Towards Scalable Sign Production: Leveraging Co-Articulated Gloss Dictionary for Fluid Sign Synthesis

Aparna Agrawal
International Institute of information technology Hyderabad
Hyderabad, Telangana, India
aparna.agrawal@research.iiit.ac.in

Seshadri Mazumder
International Institute of information technology Hyderabad
Hyderabad, Telangana, India
seshadri.mazumder@research.iiit.ac.in

C.V.Jawahar
International Institute of Information Technology Hyderabad
Hyderabad, Telangana, India
jawahar@iiit.ac.in

Vinay Namboodri
University of Bath
Claverton Down, Bath, United Kingdom
vpn22@bath.ac.uk

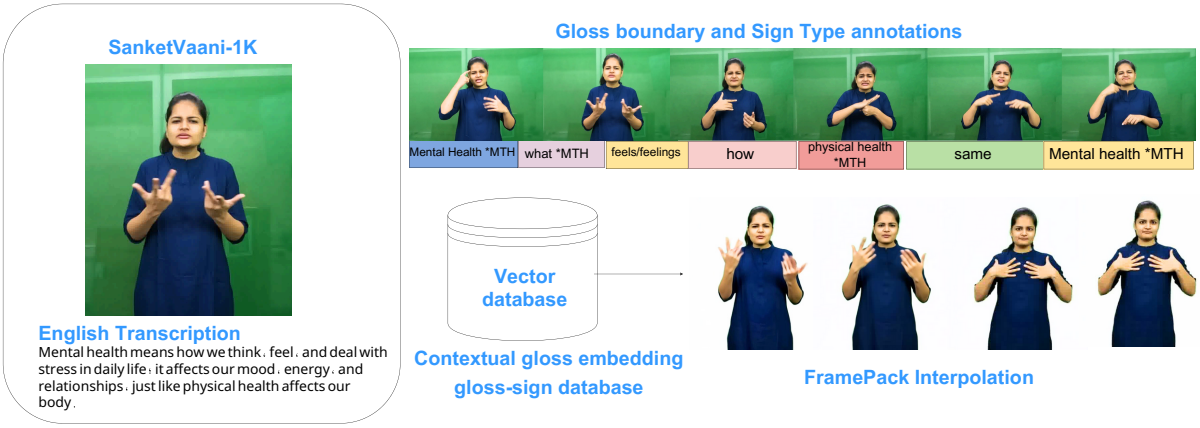

Figure 1: For Indian Sign Language Production, we exploit fine-grained gloss-boundary and sign-type annotations with SOTA interpolation method; we show it enables smooth, context-aware visual synthesis for sign-language generation, and to support this, we collected a carefully curated 1,000-sentence interpreted ISL corpus - SanketVaani-1K.

Abstract

Sign Language Production (SLP) systems can significantly improve accessibility for Deaf communities by translating spoken or written language into sign language videos. For millions of Indian Sign Language (ISL) users, such systems could bridge persistent communication gaps in education, healthcare, and public services. However, SLP in ISL is hindered by the scarcity of annotated datasets and the expressive complexity of the language. Existing ISL datasets, such as iSign [24] and ISL-CSLRT [37], are designed for recognition and lack gloss-level[1] annotations, forcing full-sentence modeling that struggles to generalize to new constructions.

We propose an interpolation-based approach that stitches together individual gloss-level signs using modern text-to-motion frameworks like FramePack [56]. This produces contextually accurate and grammatically consistent sign sequences while retaining fine-grained articulation. To support this pipeline, we curate a 1,000-sentence ISL dataset, translated by a native interpreter and annotated with precise gloss boundaries. Results show that with high-quality gloss-level supervision, interpolation-based synthesis offers a practical, scalable path for inclusive SLP in low-resource sign languages like ISL. The project page can be found here https://cvit.iiit.ac.in/research/projects/cvit-projects/towards_ssp.

[1]Gloss is word representation of sign.

CCS Concepts

• **Computer vision**; • **NLP**; • **Sign Language Production**;

Keywords

Sign Language Generation, Interpolation, Text to Motion, Dataset

ACM Reference Format:
Aparna Agrawal, Seshadri Mazumder, C.V.Jawahar, and Vinay Namboodri. 2025. Towards Scalable Sign Production: Leveraging Co-Articulated Gloss

ICVGIP 2025, Mandi, India

ACM ISBN 979-8-4007-1930-1/2025/12
https://doi.org/10.1145/3774521.3774576

Dictionary for Fluid Sign Synthesis . In *Indian Conference on Computer Vision, Graphics, and Image Processing (ICVGIP 2025), December 17–20, 2025, Mandi, India.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3774521.3774576

1 Introduction

According to the World Health Organization (WHO), approximately 466 million people worldwide are deaf or hard of hearing, with around 72 million using sign language as their primary means of communication [33]. In India, a WHO news report estimates that 63 million people have "significant auditory impairment" [43].[2] Sign language is a visually rich and complex mode of communication, incorporating spatial grammar, facial expressions, and contextual cues that are often unfamiliar to non-signers, making it challenging for them to learn and use effectively.

Sign language communication involves two key aspects: being understood and understanding others. Although significant progress has been made in sign language recognition, the ability to generate accurate and fluent sign language, known as Sign Language Production (SLP), remains a major challenge. The situation is exacerbated by the shortage of certified professionals; the Indian Sign Language (ISL) Research and Training Centre (ISLRTC)[3] currently lists only 304 Level-C (DISLI) certified ISL interpreters across India, divided across six regions. This stark gap underscores the urgent need for scalable SLP systems that can provide an accessible and affordable communication bridge for the deaf and hard-of-hearing community.

Although several approaches to Sign Language Production (SLP) have been proposed, most rely on avatar-based 3D animation [6, 7, 22, 36] or generative models like GANs [40, 48]. In contrast, we propose an interpolation-based approach that leverages modern text-to-motion models to interpolate between individual gloss-level signs. This method preserves the fine-grained detail and expressiveness of original signing, enabling the synthesis of novel sentences. Using state-of-the-art techniques such as FramePack, we show that realistic and coherent sign-language sequences can be generated through gloss-wise interpolation.

