International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 6, Nov-Dec-2026, ISSN (Online): 2395-566X

Silent Bridge: Real-Time Bidirectional Indian Sign Language and Kannada Translation with 3D Avatar

H S Thanush[1] Kiran M K[2] Manoj G U[3] Tejaswini K[4]
[1,2,3,]Dept. of Computer Science & Business Systems
B.M.S. College of Engineering Bangalore – 560019, India
[4]Assistant Professor & Project Guide Dept. of Computer Science & Business Systems
B.M.S. College of Engineering
Bangalore – 560019, India,

Abstract- Deaf and hard-of-hearing citizens in Karnataka communicate primarily through Indian Sign Language (ISL), a complete visual-spatial natural language, whereas public service counters, hospitals, banks, and schools operate in spoken and written Kannada. Almost no assistive software bridges ISL and Kannada bidirectionally on commodity hardware without privacy-invasive video uploads. This paper presents Silent Bridge, a real-time, privacy-preserving edge-AI assistive platform that (i) recognizes isolated ISL signs and continuous ISL sentences from a standard webcam and outputs grammatical Kannada text and speech, (ii) supports user-extensible few-shot custom sign enrollment in the browser via ONNX WebAssembly, and (iii) translates Kannada text into continuous ISL signing on a procedural 3D avatar. Hand and facial landmarks (147-D schema hands-face-v2) are extracted entirely in the browser via MediaPipe Tasks (v0.10.21) so raw camera frames never leave the client device. Trained exclusively on checksum-verified public ISL corpora (Amrita SLR, INCLUDE, and ISL-CSLTR) and evaluated strictly on unseen signers and unseen recording rounds, a Bidirectional GRU with 3-view Test-Time Augmentation (TTA) and selective HTTP-range INCLUDE expansion improved isolated-sign accuracy on 268 identical held-out videos from 59.7% (logistic regression baseline) to 79.9% (81.5% across all 314 held-out test videos, macro-F1 0.800), and on 213 videos of two completely unseen signers from 54.5% to 76.5% (exact paired McNemar p = 0.003). For continuous signing, a log-likelihood fusion of a 2-layer BiGRU-CTC decoder and a sentence classifier achieves 29.2% exact-sentence accuracy (GER 0.685) across 101 corpus sentences on unseen rounds. In the reverse direction, a multi-tier Kannada Subject–Object–Verb (SOV) grammar engine and an analytic two-bone Inverse Kinematics (IK) 3D avatar pipeline with Gaussian trajectory smoothing eliminated all 367 mid-sign arm drops across 144 dataset-derived clips and reduced wrist acceleration jitter by 36.0%.

Index Terms—Indian Sign Language (ISL), Kannada translation, MediaPipe landmarks, Bidirectional GRU (BiGRU), Connectionist Temporal Classification (CTC), ONNX WebAssembly, Inverse Kinematics (IK) signing avatar, assistive technology.

I. INTRODUCTION

Sign languages are independent natural languages possessing their own phonological, morphological, and syntactic structures; they are not word-for-word manual encodings of spoken languages [1], [2]. In India, over 63 million individuals experience significant auditory impairment according to World Health Organization estimates, and in Karnataka more than 300,000 deaf and hard-of-hearing citizens rely on Indian Sign Language (ISL) for daily communication. However, India has fewer than 400 certified ISL interpreters nationwide. When a deaf citizen visits a Primary Health Centre (PHC), a banking counter, a municipal office, or a classroom in Karnataka, a severe two-way linguistic barrier arises: the deaf citizen signs in ISL (which uses topic-comment spatial syntax, two-handed symmetric/asymmetric articulations, and facial Non-Manual Markers), while the hearing counter officer speaks and reads Kannada (a morphologically rich Dravidian language with Subject–Object–Verb constituent order).

Existing automatic Sign Language Recognition (SLR) literature

1.Problem Formulation

Given a commodity CPU-only laptop or public-service kiosk equipped with an ordinary 30 fps webcam (V = {I1, …, IN}), the engineering problem addressed by Silent Bridge comprises four coupled tasks:

© 2026 IJSRET

- **Privacy-Preserving Feature Extraction:** Map each RGB frame It in-browser to a scale-, translation-, and orientation-normalized landmark vector ft ∈ ℝ147 such that zero pixel data leaves the client device.
- **Signer-Independent Isolated & Continuous Recognition**: Segment active signing intervals [ts, te] from continuous live motion, classify isolated signs across C = 64 active classes (and 144 lexicon classes) with calibrated out-of-distribution (OOD)rejection threshold τ*, and decode unsegmented multi-sign streams over 101 continuous sentences (168 glosses).
- **Grammatical ISL-to-Kannada Meaning Translation**: Transform recognized ISL gloss sequences g = (g1, …, gk) into natural Unicode NFC Kannada sentences and synthesized kn-IN speech.
- **Reverse Kannada-to-ISL 3D Avatar Synthesis:** Parse Kannada text or speech into an ISL-ordered gloss plan with Non-Manual Markers (NMMs) and drive a 60 fps Three.js humanoid avatar with analytic two-bone Inverse Kinematics (IK) and zero mid-sign arm drops.

2. Key Technical Contributions

- **End-to-End Bidirectional Edge Platform:** A complete ISL ↔ Kannada system executing MediaPipe Tasks v0.10.21 and ONNX Runtime WebAssembly in-browser, backed by a deterministic FastAPI verification and training engine.
- **147-D Face-Relative Geometric Schema (hands-face-v2):** Combines palm-scaled 3D hand coordinates, image coordinates, face-relative vectors (with upper-center high-resolution face-crop fallback), orthonormal palm-normal vectors, and temporal velocity/path features (D = 3,668 for T = 24).
- **Selective HTTP-Range Archive Acquisition**: A bandwidth-frugal downloader that reads remote ZIP central directories via HTTP Range requests and fetches only target class videos with CRC-32 verification (downloading 5.96 GB instead of 44.7 GB of INCLUDE archives), boosting unseen-signer accuracy by +9.4 percentage points (p = 0.003).
- **Fused BiGRU-CTC + Sentence Classifier Decoder**: Combines frame-level CTC sequence log-likelihood with a 32-step sentence classifier (λ = 1.0), improving unseen-round sentence accuracy from 18.2% (greedy CTC) to 29.2% (GER 0.685).
- **Statistical Promotion Gate & Negative-Result Governance:** Enforces paired McNemar significance testing and non-regression gates on identical held-out videos, rejecting two over-parameterized or high-variance candidates (including a 144-class model that degraded unseen-signer accuracy from 76.5% to 70.9%).
- **Cleaned Dataset**-Retargeted 3D Avatar with Analytic IK:

