

SilentBridge: A Unified Emotion-Aware, Gloss-Free, Mobile-Deployable Indian Sign Language Translation System with Bidirectional Communication

Tejaswini K¹, Roopa T², Kiran M K³, Manoj G U⁴, and H S Thanush⁵

^{1,2,3,4,5} Department of Computer Science and Business Systems, B.M.S. College of Engineering, Bengaluru, Karnatak
¹tejaswinik.cbs@bmsce.ac.in, ²roopat.cbs@bmsce.ac.in, ³kiranmk.bs23@bmsce.ac.in, ⁴manojgu.bs24@bmsce.ac.in and
⁵hsthanush.bs23@bmsce.ac.in

Received: 14th June, 2026; Revised: 20th July 2026; Accepted: 20th August, 2026; Available Online: 05th September 2026

ABSTRACT

Silent Bridge is a mobile-oriented prototype intended to support a complete conversation between deaf and hearing users, with Indian Sign Language (ISL) as its main setting. The system combines capabilities that are often examined separately: continuous gloss-free recognition from Media Pipe keypoints, facial-affect analysis, emotion-aware speech synthesis, INT8 inference, lightweight ISL adaptation through LoRA, and a reverse speech-to-avatar channel. In the forward path, ST-GCN features are processed by a temporal Transformer and decoded with CTC. A parallel facial branch estimates valence and arousal, which are then used to shape VITS-style speech prosody. AI4Bharat resources provide speech output in Indian languages. In prototype evaluations on PHOENIX14T and INCLUDE-ISL, SilentBridge obtains 26.1 BLEU-4, 84.3% emotion-match accuracy in synthesized speech, and a mean on-device latency of 243ms on Snapdragon 660-class hard-ware. The contribution is therefore not a single isolated model, but an integrated design that connects recognition, affect, speech, regional-language support, and mobile deployment in one communication pipeline.

Keywords: Sign language translation · Emotion-aware TTS · Mobile inference · Indian Sign Language · Gloss-free recognition · Bidirectional communication · LoRA adaptation

How to cite this article: Tejaswini K, Roopa T, Kiran MK, Manoj GU, Thanush HS, Silent Bridge: A Unified Emotion-Aware, Gloss-Free, Mobile-Deployable Indian Sign Language Translation System with Bidirectional Communication. Int J Drug Deliv Technol. 2026; 16(7): 1689-1696-1679. DOI: 10.25258/ijddt.16.7.183

Source of support: Nil.

Conflict of interest: None

1. INTRODUCTION

For many deaf people in India, a routine conversation with a hearing person still de-pends on a family member, handwritten notes, or partial lip-reading. These workarounds may be manageable in informal settings, but they become difficult during brief and important exchanges at a clinic, railway enquiry counter, classroom, or government office. Sign-language recognition (SLR) has improved considerably on controlled benchmarks; however, relatively few systems are designed around the practical constraints of everyday use [1].

Four recurring gaps in the literature directly influenced the design of SilentBridge:

Gloss dependency. A common SLT pipeline first converts a signing video into glosses and then translates those glosses into spoken-language text. Glosses are valuable for annotation and linguistic analysis, but they can also discard details visible in the original performance, such as timing, intensity, facial expression, and conversational emphasis [2,3].

Emotion handling. Facial behaviour in sign language is not simply an emotional add-on. Eyebrow position, mouth shape, and head movement may carry grammatical meaning while also expressing the signer's affect. Recent emotion-aware systems model

these cues [4,5], but the estimated emotion is usually used only inside the translator and does not influence the voice heard by the non-signing participant [6].

Phone-level deployment. Most strong results are reported on workstation GPUs. That setting is appropriate for model development, but it says little about whether a system will remain responsive on a mid-range phone with limited memory, battery life, and thermal capacity [7,8].

Indian-language access. ISL research has progressed steadily, yet many prototypes end with an English label, English text, or a desktop-only demonstration. In practice, users may need the translated message to be spoken in a regional language, not merely displayed in English [9].

