

ISLiM: A Compact End-to-End Sign Language Translation Framework

Shoaib Ahmed and Dipti Mishra

Department of Computer Science & Engineering
MU-VT IARCT²

Mahindra University, Hyderabad, India

{se25pcse006, dipti.mishra}@mahindrauniversity.edu.in

Abstract

Sign Language Translation (SLT) is a key accessibility technology for Deaf communities. Although recent years have seen rapid progress in SLT, Indian Sign Language (ISL) and other sign languages of under-resourced regions remain under-researched and overshadowed by sign languages with a longer research history and stronger infrastructure. In this direction, we present ISLiM (ISL in Miniature), a compact end-to-end keypoint-based gloss-free SLT architecture that maps body-pose landmark sequences directly to spoken-language text. At its core, a stack of temporal convolutional layers captures low-level pose motion patterns, a Transformer-based encoder models higher-order kinematic interactions and sign-level temporal ordering, and a pretrained T5-small decoder with a prior grasp of the target language performs the semantic heavy-lifting of generating fluent text. We further explore complementary loss functions that sustain training on small datasets, and alternative qualitative evaluation metrics tailored to the particular challenges of SLT. ISLiM achieves a BLEU-4 score of 5.85 on iSign, establishing state-of-the-art results among reported keypoint-based models on the only large-scale ISL dataset. We also show promising transferability to German Sign Language on Phoenix-2014-T through modular decoder swap, pointing toward cross-lingual extensibility where both sign and spoken languages can be swapped independently. A preliminary mobile deployment study confirms that the full keypoint-based model runs on a consumer Android smartphone with ~ 384 ms mean inference latency on CPU, demonstrating a practical path toward offline, accessible assistive tools for the Deaf community in India and the world. Code is available at <https://github.com/ahmedshoaib/ISLiM>.

1 Introduction

Sign languages are the primary mode of communication for over 70 million Deaf people worldwide

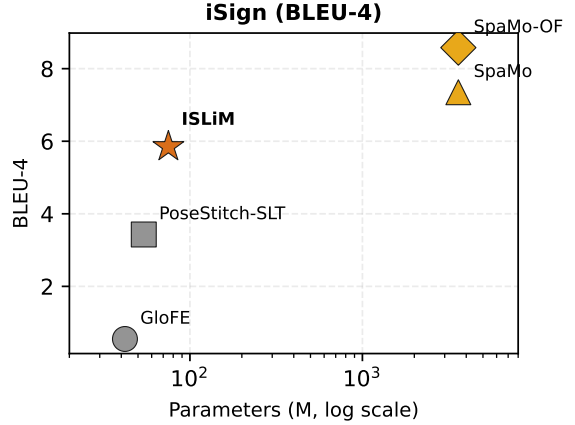


Figure 1: Parameter efficiency on iSign. ISLiM (75M, red star) improves reported keypoint SOTA by 71% (5.85 vs. 3.43). Keypoint-based methods shown as circles/squares; video-based SpaMo/SpaMo-OF as triangles/diamonds (SpaMo/SpaMo-OF only tested on 10k subset)

(WHO, 2026). India alone is home to roughly 7 million Deaf (ISLRTC, 2024a) and hard-of-hearing individuals, around 10% of the global population, yet only about 300 registered interpreters serve this community nationwide (ISLRTC, 2024b). This chronic shortage leaves most Deaf Indians without equitable access to education, employment, healthcare, and broader civic participation.

Recent advances in deep learning try to fill this shortage through Sign Language Translation (SLT), the task of mapping sign videos to spoken-language text. Large-scale SLT corpora such as iSign (Joshi et al., 2024), PHOENIX-2014-T (Camgoz et al., 2018a), How2Sign (Duarte et al., 2021), OpenASL (Shi et al., 2022), BOBSL (Albanie et al., 2021), and Multi-VSL (Dinh et al., 2025) have powered strong SLT systems. ISL, however, has just one large-scale dataset, iSign, reflecting a broader asymmetry: ISL and other sign languages of under-resourced regions remain under-researched and overshadowed by sign languages with longer re-

search history and far larger datasets (Ni et al., 2026).

Existing SLT methods have evolved along two main branches. Earlier work relies on gloss-level supervision (Camgöz et al., 2020; Yin and Read, 2020; Zhou et al., 2021; Chen et al., 2022; Alvarez et al., 2022), which provides a structured intermediate target but is costly to annotate, a particular burden for low-resource settings such as ISL where few glossed resources or qualified interpreters exist. A gloss-free approach learns a direct mapping from signed utterances to text (Zhou et al., 2023; De Coster and Dambre, 2022; Li et al., 2025) by leveraging pretrained vision-language encoders and language models. Keypoint-based approaches further improve practicality: pose landmarks are privacy-preserving, lightweight to extract, and offer an explicit geometric encoding of body-part locations that are central to sign language conveyance (Lin et al., 2023). More recently, large language models have entered the SLT space (Gong et al., 2024; Wong et al., 2024; Chowdhury and Sanyal, 2025) and promise strong translation quality, but their training and inference are compute-heavy and typically assume continuous network access, assumptions that do not hold for many Indian users with limited connectivity and low-end devices (Kumar et al., 2022).

We address these constraints with ISLiM (ISL in Miniature), a compact end-to-end keypoint-based gloss-free SLT architecture. ISLiM pairs a small visual encoder with a pretrained T5-small (Raffel et al., 2020) decoder that already understands English structure and supplies much of the semantic heavy-lifting the encoder alone cannot provide. Each frame’s landmark vector passes through stacked 1D temporal convolutions (Lea et al., 2017) that capture local pose-motion patterns, then a multi-layer self-attention Transformer (Vaswani et al., 2017) integrates higher-order kinematic interactions and resolves sign-level temporal ordering before the frame-level representations flow into T5 cross-attention, which produces autoregressive English output. This division of labor lets the lightweight encoder (14.5M of 75M total parameters) focus on visual feature extraction while the pretrained decoder handles fluency and word ordering, keeping the model compact enough for edge deployment. Figure 1 visualizes the parameter–quality trade-off for iSign: it attains BLEU-4 5.85 on our maintained split, against a published 3.43 for the prior keypoint-based state of the art at a

comparable parameter scale. Because the baselines were evaluated under their own undisclosed splits, the comparison is indicative.