II. RELATED WORK AND COMPARATIVE ANALYSIS

1. Sign Language Benchmarks and ISL Corpora

Large-scale isolated sign benchmarks such as WLASL (2,000 ASL words) [1] and MS-ASL (1,000 ASL words) [2], along with continuous corpora such as RWTH-PHOENIX-Weather [3], have catalyzed deep learning research for Western sign languages. For Indian Sign Language, Sridhar et al. [5] introduced INCLUDE (4,287 videos across 263 isolated signs recorded in Chennai), Elakkiya and Natarajan [6] released ISL-CSLTR (687 sentence videos across 101 sentences and 168 glosses), and Sasidharan et al.

[7] published the Amrita SLR dataset (isolated signs across 7 signer groups in Kerala). More recently, Joshi et al. introduced ISLTranslate [8] and iSign [9], providing large parallel video-to-English corpora without frame-level gloss annotations.

2. Appearance-Based vs. Landmark-Based Recognition

Appearance-based 3D CNNs (e.g., I3D [10]) and video Vision Transformers process raw RGB tensors directly, requiring gigabytes of GPU memory and exhibiting high sensitivity to background clutter, clothing, and lighting. Conversely, pose-based pipelines extract skeletal keypoints via OpenPose [11] or MediaPipe Hands/Face [12], [13], compressing each frame by over 103×. Skeleton-aware architectures such as SAM-SLR [14] and SignBERT

[15] demonstrate strong accuracy, while recurrent Gated Recurrent Units (GRUs) [16] and Bidirectional RNNs [17] trained with Connectionist Temporal Classification (CTC) [18] enable CPU-real-time sequence modeling on modest dataset sizes.

3. Signing Avatars and Identified Research Gaps

Traditional signing avatars synthesize motion from symbolic notation systems such as HamNoSys [19], which Wolfe et al. [20] note often appear robotic due to rigid piecewise-linear interpolation and missing Non-Manual Markers. Table I summarizes the validated ISL datasets used in this study, and Table II contrasts Silent Bridge against representative prior systems. Three critical gaps motivated our architecture: (i) the absence of a bidirectional ISL ↔ Kannada system, (ii) widespread signer leakage in random 80/20 evaluation splits in prior ISL papers, and (iii) severe mid-sign arm drops when retargeting raw 2D/3D MediaPipe landmarks to humanoid avatars.

© 2026 IJSRET

Table I
Public Isl Corpora Used (After Checksum & Integrity Validation) Table Ii

Dataset & License	Task Scope	Validated Content Used	Videos
Amrita SLR [7] (CC BY 4.0)	Isolated signs	49 signs, 7 signer groups (2 held-out)	933
INCLUDE [5] (CC BY 4.0)	Isolated signs	42 active classes (122 total), official split	816
ISL-CSLTR [6] (CC BY 4.0)	Sentences	101 sentences, 168 glosses, 5 rounds	678
Combined Active Corpus	**Bidirectional**	**64 active / 144 lexicon signs + 101 sents**	**2,427**

Comparative Analysis Of Existing Slr & Avatar Systems Vs. Silent Bridgethe Difficulty

Syste m / Study	Modality	Split Protoc ol	Kannada 3D Avatar		M
WLA SL I3D [1]	RGB 3D-CNN	Offic ial ASL	No	No	62.6% T
INCLU DE Baseline	Pose + BiLSTM	Offic ial ISL	No	No	85.6% (2
[5]					
ISL-CSLTR [6]	CNN-LSTM	Rando m 80/20	No	No	24.3 B
Amrit a SLR [7]	MediaPipe+ MLP	Rando m split	No	No	88.4% (
HamNo Sys Avatar [19]	Rule notation	N/A	No	Keyfram ed	Qualit
Silen t Brid ge (Our s)	147-D Edge BiGRU	Unse en Signe rs	Yes (SO V)	IK + NM M	81.5% / 76

III. DATASETS, SELECTIVE ARCHIVE PROTOCOL AND DATA INTEGRITY

1. Signer-ndependent Isolated Sign Protocol

The active isolated-sign model is trained on 1,749 validated videos across 64 ISL signs (49 from Amrita SLR, 42 from INCLUDE, with 27 overlapping signs present in both datasets). Fig. 2 illustrates the per-sign valid recording distribution (minimum 17, median 21, maximum 45 recordings per sign). Because Amrita SLR does not provide explicit signer IDs, signer identities were reconstructed from camera metadata and file-naming conventions into 7 disjoint signer groups. All recordings of two complete Amrita signers (213 videos) plus INCLUDE's official test split (55 initial, 101 expanded videos) form the strictly held-out test set (268 initial, 314 expanded videos). Model architecture and hyperparameters are selected exclusively via 3-fold signer-grouped cross-validation on the development set.

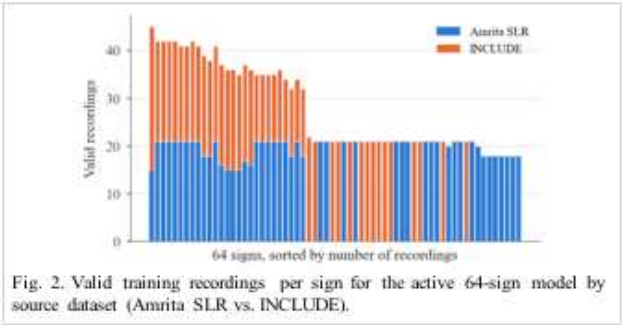


Fig. 2. Valid training recordings per sign for the active 64-sign model by source dataset (Amrita SLR vs. INCLUDE).