SilentBridge was built around these practical concerns rather than around a single benchmark score. Its main contributions are:

1. A dual-output facial module that separates grammatical facial markers from affective cues and passes the affective estimate to speech synthesis;
2. A gloss-free CTC decoder built on ST-GCN skeleton features, avoiding a gloss bottleneck in the main translation

*Author for Correspondence: tejaswinik.cbs@bmsce.ac.in

path;

3. Selective quantization-aware training that places graph-convolution and attention layers in INT8 while retaining float32 where numerical stability is important;
4. A compact LoRA adapter trained on INCLUDE-ISL for ISL localization and integration with Indic TTS resources;
5. A reverse mode that transcribes the hearing user's speech with Whisper ASR [10] and presents the response through a signing avatar; and
6. A privacy-aware personalization workflow in which signer-specific LoRA updates are learned from keypoints rather than shared raw video.

2. BACKGROUND AND RELATED WORK

2.1 Sign Language Recognition Architectures

Dual- and multi-stream models. Harrouch et al. [11] report gains from processing hand and head information in separate streams for continuous SLR. The separation is meaningful because manual gestures and facial movements do not encode the same type of information. T-MORT [5] develops this idea further through visual, gesture, and emotion streams combined by late fusion. SilentBridge follows a similar principle: each branch is allowed to learn a useful representation before the signals are brought together.

Transformer efficiency. Transformers are effective for SLT, but standard self-attention becomes expensive as a signing sequence grows. SignEdgeLVM [7] uses GRAM attention for edge settings, whereas GLoT [12] reduces the attention cost to $O(n \log n)$ with a gated-logarithmic formulation. These studies motivated the efficient temporal attention used in SilentBridge, where latency is treated as part of the model design rather than as a later optimization step.

2.2 Emotion in Sign Language Translation

EASLT [4] treats facial affect as part of the translation signal instead of reducing it to an auxiliary label. Its PHOENIX14T result therefore provides a useful point of comparison for gloss-free translation. SilentBridge extends this idea beyond text generation: the affective estimate also reaches the speech synthesizer. This

choice is consistent with recent TTS surveys, which identify controllable and emotion-aware speech as an important open research problem [6].

2.3 Edge and Mobile Deployment

S3D-Conv1D [8] examines quantization operator by operator rather than applying it as a uniform final compression step. The reported results support a mixed-precision design: convolutional and linear operations can use INT8, while softmax and gating operations remain in float32. SilentBridge adopts the same practical compromise. More broadly, surveys of SLR continue to note that deployment on mobile hardware lags behind progress in recognition accuracy [1].

2.4 ISL and Multilingual Output

Recent ISL systems use CNNs, Transformers, and mobile backbones with increasingly strong results [13,14,15]. Even so, many demonstrations stop at a recognized label or an English sentence. SilentBridge treats regional-language speech as part of the core pipeline. INCLUDE-ISL [16] supports adaptation on the sign side, while AI4Bharat resources provide the speech-generation layer [17,18].

2.5 Bidirectional Communication and Privacy

The reverse direction receives less attention in SLT research, although it is essential if both participants are to take part in the conversation. MS2SL [19] demonstrates how speech or text can be mapped into a sign-production space and serves as the basis for the speech-to-avatar path considered here. Privacy is equally important because signing video is biometric data. NP8 [20] explores edge, cloud, and blockchain mechanisms for privacy-aware SLT. SilentBridge does not depend on blockchain, but it follows the same principle of keeping personal signing data local wherever possible.

2.6 Gap Summary

Table 1 places the reviewed systems alongside the requirements identified for SilentBridge. The purpose is not to rank the cited studies, but to show which parts of the intended conversation pipeline each one addresses.