A major challenge in SLP is the lack of finely annotated datasets suitable for synthesis. Existing datasets like RWTH-PHOENIX[9], How2Sign [13], and iSign[24] are designed for recognition and lack accurate gloss-level annotations. Training end-to-end SLP systems requires large continuous corpora to capture the full range of sign-gloss combinations and co-articulations, which are scarce and expensive to annotate. To alleviate this, we adopt a modular approach and introduce a compact dataset of 1,000 sentences with gloss-level annotations, enabling compositional synthesis through interpolation to generate novel sentences. The contributions are:

- We propose a novel interpolation-based method for sign language production using gloss-level alignment and text-to-motion models.
- We introduce a compact dataset of 1000 ISL sentences with precise gloss boundaries tailored for generation.
- We conducted both quantitative evaluations and user studies with Deaf participants, showing improved realism and understandability.

[2]Estimates of India's deaf and hard-of-hearing population vary widely. The 2011 Census reports 5 million people with hearing disabilities, while WHO (2024) and Varshney (2016) estimate 63 million with "significant auditory loss." NAD (2016) reports 18 million deaf individuals. [43].

[3]https://islrtc.nic.in

2 Related works

Sign Language Production (SLP) sits at the intersection of computer vision, natural language processing, and motion synthesis. Early research primarily focused on Sign Language Recognition (SLR) [34, 58–60, 63] and Sign Language Translation (SLT) [5, 20, 27, 61]. However, with the growing need for effective communication between deaf and hearing individuals, and the shortage of interpreters, Sign Language Production (SLP) has gained significant attention in recent years.

Avatar-based systems render sign language using animated 3D characters, driven by notation-based scripting (e.g., HamNoSys/SiGML) or motion capture. They offer fine-grained control over grammatical and non-manual features, enabling linguistically accurate generation. However, these systems often fall into the "uncanny valley," with rigid or unnatural motion that reduces comprehensibility.[6, 7, 22, 36]. More recent methods like Spoken2Sign [64] leverage SMPL-X-based avatars improve motion continuity, but remain heavily reliant on tracking accuracy and still fall short of producing visually realistic, fluent sign language videos.

The advent of deep learning has enabled more data-driven methods for mapping spoken language to sign language. One of the earliest SLP pipelines decomposed the task into three stages: Text-to-Gloss (T2G) translation, Gloss-to-Pose (G2P) retrieval, and Pose-to-Sign (P2S) video synthesis [32, 48]. While effective in modularizing the process, these early systems struggled to produce smooth transitions between signs and often generated low-resolution outputs that failed to capture critical non-manual cues such as facial expressions and eye gaze.

Generative Adversarial Networks (GANs) have been applied to sign language production [40, 48] to generate realistic signing videos from pose or gloss input. While models like SignGAN [40] and FSNet [41] improve motion detail through adversarial training and temporal smoothing, they remain limited by dependence on large paired datasets, temporal inconsistencies between signs, and visual artifacts that hinder fluency and realism.

In terms of resources, several datasets have propelled SLP and SLR research. RWTH-PHOENIX-Weather 2014T [9], BSL-1K [2], and How2Sign[13] provide large-scale video-gloss pairs, with How2Sign offering multiview and RGB-D data for ASL. However, these are primarily designed for recognition tasks and lack gloss-type annotations (*P for pointing, *FS for fingerspelling, etc.) or motion-segment boundaries needed for compositional synthesis. In the Indian context, iSign [24] provides sentence-level ISL glosses, but lacks consistent signer identity, temporal alignment, or metadata for generative modeling. ISLTranslate [23] focuses on bilingual translation and omits gloss boundary annotations. CSLR2 [38] introduces some temporal annotations for signs, but is still aimed at classification and not production.

Despite this progress, current resources and methods do not adequately support semantically controlled, modular sign language generation. There remains a pressing need for datasets with rich

gloss-level annotations, temporal alignment, and motion metadata to enable scalable SLP grounded in linguistic context.

3 Proposed Synthesis Pipeline

Sign Languages (SLs) are natural visual-gestural languages used by Deaf communities worldwide, each with its own linguistic structure. ISL is the primary SL used in India, distinct from ASL or BSL in grammar, vocabulary, and regional variation. ISL is relatively under-resourced, with limited standardized datasets and formalization, making direct sign synthesis particularly challenging. These challenges motivate retrieval-based methods that operate over gloss-level annotations and leverage modular synthesis strategies.

We first build a gloss-level vector database storing gloss embeddings for fast retrieval. At runtime, an LLM translates the English sentence into ISL gloss tokens, which query the database for matching context embeddings, and a motion-interpolation module blends the retrieved segments into a fluent video Fig. 2.

3.1 Vector Database Construction

To support semantically grounded sign language generation, we construct a vector database of gloss-level embeddings that capture both the lexical identity and contextual meaning of each sign. We begin by encoding gloss sequences using a pretrained contextual language model, such as `BERT`[1], which is well-suited for producing high-quality token-level embeddings.

Given a sentence-level gloss sequence,

$$G = [g_1, g_2, \ldots, g_n],$$

we compute a contextualized embedding for each gloss token g_i as:

$$\mathbf{e}_i = f(g_i \mid g_{i-k}, \ldots, g_i, \ldots, g_{i+k}),$$

where $f(\cdot)$ is the encoder function and k is a context window size. These embeddings encode the gloss meaning in relation to its neighboring glosses, capturing critical information for sign disambiguation [4] and production.

Each embedding $\mathbf{e}_i \in \mathbb{R}^d$ is then stored in a vector database (e.g., FAISS [11]), indexed along with the following metadata:

- Gloss label and sign type (e.g., `*P` for pointing, `*MTH` for mouthing),
- Gloss position within the sentence,
- Surrounding gloss context (for local coherence),
- Motion segment pointer linking to the corresponding sign animation.

