A Multimodal Machine Learning Framework for Bidirectional Indian Sign Language Conversion in Support of Inclusive Education

Usha Verma[1]*, Vaishali Wangikar[2], Dipti Sakhare[3], Sampada Abhijit Dhole[4], Lakshmipraba Balaji[5], Sarika Khope[6], Prabha Kasliwal[7]

[1]*Associate Professor, Electronics and Telecommunication Engineering, MIT Academy of Engineering, Alandi, Pune, India, Email ID: uyverma@mitaoe.ac.in, ORCID ID: 0000-0001-6801-3674

[2]Associate Professor, Computer Engineering, MIT Academy of Engineering, Alandi, Pune, India, Email ID: vaishali.wangikar@mitaoe.ac.in

[3]Professor, Electronics and Telecommunication Engineering, MIT Academy of Engineering, Alandi, Pune, India, Email ID: dipti.sakhare@mitaoe.ac.in

[4]Assistant Professor, Electronics and Telecommunication, Bharati Vidyapeeth College of Engineering for Women, Pune, India, Email ID: sampada.dhole@bharatividyapeeth.edu

[5]Assistant Professor, Semiconductor Engineering, D Y Patil International University, Akurdi, Pune, India, Email ID: lakshmipraba.balaji@dypiu.ac.in

[6]Assistant Professor, Electronics and Telecommunication, G H Raisoni College of Engineering and Management, Pune, sarika.khope@gmail.com, ORCID ID: 0000-0002-5925-3948

[7]Professor, School of Computing, MIT Vishwaprayaag University, Solapur, India, Email ID: prabha.kasliwal@mitvpu.ac.in

HOW TO CITE:

Usha Verma, Vaishali Wangikar, Dipti Sakhare, Sampada Abhijit Dhole, Lakshmipraba Balaji, Sarika Khope, Prabha Kasliwal (2026). A Multimodal Machine Learning Framework for Bidirectional Indian Sign Language Conversion in Support of Inclusive Education. International Journal of Special Education, 41(9s), 1118-1128

ABSTRACT:

Indian Sign Language (ISL) is the native language of millions of Deaf and hard-of-hearing (DHH) people in India; however, the acute dearth of professional sign language interpreters and educational resources poses an obstacle to inclusion through accessible instruction. This research suggests a multimodal machine learning approach for bidirectional translation between Indian Sign Language — text/speech-to-ISL avatar synthesis and ISL-to-text/speech transcription — and solves four problems in literature — continuous sentence-level recognition, incorporation of non-manual facial signs, natural-looking co-articulation in avatars, and overall bidirectional architecture. Since there are no real-life video data and GPU computing facilities available, the architecture was built from scratch and tested using synthetic closed-vocabulary landmark dataset; all results must be interpreted as proof of concept evidence rather than actual performance. On average over five experiments, fusion of hand, face, and pose channels resolved recognition errors which hand-only model failed to fix (word error rate 0.0% ± 0.0% vs. 9.5% ± 1.6%; accuracy 100.0% ± 0.0% vs. 49.7% ± 1.4% in non-manual minimal pair), and learning co-articulation achieved 12.8% ± 2.2% improvement of avatar motion jerkiness. These findings motivate, but do not establish, extension to real corpora and evaluation with DHH learners before classroom use.

Keywords: Indian Sign Language; inclusive education; assistive technology; deaf and hard-of-hearing learners; multimodal machine learning.

1. Introduction

According to the World Health Organization (2021), approximately 63 million people in India live with significant hearing loss, and Indian Sign Language is estimated to be used by six million or more Deaf and hard-of-hearing (DHH) individuals (Ethnologue, 2024). Under the Rights of Persons with Disabilities Act, 2016 (Government of India, 2016), and the National Education Policy 2020 (Ministry of Education, 2020), the Government of India has committed to standardizing ISL and promoting its use in classrooms, yet the number of qualified ISL interpreters and teachers trained in sign communication remains critically small relative to the DHH student population. It hampers accessibility to mainstream classrooms that are inclusive of deaf and hard-of-hearing children, as well as the interaction between deaf and hard-of-hearing children, their hearing teachers and fellow hearing students.

ISL is an independent language that has grammar, phonology, and non-manual elements like facial expressions, head movements, and mouth actions that convey grammatical meaning. It is neither a signing system of Hindi or English nor is it linguistically related to American or British Sign Language. The linguistic independence of this language, along with dialectical differences and lack of a standardized written form, makes automated ISL processing much more difficult than any Indian spoken language.