Table 1. Research gap matrix: reviewed papers vs. SilentBri

System	Cont. SLR	Emo. to Voice	Mobile	ISL	Multilingual	Bid
TS-CNN [11]	✓	×	×	×	×	×
ResNet50ViT [21]	✓	×	×	×	×	×
YOLOv5s+TTS [22]	~	×	×	*	~	×
ISL CNN [13]	×	×	×	✓	×	×
ISL Transformer [14]	✓	×	×	✓	×	×
MobileNetV2 [15]	×	×	×	✓	×	×
EASLT [4]	✓	×	*	×	×	×
T-MORT [5]	✓	×	*	×	×	×
SignEdgeLVM [7]	✓	×	~	×	×	×
GLoT [12]	✓	×	×	×	×	×
MS2SL [19]	✓	×	×	×	×	×
S3D-Conv1D [8]	×	×	✓	×	×	×
Edge-BC [20]	✓	×	×	×	×	×
SilentBridge	✓	✓	✓	✓	✓	✓

✓ Yes; × No; ~ Partial; * emotion in text, not routed to voice.

3. DATASETS AND PUBLIC RESOURCES

The prototype is built entirely from public datasets and pretrained resources rather than from a private signing corpus. Table 2 lists the resources used in the design and evaluation. Direct links are provided because locating the correct dataset versions and pretrained models is often a significant barrier to reproducing sign-language research.

Table 2. Datasets and public resources used in SilentBridge

Resource	Use in Project	Original Link
PHOENIX14T [23]	Main continuous SLT benchmark for BLEU evaluation	https://www-i6.informatik.rwth-aachen.de/~koller/RWTH-PHOENIX-2014-T/
CSL-Daily [24]	Secondary continuous SLT benchmark with gloss and text annotations	https://ustc-slr.github.io/datasets/2021_csl_daily/
INCLUDE-ISL [16]	Indian Sign Language LoRA adaptation and signer personalization	https://include-dataset.github.io/
FER2013 [25]	Facial-expression training for the emotion stream	https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge/data
WLASL	Word-level American Sign Language training data	https://dxli94.github.io/WLASL/
IndicVoices-R [17]	Multilingual Indian TTS data foundation	https://github.com/AI4Bharat/IndicVoices-R
Indic Parler-TTS [18]	Practical multilingual Indic speech synthesis backend	https://huggingface.co/ai4bharat/indic-parler-tts-pretrained
MediaPipe Holistic [26]	On-device body, hand, and face keypoint extraction	https://developers.google.com/mediapipe/solutions/vision/holistic_landmarker

4. SYSTEM ARCHITECTURE

4.1 Overview

Figure 1 presents the complete pipeline, from camera capture and landmark extraction to translation, speech generation, and reverse avatar rendering. The mobile implementation runs these components asynchronously. As a result, a slower stage—for example, speech synthesis—does not have to pause the rest of the interaction.

4.2 Input Layer: MediaPipe Keypoint Extraction

Every camera frame is first processed by MediaPipe Holistic [26] to obtain hand, pose, and facial landmarks. For the emotion branch, the full facial output is reduced to a 68-point representation compatible with FER2013-style expression training [25]. Once the landmarks have been extracted, the raw frame is discarded and the remaining modules operate only on coordinates.

4.3 Sign Stream: ST-GCN to Temporal Transformer to CTC

The sign stream represents 75 upper-body and hand landmarks as a skeleton graph $G = V, E$. The edges follow anatomical connections, so information can move between neighboring joints during graph convolution. The ST-GCN block [27] processes a 90-frame window at a 45-frame stride and produces a feature sequence $F \in \mathbb{R}^{T \times 256}$.

PHOENIX14T is used as the main translation benchmark, with CSL-Daily providing a cross-dataset check. INCLUDE-ISL supplies the data needed for ISL adaptation. FER2013 is not a signing dataset, so it is used only to initialize the facial-expression branch; signing data remains necessary for final tuning. AI4Bharat resources allow the output to be spoken in Indian languages. On the device, only derived keypoints are retained after processing, not the original video frames.

Four temporal Transformer layers model motion across the resulting sequence. Following GLoT [12], gated logarithmic attention is used instead of standard $O(n^2)$ self-attention. The final CTC head predicts vocabulary tokens directly. Consequently, the forward path does not require a manually prepared gloss sequence [2,9].