During inference, for a given query gloss g_q, we compute its contextual embedding $\mathbf{e}_q$ and perform a nearest-neighbor search over the database:

$$\hat{s} = \arg\min_{s_j \in \mathcal{D}} \ \text{dist}(\mathbf{e}_q, \mathbf{e}_j),$$

where $\text{dist}(\cdot,\cdot)$ denotes cosine, and $\mathcal{D}$ is the set of stored embeddings with linked motion segments. Gloss position, Sign Type can be used for ranking. This allows us to retrieve motion units that are both semantically relevant and contextually appropriate for subsequent stitching.

[4]process of selecting the correct sign for a gloss based on its intended meaning in context.

By capturing gloss meanings at a fine-grained, contextual level and linking them to expressive motion data, this database serves as the semantic backbone for our retrieval-based sign language production system.

3.2 Text to Gloss Sequence

Early English-to-ISL conversion systems primarily relied on handcrafted transfer rules to reorder source text into ISL gloss sequences. While the limited and partially documented nature of ISL grammar exacerbates the challenge, the core limitation stems from the inherent brittleness of rule-based architectures: fixed if–else style rules cannot adequately capture the full variability of natural language. As a result, novel syntactic constructions require manual, ad hoc rule additions, and idiomatic or polysemous expressions are often mistranslated or omitted. When applied to unconstrained, real-world text, these systems often yield gloss output that is syntactically inconsistent and semantically incomplete.

Recent studies demonstrate that LLMs provide a context-aware alternative, consistently outperforming rigid rule-based systems. [18] generate *pseudo-gloss* labels from spoken text via few-shot prompting, eliminating costly manual annotation and improving downstream sign-language translation on PHOENIX14-T and CSL-Daily. Further work has shown that zero-shot prompting in dialogue systems [19], gloss+non-manual cue generation [55], and stepwise gloss induction [4] can all leverage LLMs to capture sign language grammar better than traditional rule sets.

LLM-based ISL Gloss Generation. We supply the LLM with a concise *ISL rule card* that summarises the core grammatical constraints reported in [17] (e.g., *TOPIC–COMMENT* word order, article omission, sentence-final WH-words, clause-initial time adverbials). Let the input sentence be $t = (w_1, w_2, \ldots, w_N)$. Conditioned on this rule card (plus one worked example), the LLM emits a gloss sequence $\boldsymbol{g}' = \{g'_j\}_{j=1}^{M}$, where M is determined dynamically by the semantic segmentation of t. Each token g'_j respects the rule-card constraints, ensuring that $\boldsymbol{g}'$ preserves the meaning of t while conforming to the syntactic and morphological conventions of ISL. Full prompt templates and representative gloss outputs are provided in the supplementary material.

3.3 Stitching

Stitching is the final and crucial step in our Sign Language Production pipeline, where retrieved motion segments for individual glosses are seamlessly combined into a coherent and natural signing sequence. Naively, stitching together individual motion segments often results in temporal discontinuities, abrupt transitions, or jerky artifacts at the boundaries between segments.

Traditional approaches for generating intermediate motion frames between two anchor poses often rely on pose-space interpolation (e.g., linearly interpolating joint angles or latent vectors in a model like MediaPipe[29] or SMPL[28]). While simple, these methods fail to capture the non-linear dynamics of natural human motion, often producing mechanical or physically implausible transitions, especially around semantically rich actions like signing, gestures, or expressive facial cues. Rendering these interpolated poses with

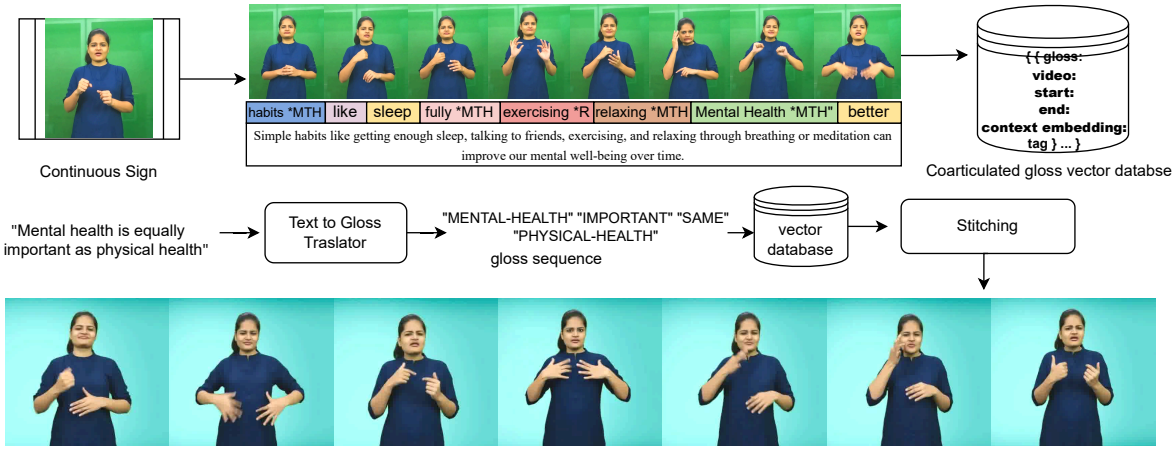

Figure 2: Stitching Pipeline: includes Text to gloss Translator, formation of Vector database, and video retrieval, and sign stitching with interpolation.

GANs or rigged avatars, FOMM [46], often produces warping artifacts or temporal inconsistency.

In contrast, unidirectional next-frame generation methods, e.g., causal video diffusion or autoregressive models, suffer from temporal drift, attention bottlenecks that limit long-range coherence, and quadratic memory/compute costs for high-resolution sequences.

To overcome this, we use FramePack, originally proposed for long-video next-frame prediction in diffusion transformers, which compresses earlier and later frames geometrically so the total context length stays fixed, letting a transformer attend to far-apart anchors without the quadratic blow-up. More importantly, its bidirectional 'antidrifting' sampling first locks in high-fidelity start and end frames, then generates the intermediate span from both directions, eliminating the drift and anchor forgetting.

We apply FramePack interpolation between glosses. For each boundary we supply only the last frame of segment i and the first frame of segment i + 1 as anchors; FramePack keeps these two frames at full resolution, treats them as bi-directional endpoints, and synthesises the missing in-betweens from both directions. FramePack runs atop a frozen Hunyuan-DIT [26] a 13 B-parameter diffusion backbone, letting us inherit strong text-to-motion alignment while achieving drift-free, low-overhead interpolation. The result is a seamless signing stream.