ISL presents a unique set of challenges as compared to other Sign languages, each of which informs a specific design choice in our architecture. Its vocabulary exhibits significant regional variation, emerging from a patchwork of local educational traditions (Jayachandran*, 2026). Consequently, the encoder must remain robust to multiple valid signs of the same gloss. We therefore employ a Transformer encoder, whose self-attention mechanism operates over temporally ordered frame-level representations rather than individual keypoint tokens; this allows the encoder to model long-range temporal dependencies directly, without the neighbourhood-constrained message passing of graph-based encoders (Yan et al., 2018), complementing the local motion patterns captured by the temporal convolutions that precede it. ISL distinguishes many signs using subtle contact locations on the face and torso (Sinha, 2018). To preserve these distinctions, each frame is represented using an expanded set of 94 landmarks, including 46 facial landmarks, together with estimated depth (z) and landmark visibility. This richer representation captures facial expressions, head pose, and mouth movements, which encode grammatical information such as negation, interrogatives, and topicalization that cannot be inferred from hand trajectories alone. Crucially, ISL belongs to the SOV family while English follows SVO order (Napoli and Sutton-Spence, 2014; Sugandhi et al., 2020): verbs, question markers, and negation appear late in a signed utterance but must reorder to earlier English positions (Ghosh and Mamidi, 2022). A faithful gloss-free translation therefore cannot be produced incrementally as the video plays; the model must observe the entire sign sequence before committing to any word of the English output. Our Transformer-based encoder layers learn this non-monotonic alignment (e.g., “I school go” becomes “I go to school”; “Your name what?” becomes “What is your name?”) before any English token is emitted, motivating a global-attention architecture over a recurrent or frame-by-frame approach.

Our main contributions are:

- We introduce ISLiM, a compact 75M-parameter keypoint-based gloss-free end-to-end SLT architecture, achieving a test BLEU-4 score of **5.85** and establishing a new state

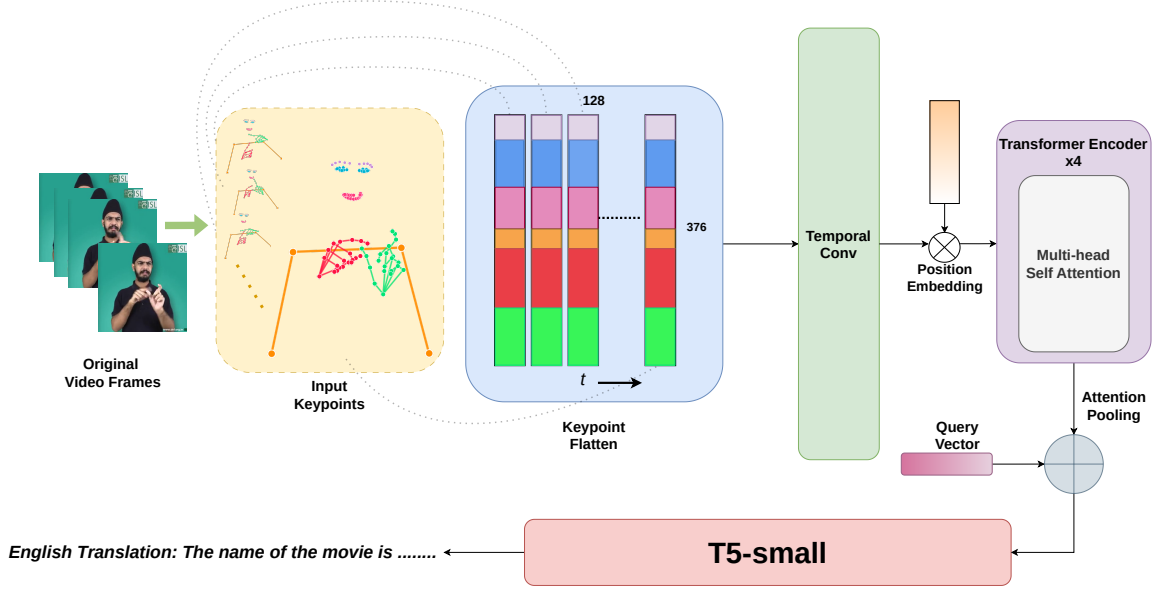


Figure 2: **ISLiM architecture overview.** The sequence of 128 frame-level representations is passed through the TCN and Transformer Encoder to the T5-small model, whose decoder generates the English translation autoregressively

of the art on iSign among reported keypoint-based models.

- We show that a cross-lingual transfer setup achieves promising results on Phoenix-2014-T (German Sign Language) at **BLEU-4 14.99**, pointing toward a modular architecture supporting multi-lingual support.
- We propose a multi-loss objective training strategy that combines cross-entropy with CLIP contrastive, semantic similarity, and entropy-regularization losses, preventing the collapse observed with single-loss baselines on smaller iSign subsets.
- We conduct a preliminary on-device deployment study on a consumer Android smartphone, reporting **384 ms** mean CPU inference latency, confirming offline feasibility without cloud connectivity for keypoint-based models.

2 Related Work

2.1 Gloss-Based Sign Language Translation

Traditional SLT systems have relied on an intermediate gloss representation, either as a necessity (Camgöz et al., 2020; Yin and Read, 2020; Zhou et al., 2021) or a desirable (Chen et al., 2022; Alvarez et al., 2022). However, manual gloss annotation is costly, and gloss vocabularies often vary

across sign languages, datasets, and annotation conventions, limiting their reuse (Colonna et al., 2026). These challenges have motivated end-to-end gloss-free SLT approaches, especially for low-resource sign languages such as ISL, where obtaining large-scale gloss annotations is impractical due to the limited availability of annotated datasets and expert sign language annotators.

2.2 Gloss-Free Sign Language Translation

Gloss-free SLT, also called end-to-end SLT, learns a direct mapping from sign video to spoken language text without requiring intermediate gloss labels (Zhou et al., 2023; De Coster and Dambre, 2022; Li et al., 2025). While the annotations in this mode are weaker than their Gloss-based counterparts (Ye et al., 2024), this paradigm has gained momentum as large-scale datasets and video-language models have become available.

Video-Based Gloss-Free Models Several recent gloss-free systems (Ye et al., 2024; Mao et al., 2026) employ large video encoders pretrained on massive vision-language corpora, combined with powerful frozen language model decoders. These approaches typically use architectures such as CLIP (Contrastive Language-Image Pre-Training) (Radford et al., 2021; Jiang et al., 2024) and Video Masked Auto Encoder (VideoMAE) (Tong et al., 2022; Chowdhury and Sanyal, 2025) to encode sign video frames, then condition a pretrained text

decoder on the visual features through adapter modules or cross-attention (Asasi et al., 2025; Hwang et al., 2025).

Keypoint-Based Gloss-Free Models An alternative approach uses skeletal keypoints rather than raw video as the input modality (Lin et al., 2023). Keypoint-based approaches extract body pose landmarks from each video frame using standard pose estimation tools, then process the resulting coordinate sequences with lightweight neural networks. This paradigm offers several advantages for deployment: keypoints are compact, more privacy-preserving than storing or processing raw video, and can be computed efficiently on-device. The current state of the art (Joshi et al., 2025) on the iSign benchmark uses a keypoint-based architecture with a standard transformer (Vaswani et al., 2017) operating on two-dimensional joint coordinates. Hybrid methods that combine keypoints with image crops have also been explored (Jang et al., 2026). Our work extends the keypoint-only direction by incorporating additional keypoint channels (depth and visibility) and leveraging a stronger pretrained decoder, while maintaining a focus on computational efficiency and mobile deployability.