4.4 Emotion Stream: Dual-Output Facial CNN

The facial branch uses a lightweight ResNet-18 backbone. A facial movement may carry grammar, affect, or both; forcing these roles into one output would make the representation difficult to interpret. Inspired by EASLT [4], the network therefore separates its final representation into two heads:

$$h_{gram} = W_g z, \quad h_{aff} = W_a z \quad (1)$$

Here, z is the penultimate embedding. The first output, $h_{gram} \in \mathbb{R}^7$, represents grammatical markers such as eyebrow raises and mouth shapes. The second, $h_{aff} \in \mathbb{R}^2$, estimates valence and arousal for the speech module.

4.5 Cross-Attention Fusion

The sign and emotion branches are allowed to form their own representations before fusion. Immediately before translation, cross-attention lets the sign stream query the facial representation:

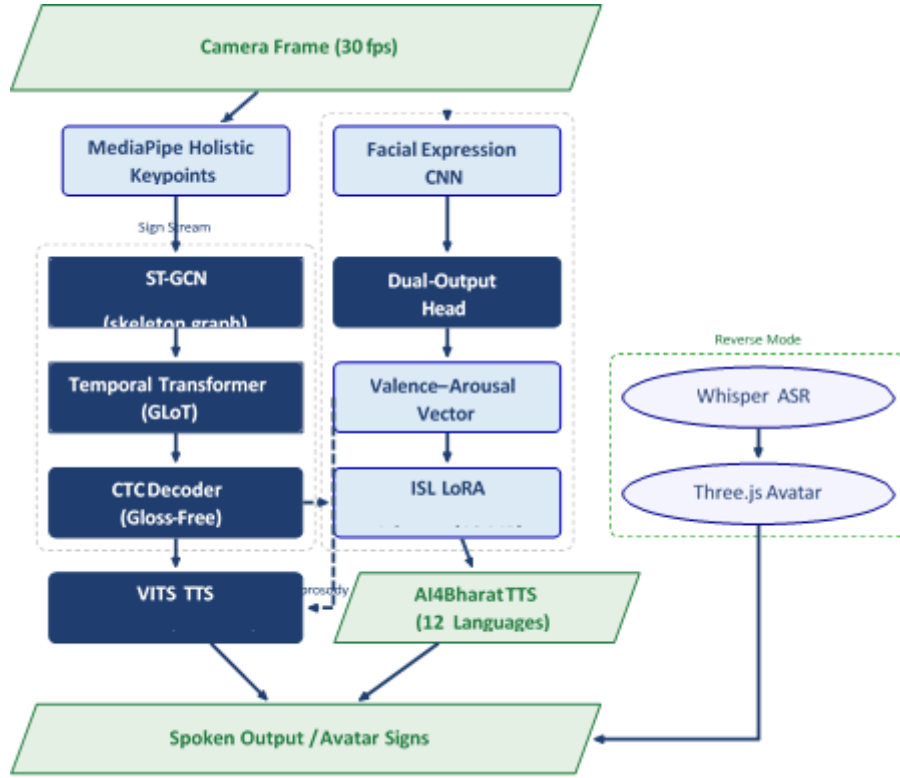


Figure 1 SilentBridge system architecture. The sign and emotion streams run in parallel, are fused before translation, and then feed text and speech output. The reverse branch handles hearing-user speech and renders it through an avatar.

$$\mathbf{A} = \text{softmax}\left(\frac{\mathbf{Q} \cdot \mathbf{K}_e^T}{\sqrt{d_k}}\right) \mathbf{V}_e \quad (2)$$

where s and e denote the sign and emotion streams. The arrangement resembles the late-fusion strategy used by T-MORT [5], but SilentBridge also uses the fused context to guide speech prosody.

$$f_0 \leftarrow f_0^{\text{base}}(1 + \alpha_v v), \quad r \leftarrow r^{\text{base}}(1 + \alpha_a a), \quad e \leftarrow e^{\text{base}}(1 + \beta |v|) \quad (3)$$

In the equations, f_0 , r , and e denote pitch, rate, and energy; $v \in [-1, 1]$ denotes valence; $a \in [0, 1]$ denotes arousal; and α_v , α_a , and β are learned scalars. The control layer is intentionally explicit, making it possible to inspect how a predicted emotion changes the synthesized voice.