2. Selective HTTP-Range INCLUDE Archive Downloader

The INCLUDE corpus is hosted on Zenodo as monolithic multi-gigabyte category ZIP archives totaling 56.8 GB, whereas only specific vocabulary classes were required. Exploiting the ZIP specification—which stores the End-of-Central-Directory (EOCD) record in the final 64 KB of the archive—we engineered an HTTP Range-request downloader (Algorithm 1) that parses remote central directories, fetches only matching member byte ranges, and validates every extracted video against its header CRC-32 checksum. As detailed in Table IV (Page 3), this reduced network transfer from 44.7 GB to 5.96 GB (86.7% bandwidth savings) with zero corrupted files.

© 2026 IJSRET

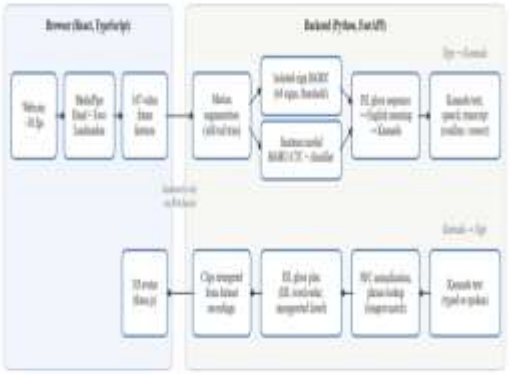


Fig 1. Silent Bridge end-to-end bidirectional architecture. Raw camera frames remain strictly inside the browser; only 147-D MediaPipe landmark vectors cross the privacy boundary. Top path: ISL video to Kannada text and speech. Bottom path: Kannada text/voice to continuous 3D avatar signing.

Algorithm 1: Selective HTTP-Range ZIP Archive Member Acquisition

Require: Remote archive URLs U, target vocabulary class paths W
Ensure: Checksum-verified local video files on disk
for each archive $u \in U$ containing classes in W **do**
  Fetch final 65,536 bytes of u via HTTP Range: bytes=−65536
  Locate EOCD signature 0x06054b50 and parse Central Directory o
  Fetch Central Directory byte range and enumerate member headers
  u
  for each member e whose normalized class folder $\in W$ **do**
    if target file already exists on disk with matching size **then** con tin
    Fetch local header + compressed payload bytes of e via HTTP Ran
    Decompress stream and compute CRC32(payload)
    if CRC32 $\neq$ e.crc32 **or** len $\neq$ e.uncompressed_size **then**
      Abort member write and log integrity failure
    else write atomically via .part temporary file rename
  end for
end for

Table Iv
Selective Http-Range Include Downloads & Bandwidth Savings

Acquisition Phase & Purpose	Videos	Fetched	Full ZIPs	Saved
Phase 1: Expand 24 core signs + 1	462	5.96 GB	44.7 GB	86.7%
Phase 2: 79 new signs + *home*	1,430	18.59 GB	48.7 GB	61.8%

3. Continuous Sentence Split and MD5 Integrity Audit

For continuous sentence recognition on ISL-CSLTR [6], recording rounds 4 and 5 form the held-out test set (154 videos), round 2 forms the validation set (73 videos), and rounds 1 and 3 form the training set (265 listed videos). A cryptographic MD5 audit of all 687 video files on disk (Table V) uncovered 5 duplicate byte-identical file groups, 4 of which carried conflicting sentence labels across folders; all 9 conflicted files were quarantined and excluded from training.

TABLE V
ISL-CSLTR CORPUS CRYPTOGRAPHIC MD5 INTEGRITY AUDIT

Audit Verification Metric	Video Count
Total raw video files discovered on disk	687
Videos listed in official corpus metadata spreadsheet	492
Unlisted videos with verified folder-name convention	195
Folder-name label agreement on listed videos	492 / 492 (100%)
MD5 byte-duplicate groups / conflicting sentence labels	5 groups / 4 conflicts
Corrupted/conflicted files excluded from training	**9**

IV. END-TO-END BIDIRECTIONAL SYSTEM ARCHITECTURE

As depicted in Fig. 1, Silent Bridge is architected around a strict client-side privacy boundary:

1) Forward Path (ISL Video → Kannada Text & Speech): The React 18 / TypeScript client captures 30 fps webcam video and executes MediaPipe Hand Landmarker (2 hands × 21 3D joints) and Face Landmarker (478 3D mesh points) via WebAssembly/WebGL. Only the extracted 147-D feature vector ft is processed by the motion segmenter and classified either in-browser via onnxruntime-web (sign_model.onnx) with few-shot cosine prototype fusion or streamed over WebSocket to the Python FastAPI backend for BiGRU-CTC continuous decoding. Accepted glosses are translated via a 187-entry phrase meaning table and SOV grammar rules into Unicode NFC Kannada text and spoken aloud via the Web Speech API (kn-IN) with a server-side gTTS fallback.

2) Reverse Path (Kannada Text/Speech → 3D ISL Avatar): Kannada input is normalized to Unicode NFC, stripped of agglutinative case suffixes (-alli, -inda, -ge, -annu), matched by longest-phrase priority, reordered into ISL Topic–Comment–Verb-Final constituent order, annotated with Non-Manual Markers (interrogative brow raise, wh-furrow, negation headshake), and rendered at 60 fps on a Three.js humanoid rig using dataset-retargeted keyframes and analytic two-bone Inverse Kinematics.