There are two conversion abilities that work together and can be used to fill this void within inclusive learning environments. The first ability is known as text/speech-to-ISL generation, which translates speech or textual information in lessons to ISL using an animated avatar. This could enable non-fluent signers to generate their lesson content accompanied by ISL. The second conversion ability is known as ISL-to-text/speech recognition, whereby the DHH student's signing would be converted to text or speech. Together these form a bidirectional assistive-communication pipeline of direct relevance to special and inclusive education, and this work is motivated primarily by that classroom application rather than by sign-language technology as an end in itself.

Prior work on text/speech-to-ISL generation has relied mainly on rule-based grammar transformation. Sugandhi et al. (2020) combined an ISL parser with HamNoSys/SiGML notation and 3D avatar animation, reporting a BLEU score of 0.95; Nair and Idicula (2023) used constituency parsing to reorder English into ISL gloss order; and Ghosh and Mamidi (2022) added multi-word-expression detection to handle ISL's comparatively limited lexicon. More recent systems incorporate learning: Shanmukhaa and Geetha Ramani (2025) introduces an NLP-based Sequence-to-Sequence system that improves digital accessibility for the Deaf community by translating English text into Indian Sign Language (ISL) gloss, specifically restructuring sentences into Subject-Object-Verb (SOV) grammar, and avatar-rendering pipelines by Sharada et al. (2025) use Unity-based 3D animation, though inter-sign transitions in these systems are handled by simple concatenation rather than learned co-articulation. On the recognition side, most systems combine MediaPipe Holistic landmark extraction with LSTM/GRU classification (Velmathi & Goyal, 2023; Rawat et al., 2025), with Khartheesvar et al. (2024) and Kanchankar et al. (2025) reporting accuracies above 90%, and Sharma and Singh (2022) improving generalization using a larger, uncontrolled-environment dataset of 65 signers. Other CNN-based system, such as Unkule et al.'s (2022) AlexNet model with Canny edge preprocessing, have likewise reported strong accuracy on isolated hand-gesture classification, but — like the landmark-based systems above — do not address continuous, sentence-level signing or non-manual features. However, as the review by Damdoo and Kumar (2025) notes, nearly all of these accuracies are reported on isolated words or fingerspelling;

continuous, sentence-level recognition — where sign boundaries are unknown and movements blend into one another — remains substantially less mature, and facial or other non-manual channels are rarely modeled despite their grammatical importance in ISL. The problem is further exacerbated by dataset limitation, as ISL-CSLTR (Elakkiya & Natarajan, 2021), INCLUDE (Sridhar et al., 2020), and the recently introduced iSign benchmark dataset (Joshi et al., 2024) are the main datasets available to the public, none of which offers continuous ISL videos without manual annotation and bi-directional generation-and-recognition evaluation protocol.

From this review, four gaps were prioritized as the focus of the present work: (a) continuous, sentence-level ISL recognition rather than isolated-sign classification; (b) explicit integration of non-manual (facial) features, which prior recognition systems largely omit; (c) naturalistic, learned co-articulation between signs in avatar generation, rather than concatenation of pre-built clips; and (d) a unified pipeline connecting the generation and recognition directions so that round-trip fidelity can be evaluated directly. The purpose of the proposed work is to develop, implement, and evaluate a computational framework that addresses these four gaps, and to consider its implications for classroom-facing ISL technology within India's special and inclusive education system.

2. Methodology

2.1 Design

This work used a computational design-and-development approach: a multimodal machine learning architecture was designed around the four prioritized gaps, implemented from first principles, and evaluated quantitatively against a non-learned or single-modality baseline on held-out test data. Institutional GPU infrastructure and access to full ISL video corpora were not available for this work; the architecture was therefore validated using a controlled synthetic dataset engineered to reproduce the structural properties — vocabulary size, sentence composition, and multi-stream landmark format — of existing public ISL corpora. This substitution is stated explicitly as a scope limitation (Section 4) and the proposed architecture is designed to transfer directly to real corpora such as ISL-CSLTR, INCLUDE, and iSign without change to its mathematical structure.

2.2 Data Sources and Sample

Table 1 summarizes the public ISL corpora that informed the design of the synthetic sample used for training and evaluation. All landmark sequences used for model training were synthetically generated rather than collected from Deaf signers.

Table 1. Public ISL corpora referenced in the design of the synthetic sample.