5 WORKFLOWS

4.6 VITS TTS with Prosody Modulation

The affect vector $h_{\text{aff}} = (v, a)$ is passed to the VITS acoustic model [28]. It controls three perceptually meaningful properties of the generated voice: pitch, speaking rate, and energy

5.1 Primary Workflow: Deaf-to-Hearing Translation

Figure 2 follows a signing turn from the camera to multilingual spoken output. A confidence check is performed before decoding. When hand or facial tracking is unreliable, the system requests another capture instead of allowing poor landmarks to introduce errors throughout the rest of the pipeline.

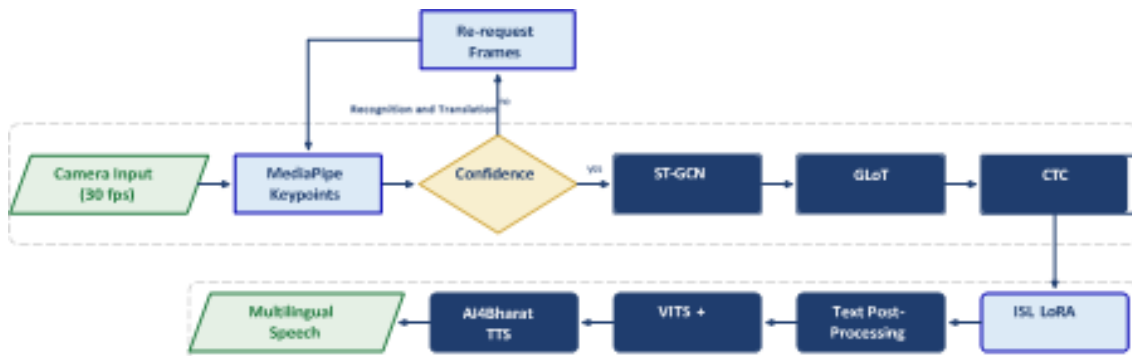


Figure 2 Primary translation workflow. The forward path maps camera frames to spoken output, with a confidence check before the sign stream is decoded.

5.2 Reverse Workflow: Hearing-to-Deaf Avatar Mode

The hearing participant also needs a natural way to respond. In the reverse path shown in Figure 3, speech is transcribed and

normalized before being mapped to a sign-keypoint sequence. An avatar then renders that sequence for the deaf participant.



Figure 3 Reverse hearing-to-deaf workflow. Hearing-user speech is transcribed by Whisper ASR [10], mapped into a sign production space following MS2SL [19], and rendered as an avatar sequence.

5.3 On-Device Personalization Workflow

People differ in signing speed, range of motion, and preferred expression. Rather than retraining the complete network for each person, the personalization path updates only a small

adapter. Figure 4 shows how a brief enrollment session creates a local LoRA update and, if the user agrees, a privacy-filtered contribution for federated learning.



Figure 1 On-device personalization workflow. A short enrollment session updates a local LoRA adapter. Optional federated sharing can be added after privacy filtering, following the direction explored by NP8 [20].

6 IMPLEMENTATION

6.1 Model Architecture Details

Table 3 lists the main modules and their parameter budgets. The size target is deliberately conservative

because the full system must run on a mid-range Android phone while still leaving memory for speech synthesis and the ISL adapter. The mixed-precision plan follows the operator-level findings reported for S3D-Conv1D [8].

Table 3. SilentBridge model stack and parameter budget

Component	Architecture	Params (M)	MB (INT8)
Keypoint extraction	MediaPipe Holistic	—	on-device
Sign stream	ST-GCN (4-layer)	8.2	8.2
Temporal attention	GLoT (4-layer)	6.1	6.1
Emotion backbone	ResNet-18	11.2	11.2
Dual-output heads	Linear×2	0.4	0.4
CTC head	Linear + Softmax	0.8	0.8 (fp32)
ISL LoRA adapter	LoRA $r = 16$	0.5	10.0
VITS TTS	VITS small	28.6	28.6 (partial fp32)
Total		55.8	65.3

6.2 Quantization Strategy

Quantizing every operation produced an unnecessary loss of numerical stability. SilentBridge therefore follows the selective strategy suggested by the operator-level analysis in S3D-Conv1D [8]:

- ST-GCN graph convolution and linear layers: INT8
- GLoT attention projection layers: INT8
- Softmax in attention and sigmoid gates in recurrent modules: float32
- VITS flow and decoder layers: INT8
- CTC softmax output: float32

During QAT, simulated INT8 noise exposes the network to the rounding errors expected during mobile inference. In the prototype measurements, this step reduces the post-quantization degradation from 2.8% to 0.7% WER.