4 Dataset: "SanketVaani-1K"

We present a richly categorized ISL dataset comprising 1,000 spoken sentences, meticulously crafted to reflect the diverse and essential communication needs of Deaf individuals in India. The dataset spans over 40 high-utility domains, including Financial Literacy, Banking, Healthcare, Legal Rights, Education, and Government Services & Policies, among others, capturing a wide range of everyday contexts that Deaf users regularly navigate. Figure 3 presents a detailed breakdown of these categories, highlighting the percentage distribution of sentences across each domain. We also analyzed the sentence structure, with Figure 4 showing the overall word count distribution; most sentences range between 10 and 30 words.

Dataset	Lang.	# Sent.	Vocab. (Text)	Domain	Gloss Annot.	Segmented	Controlled Env.
Purdue RVL [30]	ASL	2.5k	104	General	✓	✗	✓
Boston 104 [12]	ASL	201	103	General	✓	✗	✓
How2Sign [13]	ASL	35k	16k	Instructional	✓	✗	✓
OpenASL [45]	ASL	98k	33k	Web	✗	✗	✗
YT-ASL [47]	ASL	610k	60k	Web	✗	✗	✗
BOBSL [3]	BSL	993k	72k	TV	✗	✗	✗
BOBSL-CSLR [38]	BSL	5k	5k	TV	✓	✓	✗
CSL Daily [62]	CSL	20.6k	2k	General	✗	✗	✓
Phoenix14T [9]	DGS	8.2k	3k	Weather	✓	✗	✗
SWISSTXT-W [21]	DSGS	811	1k	Weather	✓	✗	✗
SWISSTXT-N [10]	DSGS	6k	10k	News	✓	✗	✗
KETI [25]	KSL	14.6k	419	Emergency	✓	✗	✓
VRT-News [10]	VGT	7.1k	7k	News	✗	✗	✗
ISL-CSLRT [37]	ISL	100	–	General	✗	✗	✓
ISLTranslate [23]	ISL	31k	11k	Education	✗	✗	✗
iSign [24]	ISL	118k	40k	General	✓	✗	✗
SANKETVAANI-1K (Ours)	ISL	**1k**	**6k**	~40 cat.	✓	✓	✓

Table 1: Comparison of sign language datasets. The *Controlled Env.* column indicates whether the dataset was recorded in a controlled environment (✓) or curated from existing sources (✗).

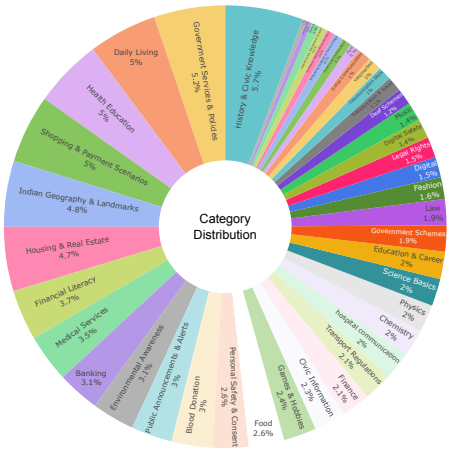

Figure 3: Comprehensive category-wise distribution of the ISL dataset.

To ensure meaningful coverage, we carefully selected each domain based on its practical relevance to the Indian socio-cultural context. Within each domain, we designed subdomains to capture

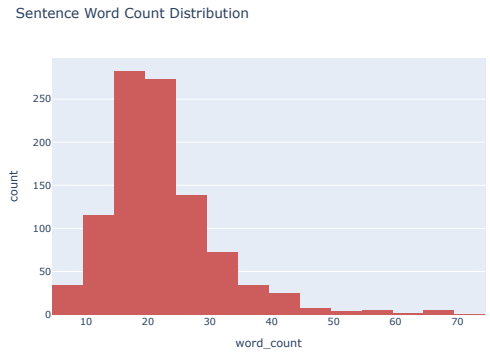

Figure 4: Distribution of sentence word counts in the ISL dataset, showing most sentences fall between 10–30 words.

finer real-world variations. For example, the Healthcare domain includes topics such as vaccination schedules, medication usage instructions, and emergency protocols. We used ChatGPT to generate contextually rich, culturally grounded, and linguistically accessible sentence prompts. Emphasis was placed on clarity, relevance, and simplicity to ensure suitability for visual translation. A certified ISL interpreter translated each spoken sentence into ISL, preserving grammatical structure, facial expressions, and non-manual cues essential for natural communication.

We recorded all clips with a single front-facing GoPro10 camera (1920×1080 at 30 fps), mounted on a tripod and centered on the signer. The fixed frontal view consistently framed the head, torso, and hands, providing clear capture of both manual and non-manual cues while keeping recordings comparable across sessions.

4.1 Data Pre-processing

4.1.1 Filtering and Syncing. To enable precise sign-token segmentation, we insert an unambiguous acoustic cue (clap) at the start and end of each sentence.

GoPro audio is down-mixed to mono PCM (44.1 kHz), normalized, and transformed to a Hilbert envelope. Peaks above 0.7 a.u. and 0.5 s apart define clap events, which segment the timeline into H.264 clips (sub00.mp4, …). Next, we run a single manual pass to filter noisy and non-important clips. All retained clips are renamed to match their annotated sentence IDs (e.g., S00001, S00002, …).

4.1.2 Signer normalization. Each frame is run through YOLO v9 [31] to detect the signer, the detection is linked across frames with ByteTrack, and the resulting box is expanded by 10 % to avoid trimming hands or face. That crop is then pasted in the center of a full-HD canvas filled with a neutral green-grey background, so every output frame shows the signer in the same position and scale.

4.1.3 Body masking & background replacement. For every cropped clip, the pipeline first extracts a foreground mask from the very first frame using isnet [35] and saves binary mask. The video is then processed in 60-frame chunks: each chunk is fed through a matting step that yields clean foreground RGB frames plus per-pixel alpha maps (MatAnyone [51]). Those alpha maps are used to blend the signer onto a chosen background image, frame by frame, preserving the original frame rate and resolution.

