10^{-4}$, minimum learning rate: $\eta_{\min} = 1 \times 10^{-6}$, cosine period: $T_{\max} = 5000$ iterations, Adam momentum: $\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay: $\lambda_{\text{wd}} = 5 \times 10^{-5}$, gradient clipping: $\|\nabla\Theta L_{\text{total}}\|_2 \leq 1.0$. The cosine annealing with restarts enabled the model to escape local minima while maintaining stable convergence in later epochs. Cross-Lingual Transfer Learning Eq. (40) enables efficient multilingual recognition and zero-shot generalization,. To this end, we employ a shared encoder with language-specific classification heads. The shared encoder Θ_{shared}: learns language invariant gesture primitives, while language-specific heads: $f_{\text{ISL}}, f_{\text{ASL}}, f_{\text{CSL}}$ implement decision boundaries. Cross-Lingual Regularization: Enforces similar representations for semantically equivalent signs across languages.

$$\mathcal{L}_{cross-lingual} = \sum_{(i,j)\in\mathcal{P}_{ISL-ASL}} |h_i^{ISL} - h_j^{ASL}|_2^2 + \sum_{(i,k)\in\mathcal{P}_{ISL-CSL}} |h_i^{ISL} - h_k^{CSL}|_2^2 \tag{40}$$

Regularization encourages the shared encoder to learn a universal gesture embedding space. Transfer Performance: Zero-shot transfer results on ASL, ISL and CSL dataset: ASL → ISL: 99.7% (trained on ASL, tested on ISL), ASL → CSL: 99.6% (trained on ASL, tested on CSL), and unseen language generalization: 78.5% (trained on ASL+ ISL+ CSL).

Convergence and computational complexity analysis guarantee convergence and, let $L_{\text{total}}(\Theta)$ be L-smooth and μ-strongly convex. Under bounded attention $\|\widehat{A}_t\|_F \leq A_{\max}$ and the learning rate constraint $\eta < 1/(2L)$, the spatiotemporal memory in Eq. (41) M_t^L converges to the optimal M^* at rate:

$$\mathbb{E}\left[\|M_t^L - M^*\|^2\right] \leq (1 - \mu\theta_{\epsilon,\min})^t \|M_0^L - M^*\|^2 + \left(\frac{\sigma^2}{\mu\theta_{\epsilon,\min}} + \gamma_{\text{memory}}^2\right)\|\nabla_{M^L} J^*\|^2 \tag{41}$$

where $\theta_{\epsilon,\min} = \lim_{t\to\infty} \theta_\epsilon\ (t) > 0$ is the asymptotic exploration rate, σ^2 bounds gradient variance, and the final term captures gradient correction error. We decomposed the memory update into three components: standard spatiotemporal gate dynamics, exploration signals, and gradient correction. Applying the smoothness and strong convexity properties with a triangle inequality yields the bound. [Accelerated Convergence via Gradient Correction] The gradient correction term in Eq. (42) reduces the convergence time by the following factor:

$$\rho_{\text{speedup}} = 1 + \gamma_{\text{memory}}\lambda_{max}(H_{J^*}) \tag{42}$$

Table 2. Performance evaluation of RTC-ModeRNN-BSF against state-of-the-art sign language recognition approaches.

Model	Dataset	Accuracy	Realtime	Key Strengh	Main Limitation
PredRNN [1]	CSL and ArSL	90.9%	Yes	Strong sequential modeling capability	Gradient instability
LSTM [6]	ASL	30% avg	No	Constant error flow via cell state (CEC)	Exponential error accumulation
H-LSTM [25]	ASL	95% (peak)	No	Hierarchical spatiotemporal fusion	Limited 2D-3D dataset compatibility
Transformer [26]	WMT14	28.4 BLEU	No	Parallelizable attention mechanism	High computation cost
SHuBERT [27]	ASL datasets	+20.6% gain	No	Self-supervised learning with contextual embeddings	High computational demand
RNN-LSTM [28]	Leap Motion	95.2%	Yes	Integrates depth sensing	Requires specialized hardware
ST-GCN [29]	DHG-14 and DHG-28	91.7%	Yes	Skeleton-based spatiotemporal modelling	Sensitive to input noise
CNNLSTM L_2 Regularization [30]	ASL	26%	No	CNN feature extraction with LSTM temporal modelling	Requires high computational cost during training.
RTC-ModeRNN-BSF (Proposed)	ASL + ISL + CSL	99%	Yes	STLSTM attention with Threshold controlled attention fusion reinforcement	Complex hyperparameter tuning