V.MATHEMATICAL FORMULATION AND METHODOLOGY

1. 147-DGeometric Landmark Schema (hands-face-v2) Let pi = (xi, yi, zi) ∈ ℝ3 denote the raw MediaPipe coordinates of hand joint i ∈ {0, …, 20}, where p0 is the wrist root and p9 is the middle-finger MCP joint. To achieve invariance to camera distance and hand size, we normalize all 20 non-wrist joints by the apparen

© 2026 IJSRET

$$p_i = (p_i - p_0)/(s_{palm}), \quad i \in \{1, \ldots, 20\} \Rightarrow 60 \text{ dims/hand}. \tag{1}$$

Many ISL signs are distinguished solely by hand height relative to the head (e.g., MOTHER vs. FATHER, or EAT vs. NOSE). Given the face bounding box center cface = (cx, cy) and dimensions (wface, hface), the face-relative wrist vector rface ∈ ℝ3 is:

$$r_{face} = \left[(p_{0,x} - c_x)/(\max(w_{face}, 10^{-3})), (p_{0,y} - c_y)/(\max(h_{face}, 10^{-3})), p_{0,z} - c_z \right]. \tag{2}$$

Three-dimensional hand rotation is captured via the unit finger-axis vector d̂ and unit palm-normal vector n̂:

$$\hat{d} = (p_9 - p_0)/(\|p_9 - p_0\|_2), \quad \hat{n} = ((p_5 - p_0) \times (p_{17} - p_0))/(\|(p_5 - p_0) \times (p_{17} - p_0)\|_2). \tag{3}$$

. Concatenating left and right hand descriptors (2 × 73 = 146) with the binary face-visibility indicator vface ∈ {0, 1} yields ft ∈ ℝ147 (Table III). A variable-length sign segment of L frames is temporally resampled via linear interpolation to T = 24 steps and augmented with inter-frame wrist velocities Δ p0,t ∈ ℝ6(T-1) and cumulative path lengths ℓL, ℓR ∈ ℝ2, yielding total segment dimension:

A

$$D = 147\,T + 6\,(T - 1) + 2 = 3{,}668 \quad (\text{for } T = 24). \tag{4}$$

Table Iii

147-D Landmark Feature Schema (Hands-Face-V2) Per Frame

Feature Sub-Block	Mathematical Definition	Dims
Wrist-relative shape	$(p_i - p_0)/s_{palm}$ for $i=1\ldots20$ (× 2 hands)	120
Image wrist position	Normalized $(p_{0,x}, p_{0,y}, p_{0,z})$ (× 2 hands)	6
Face-relative wrist	r_{face} scaled by face width/height (× 2 hands)	6
Palm orientation	Orthonormal $(\hat{d}, \hat{n}) \in \mathbb{R}^6$ (× 2 hands)	12
Visibility flags	Left hand v_L, right hand v_R, face v_{face}	3
Total Frame / Segment	$f_t \in \mathbb{R}^{147}$ per frame; $x \in \mathbb{R}^{3668}$ at $T=24$	**147**

BiGRU Classifier & 3-View Test-Time Augmentation

Each frame ft ∈ ℝ147 is projected through a LayerNorm-ReLU linear embedding to et ∈ ℝ96 and processed by a 2-layer Bidirectional GRU (96 hidden units per direction, ht ∈ ℝ192). Temporal mean pooling and max pooling are concatenated with summary wrist velocity statistics svel ∈

$$z = [(1/T)\textstyle\sum_{t=1}^{T} h_t \,\|\, \max_{1 \le t \le T} h_t \,\|\, s_{vel}] \in \mathbb{R}^{398}. \tag{5}$$

At inference, 3-view Test-Time Augmentation (TTA) averages probabilities over the raw segment $X^{(1)}$, a 3-tap temporally denoised segment $X^{(2)}$, and an 84% central core-stroke crop $X^{(3)}$:

$$\bar{P}(y \mid X) = (1/3)\textstyle\sum_{v=1}^{3} \mathrm{Softmax}(W_{cls}\, z(X^{(v)}) + b_{cls}). \tag{6}$$

3.Out-of-Fold Threshold Calibration & Live Segmentation

To reject non-sign gestures and ambiguous motions, we compute out-of-fold (OOF) cross-validation probabilities on all 1,435 development videos and select threshold τ* maximizing accepted-word F1(τ)

$$\tau^* = \arg\max_{\tau \in [0.05,\, 0.95]} (2\,\mathrm{Prec}(\tau)\,\mathrm{Rec}(\tau))/(\mathrm{Prec}(\tau) + \mathrm{Rec}(\tau)) = 0.50. \tag{7}$$

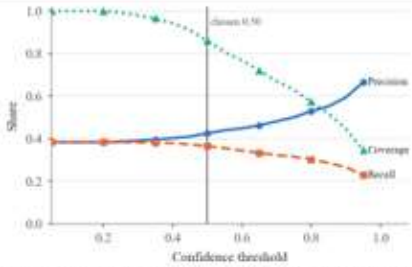


Fig. 3. Precision, recall, and coverage of accepted signs against confidence threshold τ across 1,435 out-of-fold development predictions; $\tau^* = 0.50$ maximizes F1.

Algorithm 2: Motion Segmentation with Still-Tail Trimming

for each incoming frame f_t with inter-frame wrist speed m_t **do**
  if state = IDLE **and** hands present **and** $m_t > \theta_{start}$ $(0.014\ s^{-1})$ **then**
    state ← ACTIVE; initialize buffer $B \leftarrow [f_t]$
  else if state = ACTIVE **then** append f_t to B
    if ($m_t < \theta_{still}$ for T_{still}) **or** hands absent ≥ 200 ms **or** $|B| > 4$ s **then**
      Trim trailing stillness tail where $m_k < \theta_{still}$; classify B via Eq. (6)
      if $\max_y \bar{P}(y) \ge \tau^*$ **then** emit Kannada word; **else** mark uncertain
    end if
  end if
end for

4. Continuous Sentence Recognition via BiGRU-CTC Fusion

For continuous unsegmented ISL sentences, a 2-layer BiGRU (128 units per direction) processes every second frame of sequence X and outputs per-step probabilities over 168 glosses plus the CTC blank token ε [18]. Because pure CTC decoding can produce ungrammatical gloss combinations on small corpora, we fuse the exact CTC prefix log-likelihood of each candidate corpus sentence s

∈ {1, …, 101} (with gloss sequence gs) with a sentence-level

classifier Pcls(s | X) operating on X resampled to 32 steps:

$$\text{score}(s) = \log P_{\text{CTC}}(g_s \mid \boldsymbol{X}) + \lambda \log P_{\text{cls}}(s \mid \boldsymbol{X}), \quad \lambda = 1.0. \tag{8}$$

Sentences whose posterior Softmax(score(s)) < 0.50 are flagged as uncertain with top-3 candidates rather than silently misclassified.