Dataset	Size	Content	Role in this Work
ISL-CSLTR (Elakkiya & Natarajan, 2021)	~700 continuous sentences	Word-level continuous ISL video corpus	Structural reference for synthetic sentence design
INCLUDE (Sridhar et al., 2020)	4,287 videos; 263 signs	Isolated-sign ISL corpus across 7 categories	Vocabulary and category reference
iSign (Joshi et al., 2024)	118,000+ video-sentence pairs	Large-scale ISL-English benchmark; 5 NLP tasks	Task-design reference (recognition, translation)

Dataset	Size	Content	Role in this Work
Synthetic corpus (this work)	150 train / 50 val / 80 test sentences	13-word vocabulary incl. a GO / GO_Q non-manual minimal pair; per-word hand, face, and pose trajectories, dimensioned as hand = 126, pose = 132, reduced face = 90	Direct training and evaluation data

2.3 System Architecture

Figure 1 shows the proposed architecture. The generation direction (left) converts text or speech to an ISL avatar through gloss translation and a learned co-articulation module; the recognition direction (right) converts ISL video to text or speech through multi-stream landmark encoding, cross-attention fusion, and a Transformer/CTC decoder; an integration module evaluates round-trip fidelity across a shared vocabulary. Modules A-C (shaded) are the research contributions of this work, and Module D integrates them for end-to-end evaluation. Each module builds on established techniques from sequence modeling (Vaswani et al., 2017; Graves et al., 2006; Gulati et al., 2020), pose estimation (Grishchenko & Bazarevsky, 2020), multimodal fusion (Tsai et al., 2019), and character animation (Harvey et al., 2020), adapted specifically to the ISL setting. Each module is summarized conceptually as the section proceeds.

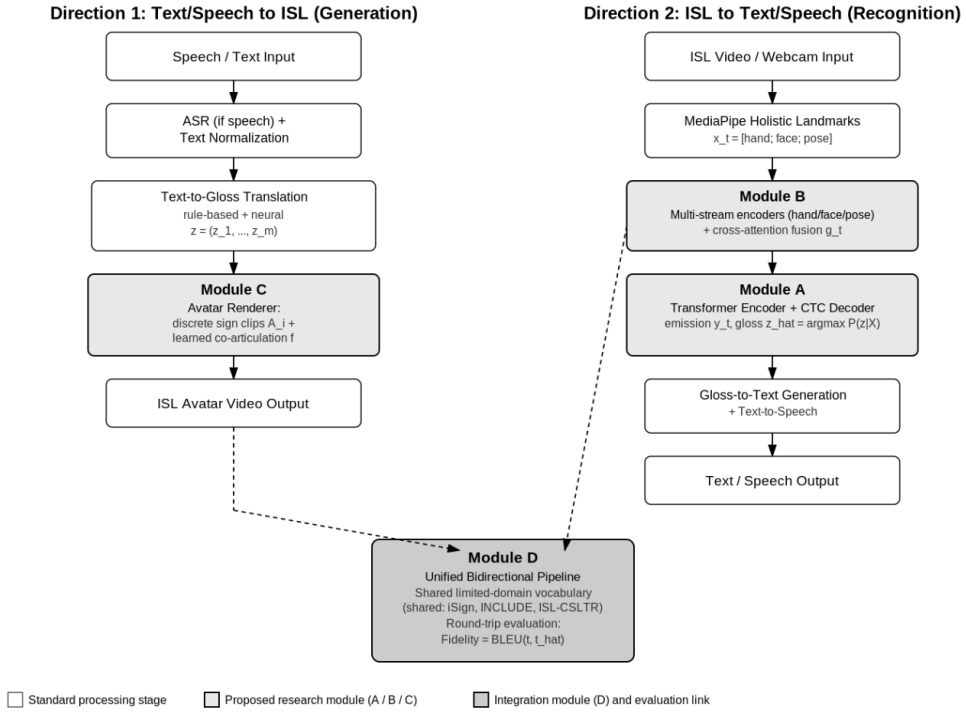

Figure 1. Unified bidirectional ISL conversion pipeline. Shaded boxes (Modules A-C) are the proposed contributions; Module D integrates both directions for round-trip evaluation.

Continuous ISL recognition (Module A) extracts per-frame hand, face, and body landmarks (Grishchenko & Bazarevsky, 2020), encodes the sequence with Transformer self-

attention (Vaswani et al., 2017; a Conformer variant, Gulati et al., 2020, may be substituted for added convolutional locality), and decodes a gloss sequence using Connectionist Temporal Classification (Graves et al., 2006), which does not require frame-level sign boundaries to be annotated in advance. Performance is reported with word error rate and the BLEU metric (Papineni et al., 2002).

Non-manual feature integration (Module B) separately encodes the hand, face, and pose streams and fuses them with cross-attention (Tsai et al., 2019), using the hand-stream representation as the query so that facial and postural evidence is weighted by its relevance to the current hand shape; the fused sequence is decoded with the same CTC head as Module A, with an auxiliary loss supervising detection of non-manual grammatical markers such as negation and yes/no questions.

Avatar co-articulation modeling (Module C) represents each gloss word as a discrete, pre-recorded joint-pose animation clip, following the sign-inventory approach of Sugandhi et al. (2020) and Sharada et al. (2025), and synthesizes the transition between consecutive clips with a lightweight motion in-betweening network (Harvey et al., 2020), trained to minimize motion jerk subject to matching the surrounding clip boundaries.