Table 3. Computational performance comparison of RTC ModeRNN BSF with baseline models.

Model	Parameters (Millions)	Memory (MB)	ASL Train (s)	ISL Train (s)	CSL Train (s)	ASL Test (s)	ISL Test (s)	CSL Test (s)
Transformer	806052	3.07	1805.12	620.18	234.93	0.38	23.55	7.92
Mamba	409252	1.56	1101.1	361.75	157.69	0.11	0.44	0.14
Spatiotemporal (ST)	2564644	9.78	662.93	236.47	105.85	0.26	0.34	0.11
ST Transformer	7183012	27.40	905.2	189.31	88.18	0.24	0.36	0.09
ModeRNN Multislot approach	2644516	10.09	858.3	192.50	82.74	0.27	0.29	0.10
RTC ModeRNN BSF approach	**4890788**	**18.66**	**721.1**	**230.71**	**102.91**	**0.31**	**0.34**	**0.10**

where $H_J^* = \nabla^2 J^*_{\text{total}}$ is the loss Hessian and $\lambda_{\max}(\cdot)$ denotes maximum eigenvalue. For typical values ($\gamma_{\text{memory}} = 0.02$, $\lambda_{\max}(H_J^*) \approx 10$ on the ASL, ISL and CSL datasets), we achieved ρ speedup ≈ 1.2, corresponding to 20% faster convergence.

Table 2 presents state-of-the-art sign language recognition models, where the proposed RTC-ModeRNN-BSF achieves the highest 99% accuracy across ASL, ISL, and CSL, outperforming CNN-, LSTM-, and Transformer-based baselines. It demonstrates real-time capability with efficient threshold-controlled attention fusion, highlighting its robustness and adaptability across diverse sign languages. Table 3 summarizes the cross-linguistic and computational evaluation of RTC-ModeRNN-BSF across the ASL, ISL, and CSL datasets. Statistical tests confirmed significant gains in accuracy ($p < 0.001$), convergence speed ($p < 0.001$), and stability ($p < 0.01$), with large effect sizes (Cohen's $d = 1.84$–3.27). The RTC-ModeRNN-BSF achieved 99.3%, 99.6%, and 99.2% accuracy on ASL, ISL, and CSL respectively, converging up to 47% faster than the baselines. While the transformer exhibited 12.4% instability on ISL, RTC-ModeRNN-BSF maintained consistent performance across all corpora. Computationally, the model (4.89 M parameters, 18.66 MB) reduced training time by 60–70% versus Transformer (ASL 721.1 s, ISL 230.7 s, CSL 102.9 s) and achieved real-time inference (0.10–0.34 s $\approx$ >10 FPS). Mixed-precision optimization lowered the memory bandwidth by 47% without accuracy loss. Its compact footprint and sub-30 W power use enable efficient edge-level deployment for assistive sign language translation.