5. Block-Balanced Few-Shot Custom Sign Enrollment

To allow deaf users and domain officers to enroll custom regional signs in 3–5 webcam repetitions without retraining the neural network, we compute a block-balanced cosine similarity across B = 5 semantic feature subspaces (left handshape, right handshape, face-relative position, palm orientation, and velocity trajectory) so the 120-D handshape block does not drown out the 6-D face-relative location cue:

$$S_{\text{block}}(\boldsymbol{u}, \boldsymbol{v}) = \sum_{b=1}^{5} w_b \, \langle \boldsymbol{u}^{(b)}, \boldsymbol{v}^{(b)} \rangle / (\|\boldsymbol{u}^{(b)}\|_2 \, \|\boldsymbol{v}^{(b)}\|_2 + 10^{-8}), \quad \sum_{b=1}^{5} w_b = 1. \tag{9}$$

6. 3D Avatar Clip Cleaning & Analytic Two-Bone Inverse Kinematics

Given upper-arm length L1 = 0.29 m, forearm length L2 = 0.27 m, shoulder position ps, and target 3D wrist position pw retargeted from dataset landmarks (with clamping d = min(‖pw - ps‖2, 0.995(L1 + L2))), the interior elbow flexion angle θelbow is solved analytically via the Law of Cosines:

$$\cos\theta_{\text{elbow}} = \text{clip}\left(\frac{L_1^2 + L_2^2 - d^2}{2 L_1 L_2}, -1, 1\right). \tag{10}$$

Consecutive signs in a sentence are joined by a 0.20 s minimum-jerk Quintic Hermite spline h(u) = 10u3 - 15u4 + 6u5 (u ∈ [0, 1]) with an outward/upward ballistic clearance arc Δparc(u) = 4u(1-u)(0,

0.035, 0.045) m to prevent torso mesh self-intersection

$$\boldsymbol{p}_w(u) = (1 - h(u))\, \boldsymbol{p}_{\text{end}}^{(k)} + h(u)\, \boldsymbol{p}_{\text{start}}^{(k+1)} + \Delta \boldsymbol{p}_{\text{arc}}(u). \tag{11}$$

Algorithm 3: Avatar Landmark Trajectory Cleaning (Per Hand)

Require: Raw per-frame hand poses $\boldsymbol{p}_1, \ldots, \boldsymbol{p}_n$ (or missing)

1: **if** hand is present in < 15% of clip frames **then return** no track (detector noise)

2: Linearly interpolate tracking dropout gaps of duration ≤ 0.40 s between valid frames

3: Apply 1D Gaussian smoothing (σ = 1.5 frames) to wrist $\boldsymbol{p}_w$, $\boldsymbol{d}$, $\boldsymbol{n}$, and 15 finger curls

4: Trim leading and trailing motionless frames (retaining 0.12 s onset/offset context)

5: Resample trajectory to at most 15 keyframes/s and serialize compact JSON clip

7. Paired McNemar Test & Non-Regression Promotion Gate

Every candidate model is evaluated against the active production model on the exact same held-out test videos. Let n10 denote videos where only the new candidate is correct and n01 denote videos where only the active model is correct. Statistical significance is evaluated via the exact two-sided McNemar test [21]:

$$p = 2 \sum_{k=\min(n_{10}, n_{01})}^{n_{10}+n_{01}} C(n_{10}+n_{01}, k)\, 0.5^{n_{10}+n_{01}}. \tag{12}$$

A candidate is promoted to active status only if (i) its held-out macro-F1 ≥ 0.40, (ii) its accuracy on shared test videos does not regress by more than 0.02, and (iii) it passes an automated live-pipeline smoke test.

VI. FULL-STACK IMPLEMENTATION AND EDGE-WASM ENGINE

Silent Bridge is implemented as a modular 7-view web platform (Live Recognition, Custom Signs Studio, Kannada-to-Sign 3D Avatar, Sign Reference Dictionary, Dataset & Model Governance Registry, About & Testing Benchmark Suite, and CSBS/IEEE Report Viewer).

1) Client-Side ONNX WebAssembly Inference: The trained PyTorch BiGRU is exported to ONNX opset 17 (sign_model.onnx, 3.1 MB) and executed in the browser via onnxruntime-web WASM SIMD threads. When offline or disconnected from the Python server, the browser executes the full 6-stage pipeline—MediaPipe landmarking, 147-D feature extraction, Algorithm 2 segmentation, 3-view TTA ONNX inference, few-shot custom prototype matching, and Kannada SOV translation—locally on the client CPU.

2) Sovereign Kannada Linguistic Engine: In the forward direction, recognized gloss sequences are matched against a 187-entry phrase meaning table that resolves ISL topic-comment syntax into natural Kannada sentences (e.g., [YOU, NAME, WHAT] → ನಿನ ಹೆಸರು ಏನು?). In the reverse direction, Kannada input undergoes Unicode NFC canonicalization, regex agglutinative suffix stripping, longest-phrase matching across 162 vocabulary entries, SOV-to-ISL reordering (placing temporal nouns first and wh-interrogatives/negations at the clause boundary), and procedural Kannada finger-spelling for out-of-vocabulary proper nouns. Expressive 3D Avatar Rig: Built in Three.js (AvatarCanvas.tsx), the humanoid character features anatomical proportions, dual-hemisphere studio lighting, 55-joint hierarchical skeleton (hips, spine, neck, head, clavicles, upper/forearms, wrists, and 30 phalangeal finger joints), 6 Kannada mouth visemes, blink controller, and NMM brow/head tracks

© 2026 IJSRET

VII. EXPERIMENTAL EVALUATION AND RESULTS

1. Isolated-Sign Recognition on Unseen Signers

Table VI and Fig. 4 report the progression of isolated-sign top-1 accuracy on the identical 268 held-out test videos across four major engineering iterations. Adding temporal BiGRU modeling and upper-center face-crop fallback raised accuracy from 59.7% (multinomial logistic regression baseline) to 72.4%. Expanding the training set with 462 selectively downloaded INCLUDE recordings (Algorithm 1) further boosted accuracy on the same 268 videos to 79.9% (and 81.5% across the full 314-video test set, macro-F1 0.800). On the most challenging subset—213 videos recorded by two completely unseen Amrita signers—accuracy rose from 54.5% to 76.5% (+22.0 percentage points overall; +9.4 points from data expansion alone, with n10=34 newly correct vs. n01=13 regressed, exact paired McNemar p = 0.003).