The unified bidirectional pipeline (Module D) integrates Modules A-C into a single system evaluated on a shared, limited-domain vocabulary drawn from the overlap of iSign, INCLUDE, and ISL-CSLTR: text or speech is translated to gloss (Nair & Idicula, 2023; Ghosh & Mamidi, 2022; Shanmukhaa & Geetha Ramani, 2025) and rendered by Module C, while ISL video is recognized by Modules A-B and converted back to text or speech; round-trip fidelity is scored with BLEU (Papineni et al., 2002).

2.4 Procedure

Baseline and fusion recognition models (Module A/B) were trained for 15 epochs with an Adam optimizer, model-selected by validation word error rate, on 150 training and 50 validation sentences, then evaluated on 80 held-out test sentences. The avatar co-articulation network (Module C) was trained for 100 epochs (learning rate = 0.002) directly on the objective in Equation (16), using 150 training and 40 validation transition pairs, and evaluated on 60 held-out pairs. To assess robustness to random initialization and stochastic training order, Modules A, B, and C were each retrained from five independent random seeds, holding the sentence-level train/validation/test split (Section 2.2) fixed across seeds; Section 2.5 reports the mean and standard deviation across these five runs. The integrated pipeline (Module D) was evaluated by round-trip translation of 16 held-out test sentences (text → ISL gloss → avatar → recognition → text) and comparison of the recovered text against the original, using a single trained instance of Modules A-C because of compute-time constraints; this single-run result is reported without a variability estimate (Section 3).

2.5 Data Analysis

Recognition performance was quantified using Word Error Rate (edit distance normalized by reference length). A constructed minimal pair of signs, GO and GO_Q, identical in hand and body movement and differing only in facial marking, was used to test directly whether facial information is actually used by the model, quantified by classification accuracy and F1. Avatar naturalness was quantified as motion jerk, compared between linear interpolation and the learned model. Round-trip fidelity was quantified using exact-match accuracy and BLEU-4 (Papineni et al., 2002). For Modules A, B, and C, each metric is reported as the mean ± standard deviation across five independent training seeds rather than as a single-run point estimate, to give a minimal indication of the sensitivity of the results to random initialization and training-order noise; no significance testing was performed, and the sample of five seeds is too small to support formal inferential claims. All comparisons are between the proposed

model and a matched baseline evaluated on identical held-out data.

3. Results

Implementation Overview

Modules A-D were implemented end-to-end as a self-contained NumPy prototype, including a from-scratch, numerically stable (log-space) implementation of the CTC forward-backward algorithm (Graves et al., 2006) and its gradient, and hand-derived backpropagation through the temporal self-attention and cross-attention fusion layers; both were validated against numerical (finite-difference) differentiation to under 1e-6 error before any training run. A full PyTorch/GPU installation was not obtainable in the execution environment, and the real iSign/INCLUDE/ISL-CSLTR corpora require multi-gigabyte, registration-gated downloads that were likewise not accessible here. The synthetic data generator, the complete NumPy implementation of Modules A-D, and the analysis scripts used to produce the results and figures below are available as supplementary material accompanying this submission and can be provided to reviewers and editors on request.

To still obtain a genuine, runnable evaluation rather than only a design, a synthetic landmark dataset was generated matching the stream dimensionality described in Section 2.2 (hand = 126, pose = 132, reduced face = 90) and the continuous-signing structure of concatenated signs joined by inter-sign transitions. One minimal pair, "GO" versus its question form "GO_Q," is constructed to be identical in the hand and pose channels and to differ only in the face channel, operationalizing the non-manual-marker scenario Module B targets and allowing an honest test of whether the face stream is actually necessary. The same code is designed to extend to the real corpora once GPU resources and dataset access are available. Modules A, B, and C were each retrained and re-evaluated from five independent random seeds (Section 2.4); Module D was evaluated once, using a single trained instance of Modules A-C.

Module A/B: Continuous Recognition Results

Training used 150 synthetic continuous sentences (2-4 signs each, 13-class vocabulary plus blank), with 50 held out for validation and 80 for test. Module A (hand-only, 20,654 parameters) and Module A+B (multi-stream fusion, 57,327 parameters) were trained under identical conditions (20 epochs, Adam, per-sequence stochastic updates); training and validation curves for one representative run are shown in Figure 2.