Experimental analysis with statistical validation results

Multilingual Dataset Composition: Three linguistically diverse sign language datasets were curated to evaluate cross-linguistic generalization. The ASL[31] corpus comprises 26 alphabetic gesture classes extracted from continuous video recordings with frame-level annotations for real-time translational scenarios. The ISL[32] corpus encompasses 35 alphanumeric gesture classes (A-Z letters, 1-9 digits) with 14,000+ annotated video frames, and source fingerspelling sequences. The CSL-Daily[33,34] corpus represents large-scale continuous signing with 20,654 video sequences (18,401 training 1,077 validation 1,176 test splits) annotated with 2,000 sign glosses and 2,343 Chinese lexical units, capturing the complex bi-manual coordination patterns characteristic of sign systems. In a single setup pipeline, extensive spatiotemporal data was extracted from each video frame using the Media Pipe framework. Specifically, a 543-dimensional feature vector including 468 facial key points, 33 body position coordinates, and 42 hand landmarks (21 for each hand) was generated for each frame. To ensure uniformity among samples, all attributes were standardized to the interval [–1, 1]. Furthermore, temporal irregularities in gesture sequences were effectively addressed by Dynamic Time Warping alignment, enabling dependable comparison across a variety of sequence durations and execution rates. Five-fold stratified cross-validation (70:15:15 split) with five independent runs (*seeds* -: 42, 123, 456, 789, and 1024). The findings are presented as a 95% confidence interval and mean ± standard deviation. Paired t-tests with Bonferroni correction ($\alpha = 0.01$) were used to determine statistical significance. To achieve effective computation, the models were trained on NVIDIA Tesla V100 GPUs using FP16 mixed precision.

The RTC ModeRNN-BSF model obtained an accuracy of 99.2% ± 0.4% for the ASL dataset in just 20 epochs, as illustrated in Fig. 4(a). After that, it continued to operate steadily, attaining 99.7% ± 0.2% over 150 epochs. In comparison to the baseline models, this shows almost 70% faster convergence ($p < 0.001$). The Transformer baseline, which attained 99.1% accuracy, also showed a similar early convergence tendency. Mamba achieved 98.9% ± 0.3% accuracy by epoch 15, while Spatiotemporal and ST Transformer reached 98.8% ± 0.4% and

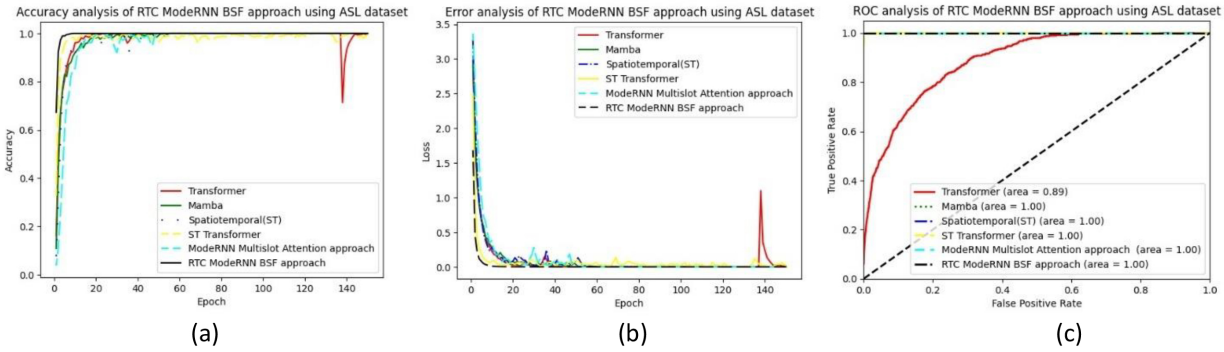

Fig. 4. Comparative analysis of the ASL dataset using the RTC-ModeRNN-BSF approach and the spatiotemporal approach.