4.2 Annotations

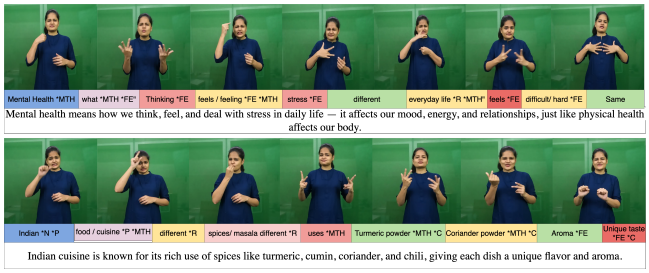

Figure 5: Example sentence from the dataset and corresponding annotation with marked boundaries for each gloss.

The annotation procedure uses a web-based interface built on the VIA video annotation tool [14], as also adopted in prior CSLR annotation work [38]. With the help of the same ISL interpreter who translated the original sentences, we performed detailed gloss-level annotations. For each sign token, the interpreter entered free-form glosses while observing the full signing context. The original spoken sentence, generated via ChatGPT, was shown alongside the video to guide semantic alignment, and glosses were chosen to reflect both semantic accuracy and contextual relevance. If a spoken word was expressed using multiple signs, both the combined and individual glosses were annotated. The annotator additionally marked the sign types when appropriate, for example, *MTH for mouthing, *P for pointing, and *NS for no sign, inspired by the work [38] with some additional types. See the Supplementary for more details. Each gloss was temporally aligned with the corresponding signing segment using the annotation tool, ensuring an accurate alignment suitable for downstream generation and synthesis.

While our primary focus is on sign language generation, the dataset's structure and annotation granularity also make it well-suited for other tasks such as sign language translation (SLT), gloss spotting, sign retrieval. Figure 5 illustrates the annotation process, including gloss input, sign-type tagging, and temporal alignment.

5 Experiments

5.1 Data preprocessing

Figure 6: Qualitative results of our preprocessing pipeline. Each column shows key stages, including initial frame mask, initial segmented frame, signer detection and normalization, alpha matting, and final composite with background replacement.

We qualitatively evaluate our preprocessing pipeline by examining visual consistency, signer isolation, and background replacement across multiple sign language video clips. The pipeline ensures temporal alignment, consistent framing, and clean signer isolation across all clips. Figure 6 presents exemplar frames from the preprocessed video, highlighting YOLO v9 + ByteTrack-based cropping and centering procedure successfully standardizes signer positioning and scale across clips, leading to uniform framing regardless of signer movement. The matting and blending pipeline with MatAnyone effectively isolates the signer and integrates them onto a neutral background, preserving edge details such as hands and hair.

5.2 Text to Gloss Sequence

BLEU ↑				ROUGE-L ↑	METEOR ↑	CHrF ↑
1	2	3	4			
0.350	0.182	0.114	0.086	0.393	0.294	0.473

BERTScore	F1	Recall	Precision
	0.784	0.789	0.781

Table 2: Automatic evaluation of generated gloss sequences on the RWTH-PHOENIX-Weather 2014T [16] test set using BLEU, ROUGE-L, METEOR, CHrF, and BERTScore. Metrics compare the ground truth glosses with generated (ChatGPT).

Gloss Seq Variant	TSOV Accuracy ↑	BERTScore ↑		
		F1	Recall	Precision
RWTH-PHOENIX-Weather 2014T(Test set)	0.72	0.723	0.687	0.704
ChatGPT 4o	0.76	0.791	0.740	0.764

Table 3: Evaluation of German Sign Language gloss sequence generation using TSOV (Time-Subject-Object-Verb) word order accuracy and BERTScore. BERTScore is computed between the original sentence and the generated gloss sequence.

We convert Text into a sequence of glosses following the respective linguistic rules as discussed in section (3.2) for ISL. Since there is no parallel text dataset for ISL, we evaluated our method on RWTH-PHOENIX-Weather 2014T [16], a German Sign Language (GSL) corpus. As shown in Table 2, our generated glosses achieve competitive scores in BLEU, ROUGE-L, METEOR, CHrF, and BERTScore metrics. Notably, while PHOENIX glosses are widely used for evaluation, they were designed for recognition tasks and often diverge from canonical GSL structure, reflected in their lower TSOV accuracy (0.72) compared to ChatGPT-generated glosses (0.76) in Table 3. The BERTScore F1 is also higher (0.791 vs. 0.723), indicating better semantic alignment. German compound words (e.g., Sturmwarnung) further challenge lexical matching, emphasizing the need for gloss evaluation benchmarks for generation tasks.

5.3 Vector Database Retrieval

A single English word may map to several distinct signs, depending on context. Our experiment therefore measures the ability of the gloss-level vector database to perform sense disambiguation and return the correct sign for each occurrence of a polysemous word.

For every unseen sentence we (1) convert the sentence to an ISL gloss sequence using the LLM , and (2) query the vector database with the contextual embedding of each gloss. The database returns the top-k candidate signs together with metadata. To handle low-confidence retrievals, we also collect isolated alphabet signs from the signer. If the retrieval match score for a given gloss is below 40%, the system defaults to fingerspelling the word using the recorded alphabet signs.