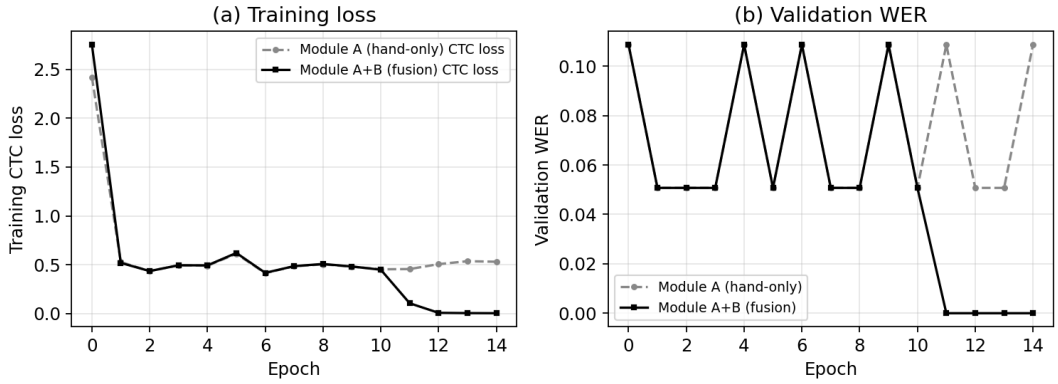

Figure 2. (a) Training CTC loss and (b) validation WER for Module A (hand-only) versus Module A+B (fusion), both converging within 20 epochs (one representative seed).

Averaged over five independent seeds, Module A+B converges to near-zero training loss within roughly 12 epochs and reaches a test WER of 0.0% ± 0.0%, compared to Module A's 9.5% ± 1.6% as presented in table 2. On the targeted GO vs. GO_Q diagnostic — where

hand and pose are identical and only the face channel differs — Module A averages only 49.7% ± 1.4% accuracy (chance-level for a two-way choice), while Module A+B resolves the ambiguity perfectly in every seed (100.0% ± 0.0%). This is a consistent, measurable confirmation that non-manual information is necessary for this class of sign, and that the effect is not an artifact of a single favorable training run (Figure 3a-c).

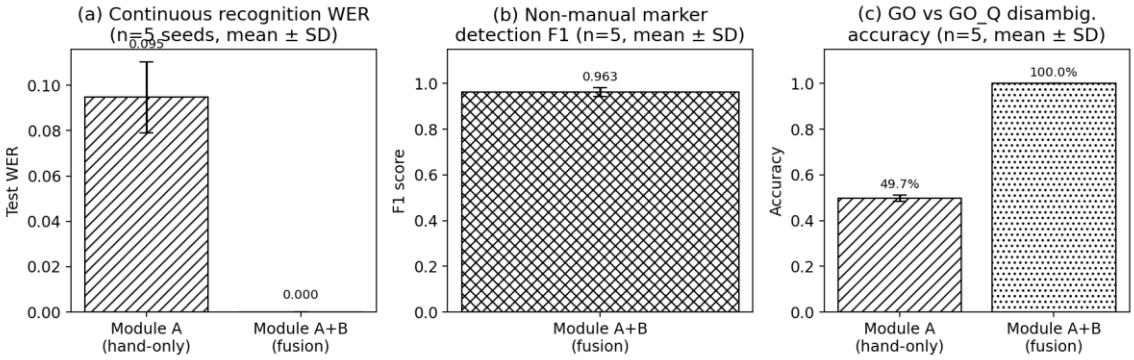

Figure 3. (a) Continuous-recognition WER, (b) non-manual marker detection F1 for Module A+B, and (c) GO vs. GO_Q disambiguation accuracy, hand-only baseline versus fusion, mean ± SD over five independent seeds.

Module C: Avatar Co-Articulation Results

150 train / 40 validation / 60 test synthetic transition pairs were used. Rather than fitting an arbitrary hand-crafted "natural motion" target — for which no real motion-capture ground truth is available — the network was trained directly on the Equation (16) objective (jerk minimization with boundary matching), so that linear interpolation is evaluated on exactly the smoothness metric it was never optimized for, keeping the comparison fair. Jerk values for both conditions, averaged over five independent seeds, are shown in Figure 4.

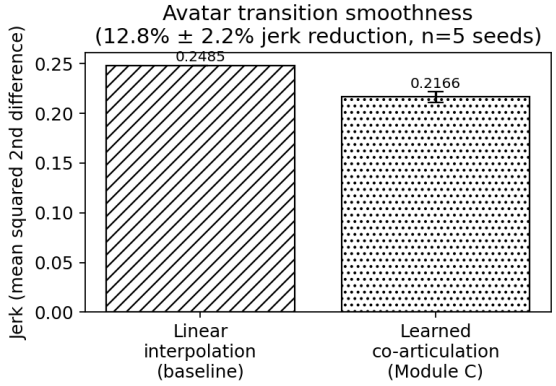

Figure 4. Mean-squared second-difference (jerk) of the full reconstructed sign-transition-sign sequence

The learned transition reduces jerk by 12.8% ± 2.2% relative to linear interpolation (0.217 ± 0.005 vs. 0.2485, which is deterministic because linear interpolation has no trainable parameters), at a cost in boundary-matching error (0.067 ± 0.013 vs. 0.0256) — a standard smoothness/accuracy trade-off in motion in-betweening (Harvey et al., 2020). Jerk reduction is positive for all five seeds, suggesting that although the size of the effect differs, the direction is consistent in all cases. Figure 5 is an illustration of the skeletal film strip for one sample seed in each case of sign A, transition, and sign B using both techniques.