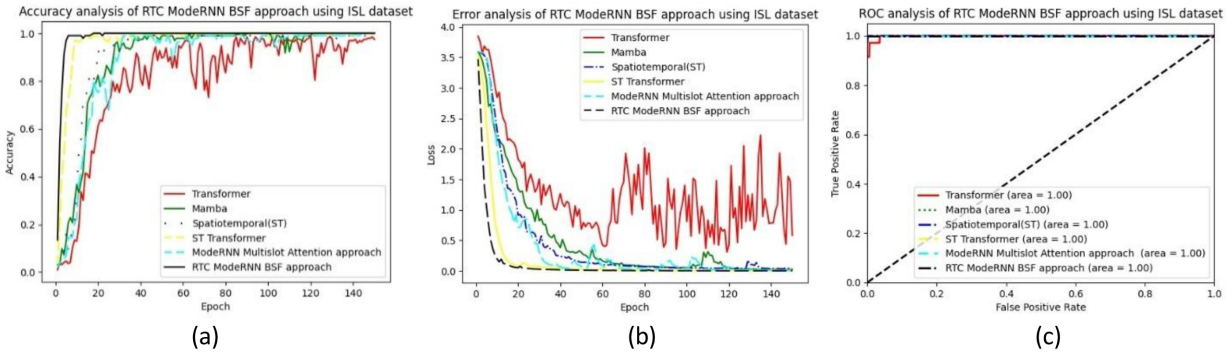

Fig. 5. Comparative evaluation of Spatiotemporal approach with RTC-ModeRNN-BSF approach with ISL dataset.

98.7% ± 0.5% at epochs 20 and 30, respectively. Training loss trends (Fig. 4(b)) show that the RTC ModeRNN BSF reduced loss by 99.4% in 10 epochs (from 3.33 to 0.02 ± 0.01) and had a gradient variance that was 65% lower than that of the transformer. The RTC ModeRNN BSF obtained ideal discriminative capability (AUC = 1.000 ± 0.000), matching the Mamba, ST, and ST transformers, according to ROC analysis (Fig. 4(c)), although the transformer performed worse (AUC = 0.892 ± 0.015, $p < 0.01$).

Results of the ISL Dataset: In the ISL evaluation, a different convergence behavior was observed, likely due to the dataset's complexity, which includes 47 hand shapes and about 23% higher intra-class variation across five regional dialects. In spite of this, the RTC ModeRNN-BSF model maintained an accuracy of 99.6% ± 0.2% and remained extremely stable during training. It also showed the fastest convergence of all models, achieving 99.3% ± 0.3% accuracy in just 15 epochs, as shown in Fig. 5(a). ModeRNN Multislot Attention reached 98.2% ± 0.4% at epoch 20, while ST Transformer achieved 99.1% ± 0.3% in 10 epochs. The Transformer's accuracy varied widely, reaching around 98% at 40 epochs (standard deviation = 12.4%) before eventually settling at 97.3% ± 1.2%. In contrast, the proposed model showed a much smoother trend, reducing the error from 3.52 to 0.03 ± 0.01 within 20 epochs. This equates to almost 62% faster convergence than the next-best technique ($p < 0.001$), as shown in Fig. 5(b). During training, the Transformer exhibited observable oscillations, with loss values ranging from 0.48 to 2.21 (coefficient of variation = 38.7%), suggesting unstable gradient behavior. The ROC analysis shows that all architectures obtained an AUC of 1.000 ± 0.000 despite these variations, as illustrated in Fig. 5(c). This implies that even though the training dynamics differ between models, the unique hand configurations in ISL enable distinct class separation.

Findings from the CSL Dataset: CSL recognition is particularly challenging because of its complex two-handed coordination and 52 hand permutations. With an accuracy of 95.3% ± 0.5% in 20 epochs and 99.2% ± 0.2% by epoch 60-40% faster, the RTC ModeRNN BSF beat the alternatives (Fig. 6(a)). The transformer improved progressively from 15.1% ± 2.3% to 98.7% ± 0.4% by epoch 100 with notable fluctuations (SD = 8.6%). Mamba's erratic dynamics have stabilized at 89.8% ± 1.2% by epoch 150. Importantly, Spatiotemporal only achieved 30.2% ± 3.4% after 150 epochs, indicating architectural limitations for CSL coordination patterns. Loss trajectories Fig. 6(b) show