3 Methodology

This section presents the architecture and training methodology of ISLiM. We begin by formalizing the SLT problem, then describe the input data and pre-processing steps, followed by the encoder-decoder design. The overall architecture of the proposed framework is shown in Figure 2.

3.1 Problem Definition

SLT, unlike conventional Machine Translation, translates a given Sign Language Video into Spoken Language Text. More formally, given a sequence of T video frames $V = \{frame_1, frame_2, frame_3, \dots, frame_T\}$, we attempt to generate a spoken-language translation $S = \{word_1, word_2, word_3, \dots, word_n\}$ in the target language.

3.2 Input Pre-processing

We use Mediapipe Holistic (MediaPipe) to extract 94 upper-body landmarks per frame (see Figure 3 for a comparison with the 76-keypoint representation used by PoseStitch-SLT) (21 per hand, 6 body-pose joints, 46 face points), each with (x, y, z, visibility), forming a 376-dimensional vector.

Sequences are normalized per-axis to $[0, 1]$ and padded or sub-sampled to 128 frames.

All reference translations and inferred translations are processed according to the criteria defined in the PoseStitch paper (Joshi et al., 2025) including: lowercasing, contraction expansion, punctuation removal, digit-by-digit number splitting, person name anonymization, and replacing words with fewer than 3 occurrences with <UNK>. Pre-processing is applied to all input references and outputs generated by the model to maintain uniformity when comparing. The rare-word dictionary is computed on the training split only and applied, frozen, to the validation and test references.

3.3 Encoder

The encoder transforms keypoint sequences into contextualized representations in two stages. The visual encoder focuses purely on pose feature extraction

Temporal Convolution Each frame’s 376-dimensional vector is passed through two stacked 1D convolutional layers (kernel size 5 along the time axis), reducing dimensionality to 256. For mean-duration videos (~ 5 seconds at 25 fps, sub-sampled to 128 frames), the kernel spans approximately 200 ms of signing; for longer videos the effective span is wider due to subsampling. These layers capture short-range motion patterns, while longer-range temporal dependencies are handled by the transformer encoder that follows.

Self-Attention Transformer The convolution output is projected to 512 dimensions, combined with learned positional embeddings, and processed by a 4-layer transformer encoder. Self-attention enables the non-monotonic SOV-to-SVO alignment without imposing a strict ordering prior on the encoder.

3.4 Decoder

We fine-tune a pretrained T5-small (Raffel et al., 2020) model as the decoder. We pass the full 128-frame representation sequence into the T5 model, which conditions on the temporally aligned frames through cross-attention rather than compressing them into a single vector. This design offloads the temporal alignment and translation burden to the pretrained decoder. For cross-lingual transfer to German, we initialize from the iSign-trained checkpoint, swap the decoder for a German T5 (S et al., 2022) and fine-tune on Phoenix-2014-T.

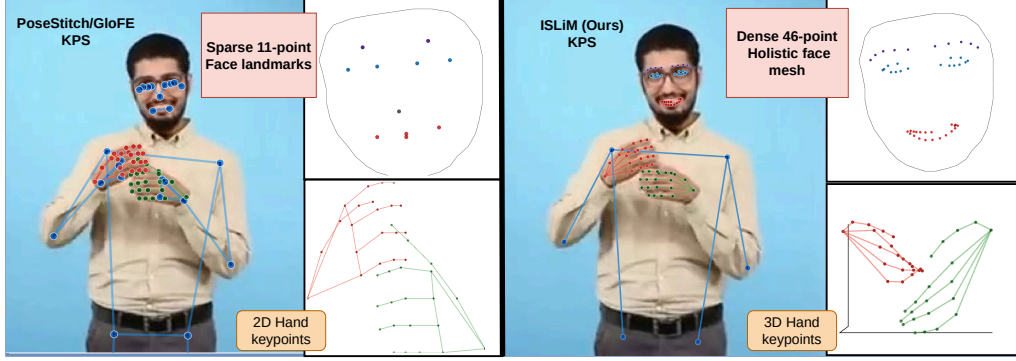


Figure 3: **Keypoint comparison of PoseStitch-SLT ($76 \times 2D$) and ISLiM ($94 \times 4D$).** Left: 76 landmarks (33 full-body pose incl. legs/feet (including **11 face**), 42 hands, 1 face reference), x,y only, no depth or visibility. Right: 94 landmarks (6 upper-body pose, 42 hands, **46 face mesh points** capturing lips, eyes, and eyebrows), plus estimated z-depth and per-landmark visibility scores. Same signing frame shown on both sides.

3.5 Multi-Loss Training Strategy

On subsets of iSign smaller than approximately 15K samples, CE-only training regularly collapses: the decoder converges to outputting one or two high-frequency tokens for every input, and once this happens, the gradient signal from CE is too weak to recover. We address this with three auxiliary losses, each mitigating a different aspect of the collapse. A **CLIP contrastive loss** aligns pose and text embeddings in a shared latent space, forcing the encoder to produce distinct representations for different inputs rather than collapsing to a single point. A **Semantic Similarity Loss** compares the decoder’s hidden state to a SentenceBERT (Reimers and Gurevych, 2019; Hamidullah et al., 2024) reference embedding, providing a gradient signal even when token-level overlap (Papineni et al., 2002) is zero. **Entropy Regularization** penalizes the output distribution when it becomes too peaked too early, keeping the decoder in an exploratory regime for longer. Detailed formulations, weights, and ramping schedules are provided in Section 4.2.

4 Implementation Details

4.1 Dataset

The iSign dataset (Joshi et al., 2024) contains 127,237 parallel ISL-to-English samples collected from Deaf signers and interpreters across India, covering news, daily conversation, and educational domains. We apply the PoseStitch (Joshi et al., 2025) text preprocessing pipeline to all references and create our own 90/5/5 random split (seed = 42) as the PoseStitch data split is not available. The split is per-clip, as iSign provides no signer-

level metadata; the sample identifiers, which encode each clip’s source video, are released with the code.

The iSign 10K subset, introduced in SpaMo-OF (Chowdhury and Sanyal, 2025), is a curated 10,000-sample subset used for controlled small-data experiments. Phoenix-2014-T provides German weather-forecast translations used for cross-lingual transfer experiments; its constrained vocabulary makes it a dramatically easier benchmark, as discussed in Section 5.2.

4.2 Training Objectives and Hyperparameters

Primary objective : Cross-Entropy (CE) Standard next-token prediction with label smoothing ($\epsilon = 0.1$). This is the sole training objective when sufficient data is available (full iSign, 113K training samples).

Multi-loss Configuration Each auxiliary loss contributes to the combined objective.

The **CLIP contrastive loss** aligns mean-pooled encoder features $V \in \mathbb{R}^{B \times d}$ with mean-pooled decoder text embeddings $T \in \mathbb{R}^{B \times d}$:

$$S = (V \cdot T^\top) \cdot \tau,$$

$$\mathcal{L}_{\text{CLIP}} = -\frac{1}{2} \left(\sum_i \log \frac{e^{S_{ii}}}{\sum_j e^{S_{ij}}} + \sum_i \log \frac{e^{S_{ii}}}{\sum_j e^{S_{ji}}} \right)$$

where $\tau = \exp(2.6592) \approx 14.3$ is a learnable temperature. The computation is performed in float32 to prevent an observed float16 drift in the similarity matrix.