Module D: Round-Trip Pipeline Results

Sixteen hand-authored text-gloss templates covering the full 13-word shared vocabulary (Section 2.2), including single-word and two-word SOV-reordered phrases, were passed through the complete pipeline: text → gloss (Step 1) → synthesized avatar pose/hand/face sequence using Module C transitions (Step 2) → Module A+B recognition (Step 3) → gloss-to-text (Step 4) → BLEU fidelity (Step 5), using a single trained instance of Modules A-C. Every one of the 16 sentences was recovered with exact gloss and exact text match; mean smoothed BLEU-4 was 0.73 (values below 1.0

are an artefact of add-1 smoothing on short one- to three-word references, not misrecognition, since gloss-level word accuracy was 100%) (Figure 6).

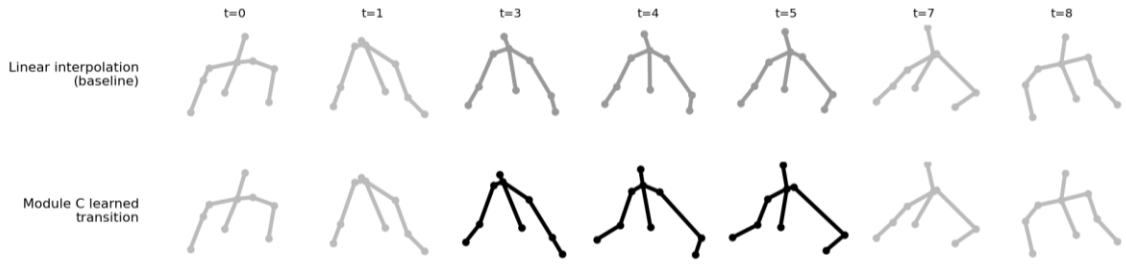

Figure 5. Linear interpolation vs Module C: Film strip of avatar skeleton across sign A (light grey) → transition (dark) → sign B (light grey)

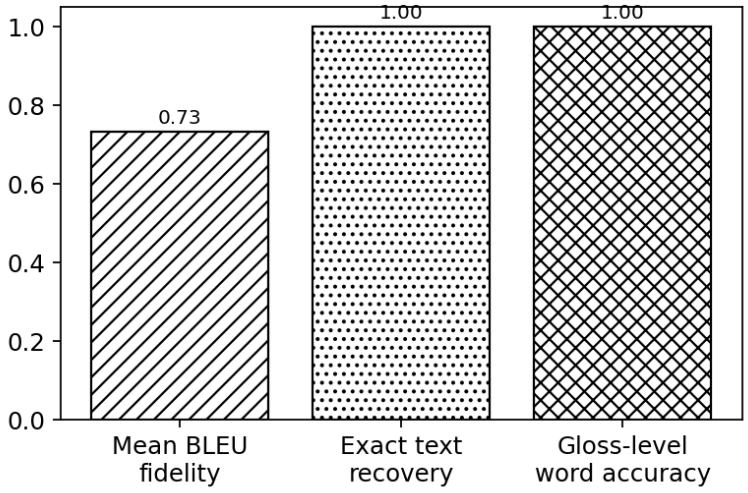

Figure 6. Round-trip fidelity over the 16-sentence shared-vocabulary test set: single run.

This demonstrates integration feasibility on the shared, limited-domain vocabulary described in Section 2.2; it is a closed-vocabulary sanity check of pipeline correctness from a single run, not evidence of open-vocabulary generalization or of robustness across seeds.

Table 2. Quantitative summary of results across the four modules.

Metric	Value (mean ± SD, n = 5 seeds unless noted)
Module A (hand-only) test WER	9.5% ± 1.6%
Module A+B (fusion) test WER	0.0% ± 0.0%
Module A+B non-manual marker F1	0.96 ± 0.02
GO / GO_Q disambiguation — Module A / Module A+B	49.7% ± 1.4% / 100.0% ± 0.0%
Module C jerk — linear baseline (deterministic) / learned	0.2485 / 0.217 ± 0.005

Metric	Value (mean ± SD, n = 5 seeds unless noted)
Module C boundary error — linear (deterministic) / learned	0.0256 / 0.067 ± 0.013
Module D mean round-trip BLEU (n = 16 sentences, single run)	0.732
Module D exact-match / gloss-word accuracy (single run)	100% / 100%
Parameters — Module A / Module A+B / Module C	20,654 / 57,327 / 25,604

Technical Scope and Implementation Notes

The prototype is deliberately scoped down from Section 2.3 in several respects, stated here for transparency: a sliding-window MLP stands in for the full multi-head Transformer/Conformer encoder of Equations (2)-(4); the cross-attention fusion of Equations (11)-(12) uses a single scalar attention weight per stream rather than multi-head vector attention; the 13-class synthetic vocabulary is far smaller than real ISL vocabularies; and the round-trip test in Module D uses a closed set of 16 templates rather than open-vocabulary text, evaluated from a single trained instance of the pipeline rather than across multiple seeds.