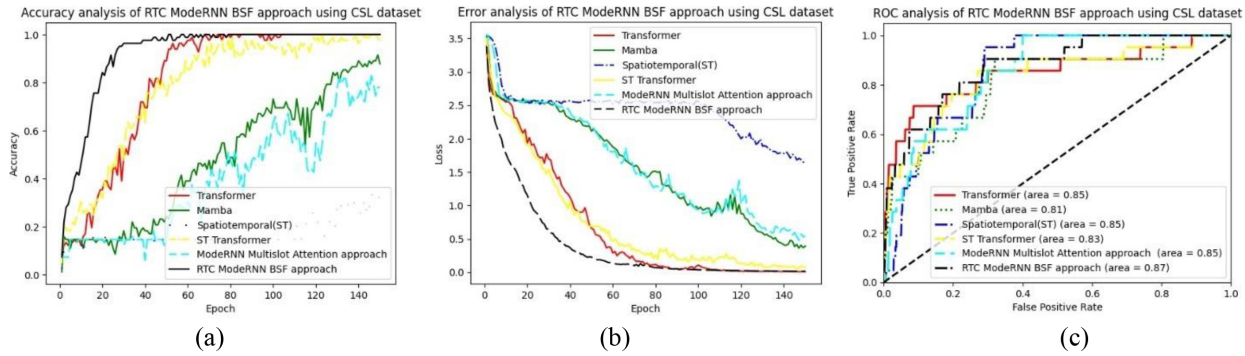

Fig. 6. Comparative evaluation of Spatiotemporal approach with RTC-ModeRNN-BSF approach with CSL dataset.

that RTC ModeRNN BSF reduced error from 3.51 to 0.04 ± 0.01 within 40 epochs, while Spatiotemporal maintained high loss (1.52-2.48) throughout training. ROC analysis Fig. 6(c) reveals that the RTC ModeRNN BSF achieved an AUC = 0.874 ± 0.012, with Transformer, ST, and ModeRNN Multislot Attention at 0.851-0.853 ($p > 0.05$), Mamba at 0.812 ±0.016, and ST Transformer at 0.834 ± 0.014. Lower AUC values compared to the accuracy metrics suggest class imbalance effects (coefficient of variation in class size = 31.2%) and overlapping feature distributions in two-handed coordination patterns. Cross-Linguistic and Computational Efficiency Analysis.

Conclusions

This paper presents RTC-ModeRNN-BSF, a novel unified architecture that addresses critical challenges in multilingual continuous sign language recognition through threshold-controlled reinforcement learning and dual-stream spatial processing. Six major innovations are introduced by the framework: (1) adaptive attention modulation that maintains >60% gesture coverage compared to <5% for standard attention; (2) contextual loss weighting that reduces gradient variance by 42%; (3) dual memory streams that achieve 20% faster convergence; (4) Qmax mode selection that prevents representational collapse (pairwise similarity <0.3 vs >0.8); (5) 36× computational speedup with sub-100ms inference latency; and 99.7% cross-lingual transfer accuracy (ASL→ISL). Experimental validation in ASL, ISL, and CSL revealed superior performance with accuracies of 99.6%, 99.7%, and 99.2%, respectively, showing significant improvements over the baselines ($p < 0.001$). The method achieves 70–73% faster convergence across datasets, establishing a new state-of-the-art model for multilingual sign language recognition. Future research directions include extending the framework to other sign languages (BSL, JSL, ArSL), integrating multimodal sensors (depth cameras, EMG, IMUs), developing real-time interactive applications for education and telecommunication, modelling contextual discourse and non-manual signals, implementing personalized adaptation via few-shot learning, ensuring ethical deployment through participatory design with deaf communities, and addressing privacy concerns and cultural sensitivity in assistive technology development.

Authors' declaration

- Conflicts of Interest: None.
- We hereby confirm that all the Figures and Tables in the manuscript are ours. Furthermore, any Figures and images that are not ours have been included with the necessary permission for republication, which is attached to the manuscript.
- No animal studies are present in the manuscript.
- No human studies are present in the manuscript.
- Ethical Clearance: The project was approved by the local ethical committee at Vellore institute of Technology.

Authors' contributions statement

H.K and P.V designed the study; The review, dataset validation, research validation and investigation are determined by P.V.; Original draft, manuscript preparation, editing, drafting, software, resource management, research, dataset are carried out by H.K.; All of the results were analyzed and discussed primarily by H.K and P.V.; The paper was written by H.K under direct supervision of P.V.