The **Semantic Similarity loss** guides the decoder output toward the reference in Sentence-BERT embedding space:

$$\mathcal{L}_{\text{sem}} = 1 - \cos(\text{proj}(\mathbf{h}_{\text{dec}}), \mathbf{e}_{\text{ref}})$$

where proj is a learned linear layer $512 \rightarrow 384$ followed by ℓ_2 normalization, and $\mathbf{e}_{\text{ref}}$ is a precomputed embedding from all-MiniLM-L6-v2 (Reimers and Gurevych, 2019). For German Phoenix, the multilingual variant paraphrase-multilingual-MiniLM-L12-v2 (Reimers and Gurevych, 2020) replaces it.

Entropy regularization penalizes overconfident per-token distributions:

$$H = - \sum_v p_v \log p_v,$$

$$\mathcal{L}_{\text{ent}} = \frac{1}{N} \sum_{t=1}^N \max(0, 3.0 - H_t)^2$$

masked over non-padding positions.

The combined objective is:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + 1.0 \cdot \mathcal{L}_{\text{CLIP}} + 0.05 \cdot \mathcal{L}_{\text{sem}} + 0.01 \cdot \mathcal{L}_{\text{ent}}$$

CLIP and semantic weights ramp linearly from 0 to their final values over the first 10 epochs. The weights are determined empirically: a CLIP weight of 1.0 is required to prevent encoder collapse (lower values 0.1–0.5 are insufficient), while semantic and entropy weights are kept low to serve as regularizers rather than dominant objectives.

Optimization. We use AdamW (Loshchilov and Hutter, 2019) ($\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay = 0.01), following standard practice for fine-tuning Transformer-based language models. The learning rate follows a linear warmup to stabilize early training, then decays via a cosine schedule, providing a smooth annealing trajectory without abrupt drops at epoch boundaries. We train with a per-device batch size of 4 and gradient accumulation over 4 steps, yielding an effective batch size of 16; the small per-step allocation keeps peak VRAM below 5 GB, making full-dataset training feasible on a single consumer GPU or a small shared cloud instance. Automatic mixed precision (float16) further reduces memory and accelerates tensor-core operations, with the CLIP loss computed in float32 to avoid an observed stability drift in the

similarity matrix. Under this configuration, training on the full 113K iSign samples requires approximately 90 GPU-hours.

4.3 Experimental Setup

iSign full-dataset training. For iSign full-dataset training, the visual encoder is initialised at random and the T5-small from pretrained weights; training proceeds on the 113K-sample PoseStitch-preprocessed split with CE-only loss. Training runs for 142 epochs; we observed convergence by epoch 103, with negligible improvement thereafter. The best checkpoint is selected by validation BLEU4.

Small-data (iSign 10K (from SpaMo-OF)).

The full 4-loss objective is trained for 500 epochs under the same initialization scheme.

Cross-lingual transfer (Phoenix-2014-T). The ISLiM checkpoint trained on full iSign is fine-tuned on Phoenix-2014-T (German) for 163 epochs. The T5-small decoder is replaced with GermanT5-efficient-small; encoder weights transfer directly.

Keypoint normalization. Each of the three spatial channels (x, y, z) is independently min-max normalized to $[0, 1]$ per video on a per-axis basis. Normalization is applied after keypoint extraction and before padding or temporal subsampling to 128 frames. This per-video normalization removes dependence on signer position, camera distance, and frame resolution.

Evaluation metrics. We report multiple BLEU variants across experiments, each serving a different comparison purpose. Two scoring functions are used: (i) PoseStitch BLEU (PS-B4) (originally from (Camgoz et al., 2018b)), the `compute_bleu()` implementation from PoseStitch with add-1 smoothing and 4-gram precision (words tokenized by `str.split()`), and (ii) SacreBLEU (Post, 2018), the standard corpus-level BLEU with `tokenize='13a'` and exp smoothing. On iSign, the primary metric is PS-B4, matching the PoseStitch-SLT evaluation protocol exactly. On SpaMo-OF, we use SacreBLEU to align with the SpaMo-OF paper. We additionally report BLEU-1 through BLEU-3 (sacrebleu), ROUGE-L (Lin, 2004) (Google `rouge_score` with Porter stemmer), METEOR (Banerjee and Lavie, 2005)(NLTK), chrF++ (sacrebleu, `word_order=2`), semantic similarity (SentenceTransformer all-MiniLM-L6-v2 cosine).

We find several different BLEU implementations that differ in smoothing (add-1 vs. exp vs. none), tokenization, and aggregation (corpus vs. per-sentence). We discuss the empirical impact of these choices in Section 5.2 alongside the results.

5 Results

5.1 Comparison with Prior Work

Table 1 compares ISLiM against prior work across three SLT benchmarks. GloFE and PoseStitch-SLT report results obtained under their own splits; ISLiM is evaluated on our released 90/5/5 split (seed = 42), with references and outputs processed identically. ISLiM attains a BLEU-4 of **5.85** on iSign, exceeding the published PoseStitch-SLT result (3.43) at a comparable parameter scale (75M vs. 54M); since the protocols differ, the comparison is indicative, and the controlled contribution is the preprocessing ablation, which shows the identical baseline already exceeds the published value (2.24 to 4.10) when evaluated under the PoseStitch pipeline (Section 5.2). The bootstrap 95% confidence interval for sacreBLEU-4 is [5.53, 6.15] (1,000 resamples over test examples), indicating stability of the estimate with respect to test sampling. Training seed variability is not reported. No large-scale visual or video pretraining is employed. Figure 1 plots the parameter-efficiency trade-off. A per-metric breakdown is given in Appendix A.

On the iSign 10K subset (SpaMo-OF), ISLiM’s 75M keypoint-based model achieves BLEU-4 of 4.12, below the 8.58 of the 3.6B video-based SpaMo-OF but substantially higher than the SLT(GF) and GFSLT-VLP video baselines presented in the paper (Chowdhury and Sanyal, 2025). For cross-lingual transfer, ISLiM is first trained on full iSign (English ISL), then fine-tuned on Phoenix-2014-T (German DGS) with only the decoder swapped for a German T5. The fine-tuned 75M model achieves BLEU-4 14.99 vs. SpaMo-OF’s reported 14.72, performing comparably to the 3.6B SpaMo-OF on Phoenix with a $48\times$ parameter disadvantage. This suggests that keypoint features capture enough language-agnostic signing structure to transfer across unrelated sign languages. We also note that video-based models are not yet benchmarked on the full iSign dataset; even the SLT (GF) baseline degrades sharply on the full iSign set (BLEU-4 0.32, per SpaMo-OF’s results), reflecting the greater difficulty of the broader-domain benchmark.