Query(sense)	Rank	Retrieved sentence
light (lamp)	1	**If you smell gas inside your home, do not switch on electrical devices. Open windows immediately and call the emergency gas service.**
	2	Electricity is the energy that powers lights, fans, and machines. It comes from wires or batteries.
	3	Photosynthesis is how plants make their food. They use sunlight, water, and air to grow.
light (sun)	1	**Photosynthesis is how plants make their food. They use sunlight, water, and air to grow.**
	2	If you smell gas inside your home, do not switch on electrical devices. Open windows immediately and call the emergency gas service.
	3	Electricity is the energy that powers lights, fans, and machines. It comes from wires or batteries.
fair (carnival)	1	**Located in Gujarat, the Rann of Kutch is a salt desert. It hosts the colorful Rann Utsav festival every year.**
	2	Punjab is famous for its fertile plains, Sikh culture, and Baisakhi festival.
	3	The local community center will hold a vaccination drive next Saturday from 9 AM to 3 PM. All residents are encouraged to participate.
fair(just)	1	**Shivaji maharaj was a brave Maratha leader who fought the Mughals and created a strong navy to protect the coast. He is admired for his clever plans and fair rule.**
	2	Beware of fake callers pretending to be bank officials. Never share your PIN or OTP with anyone.
	3	If someone cannot be present, they can give legal rights to another person to sell or manage property using POA.
bank(finance)	1	**A home loan is money borrowed from a bank to buy a house, and it must be repaid monthly with interest over several years. Before approving a loan, banks check your income, job status, credit history, and the property documents.**
	2	Beware of fake callers pretending to be bank officials. Never share your PIN or OTP with anyone.
	3	Beware of fake callers pretending to be bank officials. Never share your PIN or OTP with anyone.

Table 4: Top-3 sentence retrieval for five ambiguous queries with different meanings in the context.

Table 4 shows three illustrative cases. The **Sense** column clarifies the meaning of the ambiguous gloss in its source sentence, whereas the **Retrieved sentence** column displays the sentence that supplied the top-ranked sign. In all examples, the system selects a sign whose usage is semantically consistent with the query context, demonstrating effective disambiguation.

5.4 Interpolation

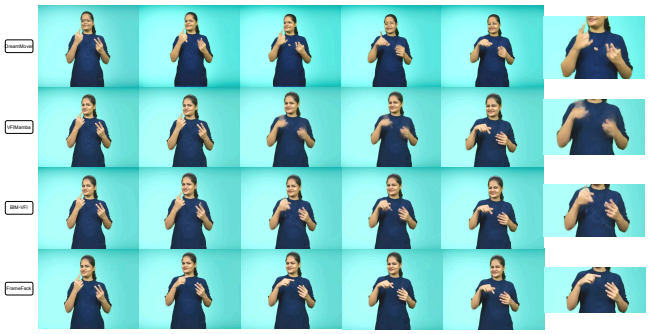

Figure 7: Qualitative comparison of DreamMover, VIFMamba, BIM-VFI and FramePack interpolation Methods.

Interpolating human motion in sign language videos poses challenges not typically found in generic frame interpolation. In natural images or object scenes, small geometric or texture inaccuracies may go unnoticed. In contrast, sign language demands high precision in handshape, palm orientation, and facial expressions: any

distortion can alter or obscure meaning. Movements often involve rapid, non-linear trajectories with complex co-articulation between consecutive signs. Both hands must remain synchronised in timing and spatial relation, and frequent self-occlusion (hands crossing or covering the face) makes optical-flow-based methods prone to warping artefacts.

We evaluate FramePack for video frame interpolation against three recent SOTA baselines DreamMover [44], BiM-VFI [42], and VFIMamba [54] to study trade-offs in perceptual quality, temporal consistency, and computational efficiency across diverse upper-body motions.

For each annotated gloss boundary, we identify the co-articulation midpoint t_m, convert it to a frame index, and extract nine frames centered around it. The two outer frames serve as anchors, and the seven inner frames are treated as ground-truth targets, resulting in 3,963 anchor–target pairs in the test set. All methods generate seven 720X1080 frames and are evaluated using PSNR, SSIM, LPIPS, FID [53], NIQE [57], MS-SSIM [39], GMSD [50] , and KID [8].

Baselines:**DreamMover** is a diffusion-based method that uses feature-level warping and two-stage latent fusion. **BiM-VFI** employs bidirectional motion fields to interpolate arbitrary-t frames with perceptually rich detail.**VFI-Mamba** combines state-space modeling and token interleaving to handle large motions and high-resolution inputs. **FramePack** uses a 13B Hunyuan-DiT[26] , conditioned on both anchor frames. Prior context is geometrically compressed to maintain memory efficiency. Interpolation is performed via bidirectional generation with anchor-based sampling.

Qualitative Results: Figure 7 compares intermediate frames generated between two anchor frames. VFIMamba shows increasing blur from the second interpolated frame onward, especially on hands and face. DreamMover largely preserves global pose but introduces mild facial distortion and hand artefacts in the third and fourth frames. BiM-VFI produces a conspicuous artefact on the subject's left hand in the third frame. In contrast, FramePack maintains facial detail and hand articulation across all frames, yielding the most coherent sequence.

Quantitative Results: FramePack scores lower on distortion-based metrics because those metrics tend to favour blur and strict pixel alignment. PSNR, SSIM, MS-SSIM, GMSD penalise even sub-pixel finger shifts, while blur reduces the measured error. LPIPS, FID, KID, NIQE use feature statistics learned from generic natural images; they often treat the high-frequency detail in hands and faces as noise. By keeping that fine detail, FramePack incurs larger numeric penalties yet looks perceptually sharper and more natural.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	NIQE ↓	MS-SSIM ↑	GMSD ↓	KID ↓
DreamMover	26.02	0.915	0.117	56.82	**46.85**	0.947	0.103	0.065
VFIMamba	27.81	0.950	**0.066**	52.96	70.12	**0.961**	**0.094**	0.079
BiM-VFI	**27.88**	**0.956**	0.103	**39.77**	58.84	0.956	0.097	**0.031**
FramePack	23.96	0.909	0.163	49.80	69.48	0.925	0.132	0.057

Table 5: Quantitative comparison of interpolation methods. Higher is better for PSNR, SSIM, MS-SSIM; lower is better for LPIPS, FID, NIQE, GMSD, and KID.