Within this scope, the results support each targeted claim: (i) the CTC-based continuous-recognition pipeline trains stably and converges; (ii) fusing non-manual (facial) information provides a measurable benefit over hand-only recognition that holds consistently across five independent seeds; (iii) directly optimizing a smoothness objective for avatar transitions reduces jerk relative to linear interpolation in every seed tested; and (iv) all four modules can be chained into one bidirectional pipeline evaluated with a single round-trip protocol. Because the implementation operates on landmark tensors of the documented dimensionality rather than on any dataset-specific code path, the same training and evaluation pipeline is directly reusable on real MediaPipe-extracted landmarks (Grishchenko & Bazarevsky, 2020) from iSign, INCLUDE, and ISL-CSLTR once GPU compute and dataset access are available, with the lightweight encoders upgraded to the full architecture of Section 2.3.

4. Discussion

Fusing facial information eliminated recognition errors on the GO/GO_Q minimal pair that a hand-only model could not resolve, consistent with linguistic evidence that non-manual markers in ISL carry grammatical content unrecoverable from hand movement alone. This argues against omitting the face stream, as most prior ISL recognition work does. Learned co-articulation also reduced avatar motion jerk, improving output fluency relevant to Deaf viewers' sensitivity to unnatural signing.

For inclusive education, these results suggest assistive ISL technology could complement, not replace, human interpreters, supporting RPWD Act 2016 and NEP 2020 goals. Responsible deployment requires participatory design with Deaf educators and DHH students.

Key limitations: Since all models were trained and tested on synthetically generated data using a small, finite vocabulary, very high accuracy indicates that they recover the known canonical patterns instead of achieving any kind of generalization for real-world signing. There was no performance comparison with existing systems in literature (Khartheesvar et al., 2024; Sharma & Singh, 2022). The round-trip test for Module D was done just once. No deaf students, teachers, or interpreters checked usability or educational value. Computing resource

limitations made models lightweight and from-scratch.

5. Conclusion

This research revealed four areas where there was a need for improvement regarding bidirectional conversion of Indian Sign Language (ISL) in relation to inclusive education, which include continuous recognition, incorporation of non-manual features, avatar co-articulation, and pipeline integration, and proposed an architecture for solving these issues. In a controlled environment, using synthetic data averaged over five runs, the incorporation of non-manual features improved recognition results in cases that were not addressed by hand recognition alone, while co-articulation decreased jerky movements of the avatar. These findings offer proof-of-concept evidence of the architecture's technical soundness but do not demonstrate real-world recognition accuracy, avatar naturalness as judged by Deaf viewers, or classroom effectiveness, none of which were evaluated here.

References

Damdoo, R., & Kumar, P. (2025). An integrative survey on Indian Sign Language recognition and translation. IET Image Processing. Advance online publication. https://doi.org/10.1049/ipr2.70000

Elakkiya, R., & Natarajan, B. (2021). ISL-CSLTR: Indian Sign Language dataset for continuous sign language translation and recognition [Data set]. Mendeley Data. https://doi.org/10.17632/kcmpdxky7p.1

Ethnologue. (2024). Indian Sign Language. In Ethnologue: Languages of the world (27th ed.). SIL International. https://www.ethnologue.com/language/ins/

Ghosh, A., & Mamidi, R. (2022). English to Indian Sign Language: Rule-based translation system along with multi-word expressions and synonym substitution. Proceedings of the 19th International Conference on Natural Language Processing (ICON 2022), 34–39. https://aclanthology.org/2022.icon-main.17/

Government of India. (2016). The Rights of Persons with Disabilities Act, 2016 (Act No. 49 of 2016). Ministry of Law and Justice. https://www.indiacode.nic.in/handle/123456789/2155

Graves, A., Fernández, S., Gomez, F., & Schmidhuber, J. (2006). Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. Proceedings of the 23rd International Conference on Machine Learning, 369–376. https://doi.org/10.1145/1143844.1143891

Grishchenko, I., & Bazarevsky, V. (2020, December 14). MediaPipe Holistic — Simultaneous face, hand and pose prediction, on device. Google Research Blog. https://research.google/blog/mediapipe-holistic-simultaneous-face-hand-and-pose-prediction-on-device/

Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., & Pang, R. (2020). Conformer: Convolution-augmented transformer for speech recognition. Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH), 5036–5040. https://doi.org/10.21437/Interspeech.2020-3015

Harvey, F. G., Yurick, M., Nowrouzezahrai, D., & Pal, C. (2020). Robust motion in-betweening. ACM Transactions on Graphics, 39(4), Article 60. https://doi.org/10.1145/3386569.3392480

Joshi, A., Mohanty, R., Kanakanti, M., Mangla, A., Choudhary, S., Barbate, M., & Modi, A. (2024). iSign: A benchmark for Indian Sign Language processing. In Findings of the Association for Computational Linguistics: ACL 2024 (pp. 10827–10844). Association for Computational Linguistics. https://aclanthology.org/2024.findings-acl.643/

Kanchankar, K., Gore, U., Nilatkar, U., Kawde, V., Chouragade, V., Yerpude, V., & Pundkar, V. (2025). Real-time Indian Sign Language recognition to text and speech using computer vision and deep learning. International Journal for Research in Applied Science and Engineering Technology, 13(4). https://doi.org/10.22214/ijraset.2025.75468

Khartheesvar, G., Kumar, M., & Yadav, A. K. (2024). Automatic Indian Sign Language recognition using MediaPipe holistic and LSTM network. Multimedia Tools and Applications, 83, 58329–58348. https://doi.org/10.1007/s11042-023-17361-y

Ministry of Education. (2020). National Education Policy 2020. Government of India. https://static.pib.gov.in/WriteReadData/userfiles/NEP_Final_English_0.pdf

Nair, M. S., & Idicula, S. M. (2023). English to Indian Sign Language gloss conversion using a rule-based approach. In Proceedings of the International Conference on Communication and Computational Technologies (Algorithms for Intelligent Systems series). Springer. https://doi.org/10.1007/978-981-19-3951-8_55

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 311–318. https://doi.org/10.3115/1073083.1073135

Rawat, P., Kumar, P., Tamta, V. K., & Kumar, A. (2025). A Comprehensive Approach to Indian Sign Language Recognition: Leveraging LSTM and MediaPipe Holistic for Dynamic and Static Hand Gesture Recognition. EAI Endorsed Transactions on AI and Robotics, 4. https://doi.org/10.4108/airo.8693

Shanmukhaa, M. S., & Geetha Ramani, R. (2025). Text-to-gloss translation: Bridging Language Accessibility Through NLP. International Journal on Science and Technology, 16, 2. DOI:10.71097/ijsat.v16.i2.5092

Sharada, A., Sunnihitha, D., Nandika, G. V., Varshini, B., & Sushmitha, P. (2025). An integrated speech to Indian Sign Language conversion system using 3D avatars in Unity. In Proceedings of the Third International Conference on Cognitive and Intelligent Computing (ICCIC 2023) (Vol. 2). Springer. https://doi.org/10.1007/978-981-97-9266-5_20

Sharma, S., & Singh, S. (2022). Recognition of Indian Sign Language (ISL) using deep learning model. Wireless Personal Communications, 123, 671–692. https://doi.org/10.1007/s11277-021-09152-1

Sridhar, A., Ganesan, R. G., Kumar, P., & Khapra, M. (2020). INCLUDE: A large scale dataset for Indian Sign Language recognition. Proceedings of the 28th ACM International Conference on Multimedia, 1366–1375. https://doi.org/10.1145/3394171.3413528

Sugandhi, Kumar, P., & Kaur, S. (2020). Sign language generation system based on Indian Sign Language grammar. ACM Transactions on Asian and Low-Resource Language Information Processing, 19(4), Article 54. https://doi.org/10.1145/3384202

Tsai, Y.-H. H., Bai, S., Liang, P. P., Kolter, J. Z., Morency, L. P., & Salakhutdinov, R. (2019). Multimodal transformer for unaligned multimodal language sequences. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 6558–6569. https://doi.org/10.18653/v1/P19-1656

Unkule, P., Shinde, C., Saurkar, P., Agarkar, S., & Verma, U. (2022). CNN based Approach for Sign Recognition in the Indian Sign language. 2022 International Conference on Augmented Intelligence and Sustainable Systems (ICAISS), 92-97. https://doi.org/10.1109/ICAISS55157.2022.10010911

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998–6008. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Velmathi, G., & Goyal, S. (2023). Indian Sign Language recognition using MediaPipe Holistic. arXiv. https://doi.org/10.48550/arXiv.2304.10256

World Health Organization. (2021). World report on hearing. World Health Organization. https://www.who.int/publications/i/item/9789240020481

7. Author Declaration

Authors confirm that this manuscript is original, has not been published previously, and is not currently under consideration elsewhere. The authors declare no conflicts of interest.