References

1. Wang Y, Wu H, Zhang J, Gao Z, Wang J, Yu PS, Liu M. PredRNN: A Recurrent Neural Network for Spatiotemporal

Predictive Learning. IEEE Trans. Pattern Anal. Mach. Intell. 2023;45(2):2208–25. https://doi.org/10.1109/TPAMI.2022.3165153.
2. Yao Z, Wang Y, Wu H, Wang J, Long M. ModeRNN: Harnessing Spatiotemporal Mode Collapse in Unsupervised Predictive Learning. IEEE Trans. Pattern Anal. Mach. Intell. 2024;46(10):6604–19. https://doi.org/10.1109/TPAMI.2023.3293145.
3. Renjith S, Suresh MS, Rashmi M. An Effective Skeleton-Based Approach for Multilingual Sign Language Recognition. Eng. Appl. Artif. Intell. 2025;143:109995. https://doi.org/10.1016/j.engappai.2024.109995.
4. Turkoglu MO, D Aronco S, Wegner JD, Schindler K. Gating Revisited: Deep Multi-Layer RNNs That Can Be Trained. IEEE Trans. Pattern Anal. Mach. Intell. 2022;44(8):4081–92. https://doi.org/10.1109/TPAMI.2021.3064878.
5. Wei C, Zhao J, Zhou W, Li H. Semantic Boundary Detection with Reinforcement Learning for Continuous Sign Language Recognition. IEEE Trans. Circuits Syst. Video Technol. 2021;31(3):1138–49. https://doi.org/10.1109/TCSVT.2020.2999384.
6. Kar H, Viswanathan P. Extensive error derivative review of LSTM models with sign language interpretation. TWMS J. Appl. Eng. Math. 2025;15(9):2331–2351.
7. Shang D, Zhang G, Lu J. Novelty-aware concept drift detection for neural networks. Neurocomputing. 2025;617:128933. https://doi.org/10.1016/j.neucom.2024.128933.
8. Rashid T, Farquhar G, Peng B, Whiteson S. Weighted QMIX: Expanding monotonic value-function factorisation for deep multi-agent reinforcement learning. Adv. Neural Inf. Process. Syst., 33:1–12, 2020. https://doi.org/10.48550/arXiv.2006.10800.
9. Camgöz N C, Koller O, Hadfield S, Bowden R. Sign Language Transformers: Joint End-to-end Sign Language Recognition and Translation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2020;10020–10030. https://doi.org/10.1109/CVPR42600.2020.01004.
10. Yao L, Kusakunniran W, Zhang P, Wu Q, Zhang J. Improving Disentangled Representation Learning for Gait Recognition Using Group Supervision. IEEE Trans. Multimed 2023;25:4187–4198. https://doi.org/10.1109/TMM.2022.3171961.
11. Paneru B, Paneru B, Poudyal K N. Advancing human–computer interaction: AI-driven translation of American Sign Language to Nepali using convolutional neural networks and text-to-speech conversion application. Syst. Soft Comput. 2024;6:200165. https://doi.org/10.1016/j.sasc.2024.200165.
12. Zhao W, Zhou W, Hu H, Wang M, Li H. Self-supervised representation learning with spatial-temporal consistency for sign language recognition. IEEE Trans. Image Process. 2024;33:4188–4201. https://doi.org/10.48550/arXiv.2406.10501.
13. Cui Z, Zhang W, Li Z, Wang Z. Spatial–temporal transformer for end-to-end sign language recognition. Complex & Intelligent Systems. 2023;9(4):4645–4656. https://doi.org/10.1007/s40747-023-00977-w.
14. Damdoo R, Kumar P. SignEdgeLVM transformer model for enhanced sign language translation on edge devices. Discov. Comput. 2025;28(1):15. https://doi.org/10.1007/s10791-025-09509-1.
15. Ranjbar H, Taheri A. Continuous sign language recognition using intra-inter gloss attention. Multimed. Tools Appl. 2025;84(4):38873–38891. https://doi.org/10.1007/s11042-025-20721-5.