Table Vi

Isolated-Sign Top-1 Accuracy (%) On Identical Held-Out Videos

Model Iteration (Train Videos)	All (268)	Amrita Unseen (213)	INCLUDE Test (55)	Macro-F1
LogReg Baseline (1,290)	59.7	54.5	80.0	0.591
LogReg + BiGRU *T=16* (1,290)	72.4	67.1	92.7	0.708
LogReg + BiGRU *T=24* (1,290)	72.0	67.1	90.9	0.707
BiGRU + INCLUDE + TTA (1,749)	**79.9**	**76.5**	**92.7**	**0.800***

**Macro-F1 on full 314-video held-out test set (overall Top-1 accuracy: 81.5%).*

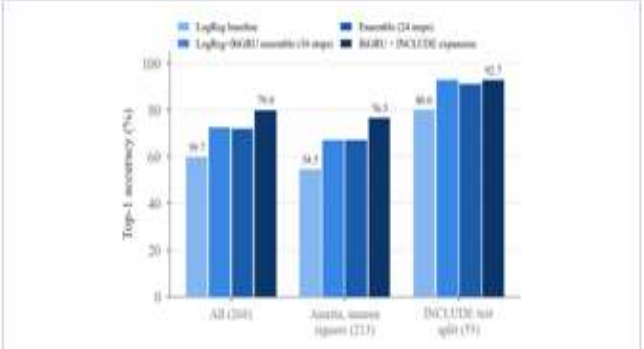


Fig. 4. Top-1 accuracy progression of successive models on identical held-out test videos (all 268 videos vs. 213 unseen Amrita signer videos).

2. Validation Ablation & Multi-Seed Stability

Table VII summarizes the 3-fold signer-grouped validation comparison used to select temporal length T and classifier weighting prior to test evaluation. Retraining the baseline and the BiGRU ensemble across three random seeds (7, 11, 23) yielded 62.2 ± 2.1% vs. 70.5 ± 1.3% test accuracy (n10=102 vs. n01=35 pooled, McNemar p < 10-8), confirming seed-invariant improvement.

TABLE VII
3-FOLD SIGNER-GROUPED VALIDATION ABLATION (DEVELOPMENT DATA)

Architecture & Resampling Variant	Val Acc	Val Macro-F1
Multinomial LogReg, *T = 12*	0.398	0.377
Multinomial LogReg, *T = 16*	0.412	0.388
Multinomial LogReg, *T = 24*	0.413	0.385
Multinomial LogReg, *T = 32*	0.416	0.391
Standalone BiGRU, *T = 16*	0.416	0.399
Ensemble (0.7 LogReg + 0.3 BiGRU)	0.424	0.402
Ensemble (0.5 LogReg + 0.5 BiGRU, selected)	**0.437**	**0.420**

Fig. 5 plots the sorted F1-score across all 60 signs with test videos: 26 signs achieve F1 ≥ 0.90 (mean 0.800), while confusions concentrate among minimal pairs sharing facial location (e.g., EAT vs. NOSE, 4 videos; HELLO vs. GOOD, 3 videos). When a 144-class candidate (3,182 training videos) was trained, it achieved 85.2% on 210 INCLUDE test videos of the new classes but regressed on the original 64 signs from 81.5% to 76.1% (and from 76.5% to 70.9% on unseen Amrita signers); the promotion gate automatically rejected it from replacing the active recognizer whilepreserving all 144 clips for the reverse avatar direction

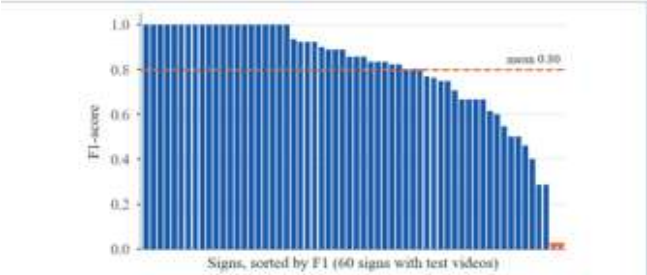


Fig. 5. Sorted per-sign F1-scores of the active model across the 60 signs with held-out test videos (mean F1 = 0.800).

Table ViiiLive-Pipeline Back-To-Back Stream Replay (Sign Error Rate)

Model Pipeline	Signs	Sub (S)	Del (D)	Ins (I)	SER
LogReg Baseline	268	55	74	1	0.485
LogReg+BiGRU (*T=16*)	268	18	88	0	0.396
LogReg+BiGRU (*T=24*)	268	50	27	2	0.295
BiGRU + INCLUDE + Trim	**314**	**54**	**8**	**24**	**0.274**

© 2026 IJSRET

4. Continuous Stream Segmentation & Perturbation Robustness

Replaying all 314 held-out test videos back-to-back through the live segmenter (Algorithm 2) reduced the stream Sign Error Rate (SER) from 0.485 to 0.274 (Table VIII), with still-tail trimming cutting deletions from 88 to 8. Fig. 6 evaluates five segmentation strategies across 50 synthetic five-sign streams (245 signs) under three boundary regimes: short pause, back-to-back, and hands never lowered. The default stillness/absence gate (800/200 ms) achieved the lowest test-signer SER (0.474) when signers pause briefly between words.