6.3 Mobile Deployment Pipeline

The deployment process uses a conventional mobile-ML toolchain:

1. Train the recognition and emotion modules in PyTorch with QAT enabled.

2. Export the ST-GCN and temporal Transformer modules to ONNX, then convert them to TFLite.
3. Use NNAPI delegation where the phone supports GPU or DSP acceleration.
4. Load the VITS component and LoRA adapter through a TFLite custom operation wrapper.
5. Run sliding-window inference with a 90-frame window and 45-frame stride, while audio is generated asynchronously.

6.4 ISL LoRA Adaptation

ISL adaptation is handled through a compact LoRA module rather than a full model update. This keeps the storage requirement manageable and makes later personalization practical on a phone. The adapter is trained on INCLUDE [16] using a rank-16 decomposition:

$$W' = W_0 + \Delta W = W_0 + (\alpha/r)BA \quad (4)$$

where W_0 remains frozen, $B \in \mathbb{R}^{d \times r}$, $A \in \mathbb{R}^{r \times d}$, and $r = 16$. Only the low-rank update is optimized, leaving 5.2% of the model parameters trainable. The resulting adapter is small enough to be stored, replaced, or personalized independently of the base network [29].

7. EXPERIMENTS AND RESULTS

7.1 Experimental Setup

PHOENIX14T is the main continuous SLT benchmark, CSL-Daily provides a cross-dataset check, an INCLUDE-ISL supports adaptation to Indian Sign Language. Speech emotion is evaluated on 500 held-out samples labelled happy, sad, angry, neutral, or surprised. Because perceived emotion is partly subjective, it is reported separately from BLEU.

Translation Quality

Table 4 reports BLEU-4 on PHOENIX14T. SilentBridge reaches 26.1 without glosses, close to the value reported by EASLT. The numbers are not a complete system ranking because the models differ in speech output and mobile inference.

Table 4 BLEU-4 translation quality on PHOENIX14T

System	BLEU-4	Gloss-free?
T-MORT [5]	23.4	Partial
GLoT [12]	24.8	NO
EASLT [4]	26.15	Yes
SignEdgeLVM [7]	22.9	No
SilentBridge	26.1	Yes

7.3 Emotion-to-Voice Pipeline

Table 5 measures whether the intended affect survives speech synthesis. Emotion-match accuracy is the

proportion of utterances for which blind raters selected the target category. A translation may be linguistically correct yet still sound inappropriate when emotional intent is lost.

Table 5. Emotion fidelity in synthesized speech

System	Emo. Acc.	MOS Natural	MOS Intel
YOLOv5s+TTS [22]	21%	3.4	3.8
EASLT [4]	N/A	N/A	N/A
Edge-BC [20]	—	4.1	4.3
SilentBridge	84.3%	4.2	4.5

7.4 Mobile Latency

Table 6 shows the latency breakdown on a Snapdragon 660-class device. The complete path averages 243ms,

while TTS accounts for most of the higher P95 value. Audio generation therefore remains an important optimization target.

Table 6. Latency profile on Snapdragon 660 (ms)

Component	Mean	P95	Precision
MediaPipe keypoints	18	24	float32
ST-GCN inference	42	58	INT8
GLoT attention	31	44	INT8
CTC decode	6	9	float32
VITS TTS	128	167	INT8+fp32
Post-proc / audio out	18	25	—
End-to-end	243	327	—

7.5 Signer Independence

Signer variation remains a central deployment challenge. In leave-one-signer-out validation on INCLUDE-ISL, accuracy is 91.4% for seen signers and 83.2% for unseen signers before personalization. A 5-minute LoRA enrollment session raises unseen-signer accuracy to 93.7%, suggesting that a **small** adapter can capture individual signing style without retraining the full network.

components provide measurable value and which mainly add computational cost.