5.2 Dataset Quality and Metric Reliability

Dataset quality The iSign dataset covers a broad range of topics: topic modeling reveals over 1,700 distinct themes spanning news, sports, education, entertainment, health, and cultural content. It includes content related to Deaf education and accessibility. Like many large-scale sign language datasets, it exhibits a long-tail vocabulary distribution: 48.2% of unique words appear exactly once, and 77.6% appear five or fewer times. With 68,029 unique words across 113K training samples, this distribution reflects the diversity of the domain rather than a data quality issue. For the SLT task, this means that some evaluation references contain vocabulary that is inherently difficult to model from limited occurrences, contributing to the observed 20–21% zero-BLEU rate. This highlights an opportunity for future dataset efforts to consider vocabulary coverage and multiple references per video, which would support more reliable model evaluation.

The iSign-10K subset in SpaMo-OF provides a useful controlled setting for small-data experiments and multi-loss ablation, enabling comparisons that would be computationally prohibitive on the full dataset. We note that its topic distribution skews more towards education and learning relative to the full iSign dataset, and it contains some near-duplicate sentences across train and test splits. These characteristics are worth considering when interpreting cross-dataset comparisons and point toward the value of designing future evaluation subsets with explicit coverage targets and controlled overlap (Alkain et al., 2026).

In general, many large-scale SLT datasets are constructed by chopping long, continuous signing videos into shorter clips. While pragmatic, such segmentation can sever contextual cues. Tanzer et al. (Tanzer et al., 2024) caution against over-reliance on sentence-segmented references as ground truth. We align with this caution and treat sentence-segmented references as a practical, though imperfect, training and evaluation target rather than as a gold standard of signing intent.

Because SLT datasets typically consist of single video-sentence pairs rather than conversational chunks, we find sentence-level BLEU (sent-BLEU) more informative than corpus BLEU for detecting collapse in our setting. We also observe that Semantic Similarity (SemSim) provides a complementary signal by evaluating meaning preservation

Method	KP-dim	KP	Video	BLEU-4	ROUGE-L	Params
iSign (full) Dataset						
GloFE	2D	✓		0.55	9.08	42M
PoseStitch-SLT	2D	✓		3.43	12.50	54M
ISLiM (ours)	4D	✓		5.85	20.12	75M
iSign 10K Subset (SpaMo-OFF) Dataset						
SLT (GF)	—		✓	0.73	9.08	~40M
GFSLT-VLP	—		✓	1.93	9.24	~680M
SpaMo	—		✓	7.35	16.23	~3.6B
SpaMo-OFF	—		✓	8.58	18.98	~3.6B
ISLiM (ours)	4D	✓		4.12	12.66	75M
Phoenix-2014-T Dataset						
SpaMo-OFF	—		✓	14.72	33.52	~3.6B
ISLiM (ours)	4D	✓		14.99	35.06	75M

Table 1: ISLiM vs. prior work. Numbers from (Chowdhury and Sanyal, 2025) except iSign GloFE/PoseStitch from (Joshi et al., 2025). KP-dim = keypoint dimensionality; KP = keypoint-based; Video = video-based. ISLiM improves previously reported iSign keypoint SOTA by 71% and matches the 3.6B SpaMo-OFF on Phoenix-14T at $48\times$ fewer parameters. Bootstrap 95% CIs (1,000 resamples): iSign sacreBLEU-4 5.84 [5.53, 6.15].

independently of exact n-gram overlap, particularly valuable in our setting where hapax words or synonym choice can inflate zero-BLEU rates. However, SemSim has its own weakness: risk of moderate similarity to generic English sentences. In our experiments, combining sent-BLEU and SemSim formed a practical evaluation kit: sent-BLEU captures lexical overlap, while SemSim captures semantic adequacy.

Mobile feasibility. Preliminary deployment on a standard Android phone shows promising results: the model achieves translation quality comparable to desktop beam-search decoding with 384 ms mean inference latency (see Appendix D for details).

6 Conclusion

We present ISLiM, a compact keypoint-based gloss-free SLT architecture that achieves state-of-the-art results among reported keypoint-only models on iSign. Our findings underscore three practical lessons: careful text preprocessing can dominate reported BLEU scores and must be rigorously controlled for fair comparison; pretrained language models supply syntactic fluency that would be impractical to learn from sign-language data alone; and lightweight keypoint-based models can run on consumer smartphones at acceptable latency. Cross-lingual transfer to German Sign Language on

Phoenix-2014-T further demonstrates that a shared visual encoder paired with language-specific decoders is a viable path toward multilingual SLT.

6.1 Future Work

Key directions include: (i) new ISL datasets with controlled vocabulary coverage and multiple references per video; (ii) SLT-specific evaluation protocols that account for high hapax rates and single-sentence settings; (iii) encoder pretraining on large-scale keypoint corpora from other sign languages or general body-motion data; and (iv) identifying robust metric settings or a combination of metrics that capture lexical overlap, semantic adequacy, and fluency in a single reliable evaluation framework for SLT.

7 Limitations

- **Single-reference evaluation.** iSign provides one reference per video; multi-reference evaluation would allow more reliable assessment when multiple valid translations exist.
- **No human evaluation.** All improvements are measured by automatic metrics, which may over-reward fluent but factually incorrect output. Human judgment is needed to confirm that measured gains correspond to genuine quality improvements.
- **Keypoint fidelity.** MediaPipe Holistic may

lose some non-manual markers that carry grammatical information in ISL.

- **Failure mode taxonomy.** Zero-BLEU predictions include a mix of collapse and paraphrases; a quantitative breakdown is left to future work.

8 Ethical Considerations

- **Data consent.** The iSign dataset is collected from ISL signers and interpreters in India. We rely on the dataset creators’ assurances regarding consent.
- **Privacy.** Our model uses MediaPipe keypoints rather than raw video, discarding identifiable pixel content. Keypoint sequences may nevertheless retain identifying information such as body proportions, facial geometry, and signing style, and no formal privacy guarantee is claimed.
- **Intended use.** ISLiM is a research tool, not a replacement for human interpreters.
- **Dialect coverage.** ISL has regional variation across India. The iSign dataset may not represent all dialects equally.