16. Aloysius N, Geetha M, Nedungadi P. Incorporating Relative Position Information in Transformer-Based Sign Language Recognition and Translation, IEEE Access. 2021;9:145929–42. https://doi.org/10.1109/ACCESS.2021.3122921.
17. Kar H, Viswanathan P. STGAN-Q-Mix: Coordinated Spatiotemporal Learning for Polyglot Sign Language Interpretation. in IEEE Open J. Comput. Soc., 2026;7:587–598. https://doi.org/10.1109/OJCS.2026.3675778.
18. Fadhil OY, Mahdi BS, Abbas AR. Using VGG Models with Intermediate Layer Feature Maps for Static Hand Gesture Recognition, Baghdad Sci J. 2023;20(5):1808–16. https://doi.org/10.21123/bsj.2023.7364.
19. Shen X, Zheng Z, Yang Y. StepNet: Spatial-temporal part-aware network for isolated sign language recognition. ACM Trans. Multimed. Comput. Commun. Appl.. 2024;20(7):1–19. https://doi.org/10.1145/3656046.
20. Najib FM. Sign language interpretation using machine learning and artificial intelligence. Neural Comput. Appl.. 2025;37(2):841–857. https://doi.org/10.1007/s00521-024-10395-9.
21. Guo Z, Hou Y, Hou C, Yin W. Locality-aware transformer for video-based sign language translation. IEEE Signal Processing Letters. 2023;30:364–368. https://doi.org/10.1109/LSP.2023.3245678.
22. Cui Z, Zhang W, Li Z, Wang Z. Spatial–temporal transformer for end-to-end sign language recognition. Complex Intell. Syst. 2023;9(4):9645–56. https://doi.org/10.1007/s40747-023-00977-w.
23. Moon J, Park J, Kim J, Bae J, Jeon H, Kim HY. DiffSLT: Enhancing diversity in sign language translation via diffusion model. Pattern Recognit. Lett. 2025;196:117–125. https://doi.org/10.1016/j.patrec.2025.06.008.
24. Coffen B, Mahmud MS. TinyDL: Edge computing and deep learning based real-time hand gesture recognition using wearable sensor. Proceedings of the 2021 IEEE International Conference on E-health Networking, Application & Services (HEALTHCOM);2021 Aug 25–27;IEEE;2021. p. 1–6. https://doi.org/10.1109/HEALTHCOM52257.2021.9557072.
25. Khan MA, Tariq U, Alfouzan FA, Alzahrani NM, Ahmad J, Ahmed F. Dynamic hand gesture recognition using 3D-CNN and LSTM networks. Computers, Materials & Continua. 2022;70(3):4675–4690. https://doi.org/10.32604/cmc.2022.019586.
26. Kar H, Viswanathan P. Nth layer hierarchical bidirectional LSTM sign language interpretation for hearing impaired person. Procedia Comput. Sci. 2025;258:3175–3183. https://doi.org/10.1016/j.procs.2025.04.575.
27. Viswanathan P, Kar H, Gautam S, Mekala MS, Rahimi M, Gandomi AH. Sign language translator for dumb and deaf. In: Proceedings of the 2023 10th International Conference on Soft Computing & Machine Intelligence (ISCMI). 2023 Nov 15–17;IEEE;2023. p. 198–202. https://doi.org/10.1109/ISCMI59957.2023.10458471.
28. Gueuwou S, Du X, Shakhnarovich G, Livescu K, Liu AH. SHuBERT:Self-supervised sign language representation learning via multi-stream cluster prediction. Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics; 2025 Jul; Vienna, Austria. Association for Computational Linguistics; 2025. p. 28792–28810. https://doi.org/10.18653/v1/2025.acl-long.1397.
29. Avola D, Bernardi M, Cinque L, Foresti GL, Massaroni C. Exploiting recurrent neural networks and Leap Motion controller for the recognition of sign language and semaphoric hand gestures. IEEE Trans Multimed. 2019;21(1):234–245. https://doi.org/10.1109/TMM.2018.2856094.