5.5 Comparison with SLP methods

Figure 8 offers a concise visual audit of our quality-comparison study across four SOTA sign-language-production (SLP) pipelines:

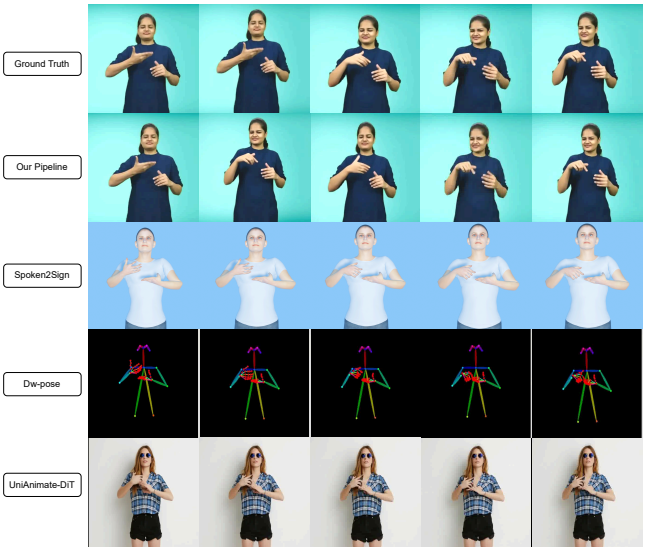

Figure 8: Qualitative comparison of SLP approaches: (1) FramePack – interpolation based; (2) Spoken2Sign – avatar-based synthesis using an SMPL-X body model; (3) UniAnimate-DiT – diffusion-based, 2D keypoint-driven.

Ground-Truth (GT): Real frames from the source signer serve as the upper-bound reference for realism. **FramePack:** Interpolates directly between two real anchor frames, preserving both identity and photorealism. **Spoken2Sign:** Fits an SMPL-X mesh to each anchor frame and linearly interpolates 3D body parameters. Small pose-estimation errors are amplified in the rendered avatar. **DW-Pose Interpolation** [52]: Detects 2D keypoints on anchor frames using DW-Pose and interpolates the skeleton; produces smooth motion, but visual quality depends on the rendering stage. **UniAnimate-DiT** [49]: Uses one reference image and interpolated DW-Pose trajectory; a diffusion model generates a full signing animation blending the reference identity and motion path.

This side-by-side layout makes it easy to spot each method's trade-offs: from fig: 8 FramePack best preserves real-world texture and photorealism, Spoken2Sign and DW-Pose offer structurally coherent motion, while UniAnimate-DiT can produce anonymous photorealistic results but drifts in fine hand details.

5.6 Step-by-Step Walk-Through

Figure 9 presents an illustrative example of the proposed SLP pipeline for a medical use case with the input sentence "Sleep, rest, meditation help mental health." The sentence is first converted into an ISL gloss sequence using our text-to-gloss translator (Section 3.2). For each gloss, contextually accurate sign clips are retrieved from the vector database; if the match score is below 40%, the corresponding word is finger-spelled using recorded alphabet clips from the same signer. The retrieved gloss segments are then interpolated and concatenated to form a continuous sentence, with neutral poses inserted at the start and end to ensure completeness. We interpolate eight intermediate frames between each pair of original frames, yielding smooth, fluid, and natural transitions between signs.

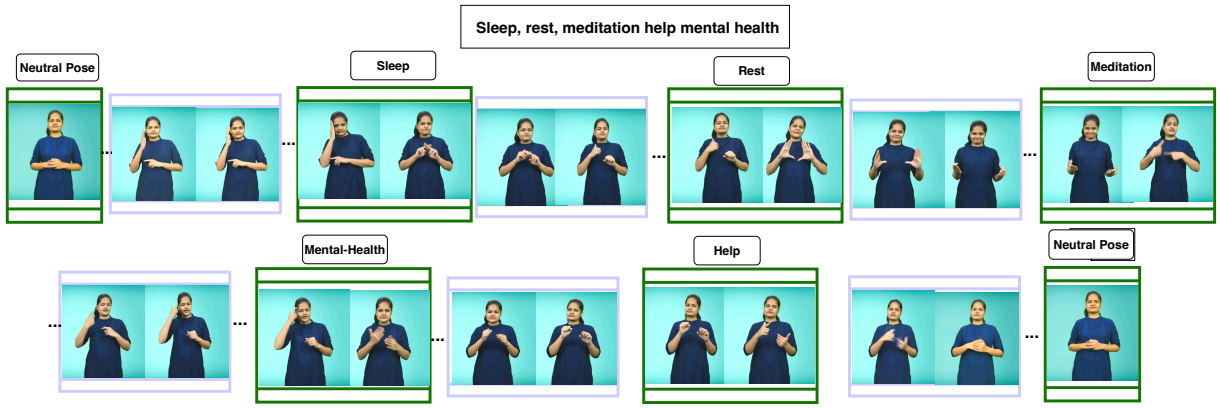

Figure 9: Illustrative example of the proposed SLP pipeline in a medical use case. The first row depicts the input sentence for inference. Green boxes indicate retrieved sign clips, purple boxes denote interpolated segments, and the first and last frames correspond to neutral poses.

6 User Study

The primary goal of our system is to narrow the communication gap between the deaf and the hearing. Direct feedback from Deaf signers is crucial. Conducting a user study allows us to assess the naturalness and accuracy of generated signing while identifying usability challenges that automated metrics cannot capture.

Method	Only interpolation			Comparision in video interpolated		
	MR↓	#1↑	B↑	MR↓	#1↑	B↑
Our Pipeline	2.47	**13**	1.53	**2.06**	**12**	**1.94**
DreamMover	**2.25**	6	**1.75**	2.89	7	1.11
VFIMamba	2.61	9	1.39	2.36	7	1.64
BMIVFI	2.67	8	1.33	2.69	10	1.31

Table 6: Subjective ranking results. Lower mean-rank (MR↓) is better; higher #1 votes and Borda[15] score (B↑) are better.

Method	Understandability↑	Fluency↑	Natural↑	Overall MOS↑
Our Pipeline	**3.12**	**3.00**	**3.33**	**3.15**
Spoken2Sign	2.62	2.42	2.38	2.47
UniAnimate	2.08	2.08	2.00	2.06

Table 7: Mean-opinion scores (MOS) from 12 raters. Scale 1 = poor and 5 = excellent.