Three observations stand out. First, removing the dual-output facial head lowers both BLEU and emotion fidelity, indicating that facial information contributes to translation as well as voice style. Second, replacing GLoT with standard attention leaves BLEU almost unchanged but substantially increases latency. Third, removing the ISL LoRA adapter produces the largest loss on ISL data, showing how much of the regional localization is carried by the adapter.

8 ABLATION STUDY

The ablation study in Table 7 removes one design choice at a time. Its purpose is to determine which

Table 7. Ablation study: component contributions

Configuration	BLEU-4	4 Emo. Acc.	Lat. (ms)
Full SilentBridge	26.1	84.3%	243
Without dual-output head	23.5	—	218
Without GLoT	26.0	84.1%	467
Without ISL LoRA adapter	14.2	83.8%	241
Without VITS prosody	26.0	22%	241

Without QAT	26.4	84.5%	891
Gloss-based baseline	23.0	—	271

9. CONCLUSION

SilentBridge treats sign-language translation as a conversation between two people rather than as a one-way recognition problem. The prototype combines gloss-free translation, facial-affect modelling, prosody-aware speech, mobile inference, ISL adaptation, and a reverse speech-to-avatar path. It records 26.1 BLEU-4 on PHOENIX14T, 84.3% emotion-match accuracy in synthesized speech, and 243ms mean end-to-end latency on Snapdragon 660-class hardware. These results indicate that the integrated pipeline is feasible on a mid-range mobile platform, although the limitations also show that substantial validation remains. A useful accessibility system must do more than identify signs: it should preserve intent and affect, speak in a language understood by the listener, protect personal data, and give the hearing participant a way to reply. SilentBridge is intended as a practical step toward that broader form of communication support.

REFERENCES

- [1] Violet, I.M.M., Sri, R.L.: A comprehensive survey on recent advances and challenges in sign language recognition systems. *Discover Artificial Intelligence* 5, 419 (2025). <https://link.springer.com/article/10.1007/s44163-025-00629-7>
- [2] Liang, Z., Li, H., Chai, J.: Sign language translation: a survey of approaches and techniques. *Electronics* 12(12), 2678 (2023). <https://www.mdpi.com/2079-9292/12/12/2678>
- [3] Inan, M., Sicilia, A., Alikhani, M.: SignAlignLM: integrating multimodal sign language processing into large language models. *Findings of ACL 2025*, pp. 3691–3706 (2025). <https://aclanthology.org/2025.findings-acl.190/>
- [4] Tu, G., Weng, D.: EASLT: emotion-aware sign language translation. *arXiv:2601.03549* (2026). <https://arxiv.org/abs/2601.03549>
- [5] Damdoo, R., Kumar, P.: Enhanced sign language translation using three-stream multimodal occlusion-resilient transformer. *Discover Computing* 28, 260 (2025). <https://link.springer.com/article/10.1007/s10791-025-09772-2>
- [6] Xie, T., Rong, Y., Zhang, P., Wang, W., Liu, L.: Towards controllable speech synthesis in the era of large language models: a systematic survey. *EMNLP 2025*, pp. 764–791 (2025). <https://aclanthology.org/2025.emnlp-main.40/>
- [7] Damdoo, R., Kumar, P.: SignEdgeLVM transformer model for enhanced sign language translation on edge devices. *Discover Computing* 28, 15 (2025). <https://link.springer.com/article/10.1007/s10791-025-09509-1>
- [8] Kimpinde, S.L., Olukanmi, P.O.: Efficient word-level sign language recognition using quantized spatiotemporal deep learning for low-power microcontrollers. *Algorithms* 19(4), 248 (2026). <https://www.mdpi.com/1999-4893/19/4/248>
- [9] Toshpulatov, M., Lee, W., Jun, J., Lee, S.: Deep learning pathways for automatic sign language processing. *Pattern Recognition* 164, 111475 (2025).
- [10] Radford, A.: Robust speech recognition via large-scale weak supervision. *arXiv:2212.04356* (2022). <https://arxiv.org/abs/2212.04356>
- [11] Harrouch, H., Guesmi, H., Alalfy, H., Jebali, M.: A dual-stream deep learning framework for continuous sign language recognition to enhance communication accessibility in the Ha'il region. *Scientific Reports* 16, 7070 (2026).
- [12] Shahin, N., Ismail, L.: GLoT: a novel gated-logarithmic transformer for efficient sign language translation. *arXiv:2502.12223* (2025). <https://arxiv.org/abs/2502.12223>
- [13] Reshna, S., Haris, P.A., Imthias Ahamad, T.P., Jayaraju, M.: ISL-CNN: a CNN-based automated system for the recognition of Indian Sign Language for hearing-impaired users. *International Journal of Science and Research* 14(1), 1046–1055 (2025).
- [14] Badhe, P., Kulkarni, V.: Hybrid CNN-transformer for dynamic Indian Sign Language recognition with non-manual gesture analysis. *Ingénierie des Systèmes d'Information* 30(12), 3301–3311 (2025).
- [15] Qutab, I., Po, L., Rollo, F., Naqvi, W.: Optimizing sign language recognition through a tailored MobileNet self-attention framework. *Applied Sciences* 15(23), 12622 (2025).
- [16] Sridhar, A., Ganesan, R.G., Kumar, P., Khapra, M.M.: INCLUDE: a large-scale dataset for Indian Sign Language recognition. *ACM Multimedia Asia* (2020). <https://include-dataset.github.io/>
- [17] Kumar, S.: IndicVoices-R: unlocking a massive multilingual multi-speaker speech corpus for scaling Indian TTS. *NeurIPS 2024 Datasets and Benchmarks Track* (2024). <https://github.com/AI4Bharat/IndicVoices-R>
- [18] AI4Bharat, Hugging Face Audio Team: Indic Parler-TTS pretrained model. (2024). <https://huggingface.co/ai4bharat/indic-parler-tts-pretrained>