References

- Samuel Albanie, Gül Varol, Liliane Momeni, Hannah Bull, Triantafyllos Afouras, Himel Chowdhury, Neil Fox, Bencie Woll, Rob Cooper, Andrew McParland, and Andrew Zisserman. 2021. BOBSL: BBC-Oxford British Sign Language Dataset. *arXiv*.
- Bittor Alkain, Adrián Núñez-Marcos, Carlos Escolano, Laura Docío-Fernández, Olatz Perez-de Viñaspre, and Gorka Labaka. 2026. [Critical analysis of datasets for sign language translation](#). *Frontiers in Artificial Intelligence*, Volume 9 - 2026.
- Patricia Cabot Alvarez, X. Nieto, and Laia Tarres Benet. 2022. [Sign language translation based on transformers for the how2sign dataset](#).
- Sobhan Asasi, Mohamed Ilyes Lakhal, and Richard Bowden. 2025. [Hierarchical feature alignment for gloss-free sign language translation](#). In *Adjunct Proceedings of the 25th ACM International Conference on Intelligent Virtual Agents*, IVA Adjunct ’25, New York, NY, USA. Association for Computing Machinery.
- Satanjeev Banerjee and Alon Lavie. 2005. [METEOR: An automatic metric for MT evaluation with improved correlation with human judgments](#). In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.
- Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018a. Neural sign language translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018b. Neural sign language translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7784–7793.
- Necati Cihan Camgöz, Oscar Koller, Simon Hadfield, and R. Bowden. 2020. [Sign language transformers: Joint end-to-end sign language recognition and translation](#). *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10020–10030.
- Yutong Chen, Fangyun Wei, Xiao Sun, Zhirong Wu, and Stephen Lin. 2022. [A simple multi-modality transfer learning baseline for sign language translation](#). *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5110–5120.
- Anal Roy Chowdhury and Debarshi Kumar Sanyal. 2025. [Enhancing Indian Sign Language translation via motion-aware modeling](#). In *Proceedings of the Workshop on Sign Language Processing (WSLP)*, pages 39–46, IIT Bombay, Mumbai, India (Co-located with IJCNLP–AAACL 2025). Association for Computational Linguistics.
- Emanuele Colonna, Ivan Rinaldi, David Landi, Gennaro Vessio, and Giovanna Castellano. 2026. [Handscribe: A gloss-free framework for sign language translation and gloss sequence generation](#). *Computer Vision and Image Understanding*, 265:104674.
- Mathieu De Coster and Joni Dambre. 2022. [Leveraging frozen pretrained written language models for neural sign language translation](#). *Information*, 13(5).
- Nguyen Son Dinh, Tuan Dung Nguyen, Duc Tri Tran, Nguyen Dang Huy Pham, Thuan Hieu Tran, Ngoc Anh Tong, Quang Huy Hoang, and Phi Le Nguyen. 2025. [Sign language recognition: A large-scale multi-view dataset and comprehensive evaluation](#). In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 7887–7897.
- Amanda Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier Giro-i Nieto. 2021. [How2sign: A large-scale multimodal dataset for continuous american sign language](#). In *2021 IEEE/CVF Conference*

- on *Computer Vision and Pattern Recognition (CVPR)*, pages 2734–2743.
- Abhigyan Ghosh and Radhika Mamidi. 2022. [English to Indian Sign Language:rule-based translation system along with multi-word expressions and synonym substitution](#). In *Proceedings of the 19th International Conference on Natural Language Processing (ICON)*, pages 123–127, New Delhi, India. Association for Computational Linguistics.
- Jia Gong, Lin Geng Foo, Yixuan He, Hossein Rahmani, and Jun Liu. 2024. [LLMs are Good Sign Language Translators](#). In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18362–18372, Los Alamitos, CA, USA. IEEE Computer Society.
- Yasser Hamidullah, Josef van Genabith, and Cristina España-Bonet. 2024. [Sign language translation with sentence embedding supervision](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 425–434, Bangkok, Thailand. Association for Computational Linguistics.
- Eui Jun Hwang, Sukmin Cho, Junmyeong Lee, and Jong C. Park. 2025. [An efficient gloss-free sign language translation using spatial configurations and motion dynamics with LLMs](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3901–3920, Albuquerque, New Mexico. Association for Computational Linguistics.
- ISLRTC. 2024a. [Annual report 23-24 | indian sign language research and training center ,government of india | india](#).
- ISLRTC. 2024b. [Isl interpreter list | indian sign language research and training center ,government of india | india](#).
- Youngjoon Jang, Liliane Momeni, Zifan Jiang, Joon Son Chung, Gül Varol, and Andrew Zisserman. 2026. Lost in translation, found in embeddings: Sign language translation and alignment. *arXiv*.
- P Jayachandran*. 2026. Indian sign language: A natural, fully developed language of global stature and cultural depth.
- Zifan Jiang, Gerard Sant, Amit Moryossef, Mathias Müller, Rico Sennrich, and Sarah Ebling. 2024. [Sign-CLIP: Connecting text and sign language by contrastive learning](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9171–9193, Miami, Florida, USA. Association for Computational Linguistics.
- Abhinav Joshi, Romit Mohanty, Mounika Kanakanti, Andesha Mangla, Sudeep Choudhary, Monali Barbate, and Ashutosh Modi. 2024. [iSign: A benchmark for Indian Sign Language processing](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10827–10844, Bangkok, Thailand. Association for Computational Linguistics.
- Abhinav Joshi, Vaibhav Sharma, Sanjeet Singh, and Ashutosh Modi. 2025. [PoseStitch-SLT: Linguistically inspired pose-stitching for end-to-end sign language translation](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 13834–13853, Suzhou, China. Association for Computational Linguistics.
- Shruthi Koratagere Anantha Kumar, Gangavarapu Vigneswara Ithita, Sachin Chaudhari, and Paventhan Arumugam. 2022. [A survey on rural internet connectivity in india](#). In *2022 14th International Conference on COMMunication Systems and NETWORKS (COM-SNETS)*, pages 911–916.
- Colin Lea, Michael D. Flynn, René Vidal, Austin Reiter, and Gregory D. Hager. 2017. [Temporal convolutional networks for action segmentation and detection](#). In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1003–1012.
- Peike Li, Trevor Cohn, and Jianyuan Guo. 2025. [Bridging sign and spoken languages: Pseudo glossgeneration for sign language translation](#). In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Chin-Yew Lin. 2004. [Rouge: A package for automatic evaluation of summaries](#). In *Annual Meeting of the Association for Computational Linguistics*.
- Kezhou Lin, Xiaohan Wang, Linchao Zhu, Ke Sun, Bang Zhang, and Yi Yang. 2023. [Gloss-free end-to-end sign language translation](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12904–12916, Toronto, Canada. Association for Computational Linguistics.
- Ilya Loshchilov and Frank Hutter. 2019. [Decoupled weight decay regularization](#). In *International Conference on Learning Representations*.
- Jiannan Mao, Chenchen Ding, Tadahiro Matsumoto, Hideki Tanaka, and Masao Utiyama. 2026. [Conditional variational autoencoder for gloss-free sign language translation](#). In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 17317–17321.
- Google MediaPipe. MediaPipe Holistic — Simultaneous Face, Hand and Pose Prediction, on Device — research.google. <https://research.google/blog/mediapipe-holistic-simultaneous-face-hand-and-pose-prediction-on-device/>. [Accessed 03-07-2026].
- Donna Jo Napoli and Rachel Sutton-Spence. 2014. [Order of the major constituents in sign languages: implications for all language](#). *Frontiers in Psychology*, 5.