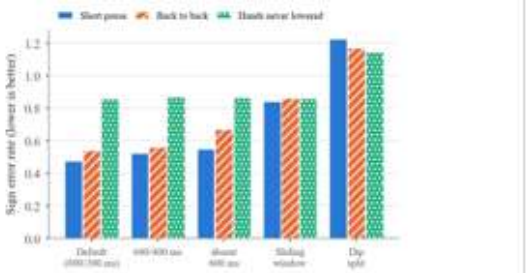


Fig. 6. Sign Error Rate (SER) of five segmentation variants on 50 test streams of five signs (245 total signs) with known ground-truth boundaries.

Fig. 7 evaluates synthetic landmark perturbations across 268 held-out videos: left-right mirroring, 20–40% random frame drops, 0.67×–1.5× speed scaling, ± 15% scale variation, and Gaussian coordinate noise degrade top-1 accuracy by at most 1.5 percentage points. Removing face landmarks drops accuracy by 19.0 points (66.8% → 47.8%), proving the importance of the face-relative vector rface in Eq. (2).

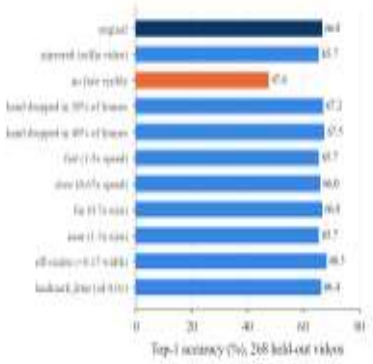


Fig. 7. Top-1 accuracy under synthetic landmark perturbations across 268 held-out test videos; removing face landmarks causes the largest drop.

Table Ixbrowser Mediapipe Vs. Offline Python Feature Extraction Parity E.

Continuous Sentence Decoding & Meaning Layer Impact

Table X compares four continuous sentence decoders on the 154 held-out test videos of unseen recording rounds (rounds 4 and 5) across 101 ISL-CSLTR sentences (~1.0% random chance). Log-likelihood fusion of the BiGRU-CTC decoder and the 32-step sentence classifier (Eq. 8) achieves 29.2% exact-sentence accuracy and reduces Gloss Error Rate (GER) from 0.883 to 0.685. On 124 held-out multi-sign phrase streams, the Kannada phrase meaning layer achieved 60.5% exact Kannada sentence match (chrF 0.669) compared to 39.5% (chrF 0.605) for word-by-word gloss translation.

Table X Continuous Sentence Decoders On 154 Unseen-Round Test Videos

Sentence Decoding Strategy (101 Classes)	Exact Acc (%)	GER ↓
32-Step Sentence Classifier Only	20.1	0.883
BiGRU-CTC Greedy Path → Nearest Sentence	18.2	0.742
BiGRU-CTC Prefix Likelihood Scoring Only	22.1	0.724
Fused BiGRU-CTC + Classifier (λ = *1.0*)	29.2	0.685

Cross-Runtime Parity Metric (N = 8 Reference Videos)	Measured Agreement
Hand presence detection agreement per video	91.8% – 100.0%
Handedness (Left/Right) label agreement	81.5% – 100.0%
Median 3D landmark coordinate discrepancy	0.002 – 0.007 units
Identical top-1 predicted sign label	7 / 8 (87.5%)

6. 3D Avatar Motion Quality & Real-Time CPU Latency

Table XI quantifies the impact of Algorithm 3 across all 144 dataset-derived 3D avatar clips. Gap interpolation (≤ 0.40 s) and Gaussian trajectory smoothing (σ = 1.5) eliminated 100% of mid-sign arm drops (from 367 drops across 114 clips down to 0) and reduced second-derivative wrist acceleration jitter by 36.0% (4.42

→ 2.83 cm/frame2). Table XII confirms that isolated sign classification requires only 3.8 ms on a commodity CPU, well within real-time interactive budgets.

Table Xi 3d Avatar Motion Quality Before & After Cleaning (144 Clips)

Objective Kinematic Metric	Raw Landmarks	Cleaned (Alg. 3)
Clips exhibiting mid-sign arm drops	114 / 144 (79.2%)	0 / 144 (0.0%)
Total mid-sign arm drop artefacts	367	0 (-100%)
Mean wrist jitter (cm/frame^2)	4.42	2.83 (-36.0%)
Mean active clip duration (s)	2.36 s	2.23 s

© 2026 IJSRET

TABLE XII
Measured Processing Latency On Commodity Laptop Cpu

Pipeline Stage / Operation	Measured Latency
Per-frame feature & segmenter processing	1.4 ms / frame
Isolated-sign BiGRU inference (with 3-view TTA)	5.8 ms / sign
Continuous greedy CTC decode per sentence	12.3 ms
Fused CTC + classifier decode (mean / p95)	37.0 / 62.0 ms
3D Avatar IK + Quintic co-articulation render	60 fps (16.6 ms)

VIII. PUBLIC SERVICE DEPLOYMENT, LIMITATIONS AND SDG IMPACT

1. Edge Economics & Sustainable Development Goals: By executing MediaPipe landmarking and ONNX WebAssembly inference directly on existing client CPUs, Silent Bridge eliminates recurring cloud GPU inference costs (~₹1.80 lakh/year per kiosk) and operates offline as a Progressive Web App (PWA). A cross-subsidized social enterprise model provides the B2C community app free forever (₹0) while licensing B2G/B2B service-counter kiosks at ₹12,000/terminal/year (92% lower than on-site human interpretation costs of ₹3.60 lakh/year). The platform directly advances UN SDG 3 (Target 3.8: accessible healthcare triage), SDG 4 (Target 4.5 & 4.a: bilingual ISL–Kannada deaf education), SDG 8 (Target 8.5: inclusive employment), SDG 9 (Target 9.c: frugal edge AI), and SDG 10 (Target 10.2: social and civic inclusion).