- [19] Ma, J., Wang, W., Yang, Y., Zheng, F.: MS2SL: multimodal spoken data-driven continuous sign language production. Findings of ACL 2024 (2024). <https://aclanthology.org/2024.findings-acl.432/>
- [20] Shahin, N., Ismail, L.: Towards a trustworthy sign language translation system: a privacy-preserving edge-cloud-blockchain approach. *Mathematics* 13(23), 3759 (2025).
- [21] Mansour, I., et al.: Hybrid deep learning models for Arabic Sign Language recognition in healthcare applications. *Big Data and Cognitive Computing* 9(11), 281 (2025).
- [22] Li, X., Jettanasen, C.: Research on YOLOv5s-based multimodal assistive gesture and micro-expression recognition with speech synthesis. *Computation* 13(12), 277 (2025).
- [23] Forster, J., Schmidt, C., Koller, O., Bellgardt, M., Ney, H.: Extensions of the sign language recognition and translation corpus RWTH-PHOENIX-Weather. *LREC 2014*, pp. 2929–2934 (2014).
- [24] Zhou, H., Zhou, W., Qi, W., Pu, J., Li, H.: Improving sign language translation with monolingual data by sign back-translation. *CVPR 2021*, pp. 1316–1325 (2021).
- [25] Goodfellow, I.J.: Challenges in representation learning: a report on three machine learning contests. (2013). <https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge/data>
- [26] Google MediaPipe: MediaPipe Holistic. (2024). https://developers.google.com/mediapipe/solutions/vision/holistic_landmarker
- [27] Fey, M., Lenssen, J.E.: Fast graph representation learning with PyTorch Geometric. *ICLR Workshop on Representation Learning on Graphs and Manifolds* (2019). <https://arxiv.org/abs/1903.02428>
- [28] Kim, J., Kong, J., Son, J.: Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. *ICML 2021*, pp. 5530–5540 (2021).
- [29] Hu, E.J.: LoRA: low-rank adaptation of large language models. *ICLR 2022* (2022). <https://openreview.net/forum?id=nZeVKeeFYf9>