- Yiming Ni, Zhi-Qi Cheng, Jiayu Li, and Wei Cheng. 2026. [Sign-language datasets at scale: A comprehensive survey on resources, benchmarks, and annotation standards](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 41578–41604, San Diego, California, United States. Association for Computational Linguistics.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [Bleu: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.
- Matt Post. 2018. [A call for clarity in reporting BLEU scores](#). In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. [Learning transferable visual models from natural language supervision](#). *Preprint*, arXiv:2103.00020.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-bert: Sentence embeddings using siamese bert-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
- Nils Reimers and Iryna Gurevych. 2020. [Making monolingual sentence embeddings multilingual using knowledge distillation](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
- Schweter S, Philip May, and Philipp Schmid. 2022. [Germant5/t5-efficient-gc4-all-german-small-el32](#).
- Bowen Shi, Diane Brentari, Gregory Shakhnarovich, and Karen Livescu. 2022. [Open-domain sign language translation learned from online video](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6365–6379, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- S. Sinha. 2018. *Indian Sign Language: A Linguistic Analysis of Its Grammar*. Gallaudet University Press.
- Sugandhi, Parteek Kumar, and Sanmeet Kaur. 2020. [Sign language generation system based on indian sign language grammar](#). *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 19(4).
- Garrett Tanzer, Maximus Shengelia, Ken Harrenstien, and David Uthus. 2024. [Reconsidering sentence-level sign language translation](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6262–6287, Miami, Florida, USA. Association for Computational Linguistics.
- Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. 2022. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems*.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.
- WHO. 2026. [Who fact sheet | deafness and hearing loss](#).
- Ryan Wong, Necati Cihan Camgoz, and Richard Bowden. 2024. [Sign2GPT: Leveraging large language models for gloss-free sign language translation](#). In *The Twelfth International Conference on Learning Representations*.
- Sijie Yan, Yuanjun Xiong, and Dahua Lin. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press.
- Jinhui Ye, Xing Wang, Wenxiang Jiao, Junwei Liang, and Hui Xiong. 2024. [Improving gloss-free sign language translation by reducing representation density](#). In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- Kayo Yin and Jesse Read. 2020. [Better sign language translation with STMC-transformer](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5975–5989, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Benjia Zhou, Zhigang Chen, Albert Clapés, Jun Wan, Yanyan Liang, Sergio Escalera, Zhen Lei, and Du Zhang. 2023. Gloss-free sign language translation: Improving from visual-language pretraining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20871–20881.
- Hao Zhou, Wengang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li. 2021. [Improving Sign Language Translation with Monolingual Data by Sign Back-Translation](#). In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1316–1325, Los Alamitos, CA, USA. IEEE Computer Society.

A Full ISLiM Metric Matrix on iSign

Metric	Val	Test
<i>Corpus BLEU (sacrebleu, tokenize='13a')</i>		
BLEU-1	22.53	22.33
BLEU-4	6.19	5.84
<i>Sentence BLEU (sacrebleu, effective_order=True)</i>		
BLEU-1 (sent avg)	18.95	18.86
BLEU-4 (sent avg)	7.71	7.76
<i>PoseStitch BLEU (add-1 smoothing, str.split())</i>		
PS-B1	22.54	22.33
PS-B4	6.19	5.85
<i>Other metrics</i>		
ROUGE-L	20.28	20.12
METEOR	16.91	16.72
chrF++	20.44	20.48
Semantic similarity	33.69	32.92
Sent zero-BLEU rate	20.3%	21.3%

Table 2: Comprehensive ISLiM metrics on iSign (val 6,302 / test 6,304).

B Qualitative Examples

Table 3 presents representative validation-set translations illustrating recurring success and failure modes.

C Input Channel Sensitivity

We evaluate ISLiM on the iSign test set with keypoint channels zeroed at inference time on a model trained with the full input, shown in Table 4. Removing z-depth causes a modest drop, while zeroing visibility or face landmarks collapses the model.

D Mobile Deployment Feasibility Study

Model weights are exported from the training checkpoint (best_ps.pt) using PyTorch’s mobile optimization module and converted to TorchScript format via the `torch.jit.trace` API. The traced modules were then serialized using the `optimize_for_mobile` utility, which applies operator fusion, constant folding, and dead-code elimination for efficient on-device execution. Six modules were traced:

The test device is a Google Pixel 6a (Google Tensor, 6 GB RAM, Android 16). Models are loaded via `LiteModuleLoader` and inference is performed on CPU with greedy decoding (argmax per step, max 64 tokens).

Inference latency and quality. Table 6 summarises the on-device evaluation over 57 keypoint-based test samples: mean latency is 398 ms (min 208, max 789). The mobile greedy decode slightly exceeds the desktop beam-search PS-BLEU-4 (6.77 vs. 5.85), which is within sampling variance due to the restricted subset size. Latency covers neural-network inference only (pose encoder and greedy T5 decoding); camera acquisition and on-device keypoint extraction are excluded.

Memory: All decoder, embedding, and lm-head modules are resident simultaneously during decoding (on-device model size in Table 6). Under `android:largeHeap="true"`, the app completes the subset evaluation and does not crash until after the evaluation completed.

E SacreBLEU Signatures

All corpus and sentence-level scores are computed with sacreBLEU v2.4.2. The canonical signatures for each metric configuration are listed below for reproducibility.

- **Corpus BLEU** (BLEU corpus_score, tokenize='13a', exp smoothing):
`nrefs:1|case:lc|eff:no|tok:13a|smooth:exp|version:2.4.2`
- **Sentence BLEU** (BLEU sentence_score with effective_order=True):
`nrefs:1|case:lc|eff:yes|tok:13a|smooth:exp|version:2.4.2`
- **chrF++** (CHRF corpus_score, word_order=2):
`nrefs:1|case:lc|eff:yes|nc:6|nw:2|space:no|version:2.4.2`

The PoseStitch BLEU (PS-BLEU) score is computed with a separate implementation (`compute_bleu()` with add-1 smoothing and `str.split()` tokenization) and does not have a sacreBLEU signature.

F Multi-Loss Training Dynamics

Figure 4 plots validation BLEU-4 across training for the SpaMo-OF 10K subset, comparing CE-only training against the full 4-loss configuration. CE-only training either collapses to near-zero (BLEU-4 < 0.3 after 24 epochs) or plateaus at ~3 even after 325 epochs. The 4-loss configuration climbs steadily past 5, reaching a maximum validation BLEU-4 of 5.44 at epoch 500 before mild overfitting. Validation BLEU-4 uses sentence-averaged

Case	Reference	ISLiM output
(a) Clean	there were 6 8 passengers and 4 crew members onboard in all there were 7 2 people	there were 6 6 passengers and 4 crew members onboard the flight
(b) Clean (almost)	sushant singh rajput s father krishna kishore singh has filed a case in the delhi high court	sushant singh rajput s parents filed a case in the delhi high court
(c) Paraphrase	they said the plane was at a very high speed which is why it went off the runway	this is because of the aircraft s failure on the runway
(d) Collapse	their internet is very slow	the person person
(e) Halluc.	his family has said once the covid outbreak ends	covid 1 9 has spread all over the world

Table 3: Qualitative validation-set examples. (a–b) Clean predictions preserving numerical and entity detail. (c) Partial semantic match with low BLEU. (d) Collapse on a short reference with hapax vocabulary. (e) Topical hallucination that semantic-similarity metrics over-reward.