2. **Limitations & Threats to Validity: First, all reported recognition accuracies are measured on held-out recorded videos of unseen signers**; live webcam accuracy depends on lighting, camera placement, and regional ISL dialect variation (Kerala in Amrita SLR, Chennai in INCLUDE vs. Karnataka). Second, continuous sentence decoding is currently closed-set over 101 corpus sentences, and single training runs exhibit seed variance (± 4.5%). Third, avatar clips retargeted from 2D/3D hand/face landmarks estimate depth heuristically from palm scale and require formal linguistic validation by native deaf signers.

IX. CONCLUSION AND FUTURE WORK

This paper presented Silent Bridge, a real-time, privacy-preserving bidirectional Indian Sign Language and Kannada translation platform with a procedural 3D signing avatar running on commodity CPU hardware. Through our 147-D face-relative geometric schema, motion segmentation with still-tail trimming, 3-view Test-Time Augmentation, and selective HTTP-range INCLUDE dataset expansion, unseen-signer isolated recognition rose from 54.5% to 76.5% (p = 0.003, overall 81.5%, macro-F1 0.800), while fused BiGRU-CTC decoding achieved 29.2% exact-sentence accuracy on unseen rounds and trajectory cleaning eliminated all 367 mid-sign avatar arm drops. Future work includes

(i) multi-signer live webcam clinical trials in Karnataka hospitals and civic centers, (ii) open-vocabulary Transformer-CTC translation trained on iSign [9], and (iii) full upper-body pose integration with native deaf community co-design.

ACKNOWLEDGMENT

The authors express their sincere gratitude to Dr. Beemsha Arya (Principal, B.M.S. College of Engineering), Dr. R Ashok Kumar (Professor & Head, Dept. of CSBS, BMSCE), and Project Guide Tejaswini K (Assistant Professor, Dept. of CSBS, BMSCE) for their guidance and institutional support. We also thank the creators and deaf signers of the open-access Amrita SLR, INCLUDE, and ISL-CSLTR corpora.

REFERENCES

1. D. Li, C. Rodriguez, X. Yu, and H. Li, “Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison,” in Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV), Mar. 2020, pp. 1459–1469.
2. H. R. Vaezi Joze and O. Koller, “MS-ASL: A large-scale data set and benchmark for understanding American Sign Language,” in Proc. Brit. Mach. Vis. Conf. (BMVC), Sep. 2019, pp. 1–13.
3. O. Koller, J. Forster, and H. Ney, “Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers,” Comput. Vis. Image Underst., vol. 141, pp. 108–125, Dec. 2015.
4. N. C. Camgoz, S. Hadfield, O. Koller, H. Ney, and R. Bowden, “Neural sign language translation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2018, pp. 7784–7793.
5. A. Sridhar, R. G. Ganesan, P. Kumar, and M. Khapra, “INCLUDE: A large scale dataset for Indian Sign Language recognition,” in Proc. 28th ACM Int. Conf. Multimedia (ACM MM), Oct. 2020, pp. 1366–1375.
6. R. Elakkiya and B. Natarajan, “ISL-CSLTR: Indian Sign Language dataset for continuous sign language translation and recognition,” Mendeley Data, vol. 1, 2021, doi: 10.17632/kcmpdxky7p.1.
7. R. Sasidharan et al., “Amrita SLR: Real-time Indian Sign Language dataset,” Mendeley Data, vol. 1, 2025, doi: 10.17632/dtg9s785v8.1.

© 2026 IJSRET

8. A. Joshi, S. Agrawal, and A. Modi, "ISLTranslate: Dataset for translating Indian Sign Language," in Findings Assoc. Comput. Linguist. (ACL), Jul. 2023, pp. 10412–10428.
9. A. Joshi et al., "iSign: A benchmark for Indian Sign Language processing," in Findings Assoc. Comput. Linguist. (ACL), Aug. 2024, pp. 8920–8938.
10. J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the Kinetics dataset," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 6299–6308.
11. [11] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Realtime multi-person 2D pose estimation using Part Affinity Fields," IEEE Trans. Pattern Anal. Mach. Intell., vol. 43, no. 1, pp. 172–186, Jan. 2021.
12. C. Lugaresi et al., "MediaPipe: A framework for building perception pipelines," in Proc. IEEE/CVF CVPR Workshop Real-Time Comput. Vis., Jun. 2019, pp. 1–9.
13. F. Zhang et al., "MediaPipe Hands: On-device real-time hand tracking," in Proc. IEEE/CVF CVPR Workshop Comput. Vis. AR/VR, Jun. 2020, pp. 1–7.
14. S. Jiang, B. Sun, L. Wang, Y. Bai, K. Li, and Y. Fu, "Skeleton aware multi-modal sign language recognition," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Jun. 2021, pp. 3413–3423.
15. H. Hu, W. Zhao, W. Zhou, Y. Wang, and H. Li, "SignBERT: Pre-training of hand-model-aware representation for sign language recognition," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2021, pp. 11087–11096.
16. K. Cho et al., "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP), Oct. 2014, pp. 1724–1734.
17. M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," IEEE Trans. Signal Process., vol. 45, no. 11, pp. 2673–2681, Nov. 1997.
18. A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in Proc. 23rd Int. Conf. Mach. Learn. (ICML), Jun. 2006, pp. 369–376.
19. T. Hanke, "HamNoSys—Representing sign language data in language resources and language processing contexts," in Proc. LREC Workshop Represent. Process. Sign Lang., May 2004, pp. 1–6.
20. R. Wolfe et al., "Sign language avatars: A question of representation," Information, vol. 13, no. 4, p. 206, Apr. 2022.
21. Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," Psychometrika, vol. 12, no. 2, pp. 153–157, Jun. 1947.
22. I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in Proc. Int. Conf. Learn. Represent. (ICLR), May 2019, pp. 1–18.
23. M. Popović, "chrF: Character n-gram F-score for automatic MT evaluation," in Proc. 10th Workshop Stat. Mach. Transl. (WMT), Sep. 2015, pp. 392–395.
24. U. Zeshan, Sign Languages in Indo-Pakistan: A Description of a Signed Language. Amsterdam, Netherlands: John Benjamins Publishing, 2000.
25. S. N. Sridhar, Kannada: Descriptive Grammars. London, U.K.: Routledge, 1990.

© 2026 IJSRET