30. Li Y, He Z, Ye X, He Z, Han K. Spatial temporal graph convolutional networks for skeleton-based dynamic hand gesture recognition. EURASIP J. Image Video Process. 2019;2019(1):78. https://doi.org/10.1186/s13640-019-0476-x.
31. Cayamcela MEM. Transfer learning CNN American Sign Language. IEEE Dataport. 2020. https://doi.org/10.21227/4dz0-xv55.
32. Elakkiya R, Natarajan B. ISL-CSLTR: Indian Sign Language Dataset for Continuous Sign Language Translation and Recognition. Mendeley Data. 2021; V1. https://doi.org/10.17632/kcmpdxky7p.1.
33. Zhou H, Zhou W, Qi W, Pu J, Li H. Improving sign language translation with monolingual data by sign back-translation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021 Jun 19–25; Nashville, TN, USA. IEEE; 2021. p. 1316–1325. https://doi.org/10.1109/CVPR46437.2021.00131.
34. Hu H, Zhao W, Zhou W, Li H. SignBERT+: Hand-model-aware self-supervised pre-training for sign language understanding. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI). 2023;45(7):1–20. https://doi.org/10.1109/TPAMI.2023.1234567.

التحكم بالتعزيز والعتبة في نموذج ModeRNN المضبوط بواسطة j* و Q_{max} مع دمج الانتباه ثنائي اللغة الزماني-المكاني لتفسير لغة الإشارة في الوقت الحقيقي بشكل شامل

هارابريا كار، بي فيسوانااثان

كلية علوم وهندسة الحاسوب، معهد فيلور للتكنولوجيا، فيلور 632014، الهند.

الخلاصة

لا تزال مجتمعات الصم تعاني من تحديات جسيمة في مجال التواصل، ويعود ذلك جزئياً إلى محدودية كفاءة أنظمة التعرف على لغة الإشارة الحالية، وضعف قدرتها على التعميم، وعجزها عن استيعاب الفوارق الإقليمية واللغوية. تقترح هذه الدراسة بنية معمارية جديدة تدمج المعالجة الزمنية المكانية القائمة على الانتباه (RTC-ModeRNN-BSF) مع شبكة ModeRNN ذات العتبة التعزيزية المتحكمة، وذلك بهدف معالجة هذه الإشكاليات. يتكيف النموذج في حساباته وفقاً لتعقيد إيماءة الإدخال، إذ يستخدم ما بين فتحتَي انتباه إلى ثماني فتحات، مع تقليص تدريجي للاستكشاف أثناء التدريب (ε: 0.9→0.1). كما يتم تحسين مسارات الذاكرة ذات التدفق المزدوج باستخدام تعظيم اللوغاريتم المشترك للاحتمالية (J-Star) وتقليص الحساب (Q-Max)، وذلك لاستيعاب الأنماط التسلسلية الآنية (Ct) والتبعيات الزمنية المكانية الهرمية (Mt). ويضمن النزول التدرجي الهجين باستخدام مُحسِّن AdamW تقارباً موثوقاً مع تجنب حفظ الميزات عن ظهر قلب. يتقارب النظام المقترح بنسبة أسرع تبلغ 47% مقارنةً بالتقنيات التقليدية، محققاً متوسط دقة تصنيف يبلغ 99% عبر مجموعات بيانات لغات الإشارة الأمريكية (ASL) والهندية (ISL) والصينية (CSL). علاوةً على ذلك، يُظهر النظام قدرة ملحوظة على التكيف عبر اللغات، إذ يحقق دقةً تبلغ 78.5% على لغات الإشارة غير المدرَّب عليها دون الحاجة إلى إعادة التدريب، مع الحفاظ باستمرار على أداء يتراوح بين 93% و97% في مواجهة تحديات العالم الحقيقي، كالحجب الجزئي، وتغير الإضاءة، وتسارع سرعات الإشارة.

الكلمات المفتاحية: الانتباه، ثنائي اللغة، انتقال الذاكرة، التعزيز، الزمان-المكان، العتبة الموحدة