Condition	BLEU-4	Δ	Zero%
Full (94×4D)	5.84	—	21.3
– z-depth	5.33	–0.51	22.9
– visibility	2.21	–3.63	33.1
– face (46 kp)	0.37	–5.47	37.7
Hand-only (42 kp)	0.21	–5.63	36.0

Table 4: Inference-time input-channel sensitivity (6,304 test samples). Zeroing visibility or face landmarks collapses the model. As channels are removed from a fully trained model, these results indicate sensitivity to feature removal under distribution shift rather than the isolated contribution of each channel; retraining-based ablations are left to future work

Module	Size
encoder_traced.ptl	56 MB
t5_encoder.ptl	73 MB
t5_decoder_first_step.ptl	159 MB
t5_decoder_step.ptl	147 MB
t5_shared_embed.ptl	63 MB
t5_lm_head.ptl	63 MB
Total	561 MB

Table 5: Sizes of the exported TorchScript artifacts (.ptl files) for mobile deployment; the on-device model footprint is reported in Table 6.

scoring; the corresponding corpus-level test set BLEU-4 at the selected checkpoint is 4.12 (Table 1).

G Parameter Efficiency on Phoenix-2014-T

Figure 5 shows the parameter efficiency of ISLiM, SpaMo, and SpaMo-OF on the Phoenix-14-T dataset.

Quantity	Value
Evaluation samples	57
Model size (on-device)	495.7 MB
Peak RAM (heap)	38.8 MB
CPU cores	8
Latency (mean)	398 ms
Latency range (min–max)	208–789 ms
PS-BLEU-1/2/3	24.3 / 13.81 / 9.19
PS-BLEU-4	6.77

Table 6: On-device evaluation of the keypoint-to-English pipeline (57 iSign test samples, Google Pixel 6a, CPU-only greedy decoding).

H Qualitative Analysis of Multi-Loss Setup

Table 7 summarizes, qualitatively, how each auxiliary loss contributed in our SpaMo-OF 10K training.

Loss	Prevents Collapse	Improves BLEU	Improves Fluency
CE only	No	Baseline	Good (when stable)
CE + CLIP	Yes	Strong	Moderate
CE + Semantic	Partial	Moderate	Strong
CE + Entropy	Yes	Slight	No

Table 7: Qualitative contribution of each auxiliary loss. CLIP contrastive loss is the primary driver of BLEU improvement; semantic similarity and entropy regularization act as stabilizers.

I Effect of Sentence Length on Translation Quality

We examine whether translation quality degrades with increasing sentence length. Table 8 reports mean and median BLEU-4 and ROUGE-L on the full iSign test set (6,304 sentences), binned by reference word count.

Sentence length has a negligible effect on translation quality. Mean BLEU-4 remains stable at

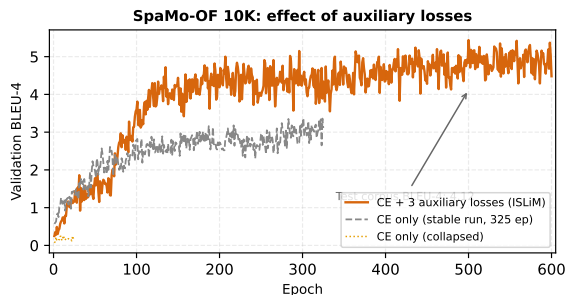


Figure 4: Validation BLEU-4 on iSign 10K. CE-only runs either collapse (orange dotted, stuck <0.3) or plateau at ~ 3 (grey dashed, 325 ep). The 4-loss configuration (orange solid, 600 ep) reaches 5.44, demonstrating that auxiliary losses, not architecture, are the source of the gain.

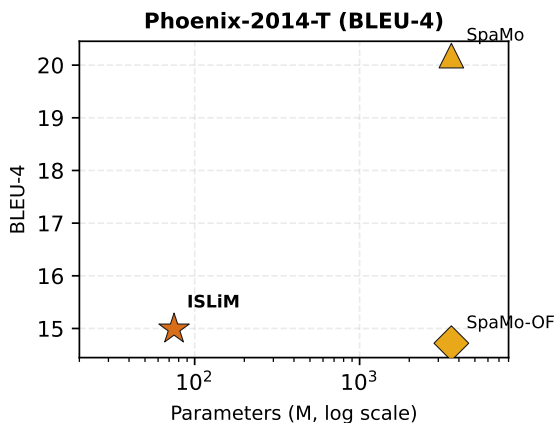


Figure 5: Parameter efficiency on Phoenix-2014-T. ISLiM (75M, red star) matches the 3.6B SpaMo-OF at $48\times$ fewer parameters (BLEU-4 14.99 vs. 14.72). Keypoint-based ISLiM appears as a star; video-based methods as triangles/diamonds.

approximately 7 across the 5-25 word range that covers 93% of the test set, dropping only to 5.6 for the small number of sentences exceeding 25 words ($n = 46$). Mean ROUGE-L increases slightly with length (16.4 to 22.5), indicating that surface-level word overlap remains consistent even when higher-order n -gram matches become more challenging.

The only notable pattern is in very short sentences (1-4 words): this bucket has the highest mean BLEU (13.84) but a median of zero, meaning more than half of these sentences score no n -gram matches at all while the remainder score high. This bimodality reflects the reduced redundancy of short targets, suggesting that there is no margin for partial credit, so the model either matches the reference exactly or receives no n -gram overlap.

Predicted sentence length slightly lags the ref-

Words	#Sent.	Mean B4	Median B4	Mean RL	Pred. Len
1-4	662	13.84	0.00	18.28	5.2
5-7	1021	7.21	4.20	16.42	7.8
8-10	1323	7.26	4.46	19.32	9.9
11-13	1211	7.31	4.07	20.63	11.7
14-16	974	6.99	3.65	21.02	13.0
17-20	734	6.76	3.60	21.20	14.6
21-25	324	6.02	3.89	21.00	16.0
26-30	46	5.61	3.81	22.46	16.6

Table 8: BLEU-4 and ROUGE-L by reference sentence length on the full iSign test set. Mean BLEU-4 is stable at approximately 7 for sentences of 5-25 words, with only a mild drop for longer sentences. Short sentences (1-4 words) show high variance: mean 13.84 but median 0.00, indicating a bimodal distribution where the model either translates them correctly or fails entirely.

erence length as sentences grow longer (e.g., 27.2 vs. 16.6 words for the 26-30 word bucket), a common length bias in neural machine translation. However, sacreBLEU’s brevity penalty largely compensates for this effect, as evidenced by the flat BLEU-4 trend. Overall, the model’s translation quality is independent of sentence length for the vast majority of test examples.