The survey collected responses from 12 participants. Table 6 lists, for each method, its mean rank (MR ↓), average Borda score (B ↑), and number of times ranked #1. The first part compares pure interpolation quality: each clip shows only the two anchor frames and the synthetically interpolated in-betweens, played back at half speed so raters can focus on motion smoothness and artefacts. The second part rates full-sentence videos where, as mentioned in Section 5.4, seven synthetic frames are replaced at every gloss boundary. The dual view separates raw interpolation quality from impact in continuous signing. 5 suggests that out of four interpolation methods, Framepack currently has the least distortion from the user's perspective.

As summarised in Table 7, the 12-rater study consistently favours FramePack: it achieves the highest mean-opinion score on every dimension and an overall MOS of 3.15/5. Spoken2Sign occupies the middle tier (overall 2.47), while UniAnimate performs least favourably (overall 2.06). The results confirm that, on average, raters find FramePack's output noticeably more understandable, fluent, and natural than the two baseline methods.

7 Conclusion and Future Work

This work presents a retrieval-based, interpolation pipeline for ISL video synthesis, leveraging a single signer and gloss-level segmentation to enable fluent, expressive outputs with minimal recording overhead. The proposed method allows for scalable content generation and efficient updates while preserving facial and articulatory features critical for sign intelligibility. To support this approach, we introduce SanketVaani-1K, a domain-diverse ISL dataset with a gloss-annotated subset containing time-aligned clips suitable for retrieval and interpolation. In a blind user study, outputs from our pipline were preferred by Deaf over baseline methods.

While effective, our approach has certain limitations. Using a single signer improves articulation but limits generalization across signer variation. Manual gloss boundaries are approximate, which may lead to minor temporal misalignments or co-articulation artifacts during segment reuse. Fixed-length interpolation may also fail to capture the true duration of glosses, especially in natural signing. Anonymity is generally not feasible because preserving the signer's expressions and articulations is central to the method. ISL lexemes are still being standardized, so some label drift is expected. In the future, to mitigate these, we aim to refine onsets/offsets with boundary-aware cues (motion and phonological landmarks), add a duration predictor for context-dependent interpolation, align labels with ISLRTC updates, and incorporate stronger Deaf representation through advisory input, annotation review, and evaluation representation in the video recording.

Acknowledgments

We gratefully acknowledge the interpreter for annotating and recording the data, and the Deaf community for their invaluable feedback. This work is supported by MeitY, Government of India, through the NLTM-Bhashini project.

References

[1] Francisca Adoma Acheampong, Henry Nunoo-Mensah, and Wenyu Chen. 2021. Transformer models for text-based emotion detection: a review of BERT-based approaches. *Artificial Intelligence Review* 54, 8 (2021), 5789–5829.

[2] Samuel Albanie, Gül Varol, Liliane Momeni, Triantafyllos Afouras, Joon Son Chung, Neil Fox, and Andrew Zisserman. 2020. BSL-1K: Scaling up co-articulated sign language recognition using mouthing cues. In *European conference on computer vision*. Springer, 35–53.

[3] Samuel Albanie, Gül Varol, Liliane Momeni, Hannah Bull, Triantafyllos Afouras, Himel Chowdhury, Neil Fox, Bencie Woll, Rob Cooper, Andrew McParland, et al. 2021. Bbc-oxford british sign language dataset. *arXiv preprint arXiv:2111.03635* (2021).

[4] Zhaoyi An and Rei Kawakami. 2025. Teach Me Sign: Stepwise Prompting LLM for Sign Language Production. *arXiv preprint arXiv:2507.10972* (2025).

[5] Safaeid Hossain Arib, Rabeya Akter, Sejuti Rahman, and Shafin Rahman. 2025. SignFormer-GCN: Continuous sign language translation using spatio-temporal graph convolutional networks. *PloS one* 20, 2 (2025), e0316298.

[6] Rotem Shalev Arkushin, Amit Moryossef, and Ohad Fried. 2023. Ham2pose: Animating sign language notation into pose sequences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 21046–21056.

[7] Maryam Aziz and Achraf Othman. 2023. Evolution and trends in sign Language Avatar systems: Unveiling a 40-Year journey via systematic review. *Multimodal Technologies and Interaction* 7, 10 (2023), 97.

[8] Eyal Betzalel, Coby Penso, and Ethan Fetaya. 2024. Evaluation metrics for generative models: An empirical study. *Machine Learning and Knowledge Extraction* 6, 3 (2024), 1531–1544.

[9] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neural sign language translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7784–7793.

[10] Necati Cihan Camgöz, Ben Saunders, Guillaume Rochette, Marco Giovanelli, Giacomo Inches, Robin Nachtrab-Ribback, and Richard Bowden. 2021. Content4all open research sign language translation datasets. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*. IEEE, 1–5.

[11] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The Faiss library. (2024). arXiv:2401.08281

[12] Philippe Dreuw, Carol Neidle, Vassilis Athitsos, Stanley Sclaroff, and Hermann Ney. 2008. Benchmark databases for video-based automatic sign language recognition. In *Proceedings of Language Resources and Evaluation Conference (LREC) 2008*. EUROPEAN LANGUAGE RESOURCES ASSOC-ELRA.

[13] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2735–2744.

[14] Abhishek Dutta and Andrew Zisserman. 2019. The VIA annotation software for images, audio and video. In *Proceedings of the 27th ACM international conference on multimedia*. 2276–2279.

[15] Peter C Fishburn. 2015. *The theory of social choice*. Princeton University Press.

[16] Jens Forster, Christoph Schmidt, Oscar Koller, Martin Bellgardt, and Hermann Ney. 2014. Extensions of the Sign Language Recognition and Translation Corpus RWTH-PHOENIX-Weather.. In *LREC*. 1911–1916.

[17] 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)*. New Delhi, India, 123–127.

[18] Jianyuan Guo, Peike Li, and Trevor Cohn. 2025. Bridging Sign and Spoken Languages: Pseudo Gloss Generation for Sign Language Translation. *arXiv preprint arXiv:2505.15438* (2025).

[19] Mert Inan, Katherine Atwell, Anthony Sicilia, Lorna Quandt, and Malihe Alikhani. 2024. Generating Signed Language Instructions in Large-Scale Dialogue Systems. *arXiv preprint arXiv:2410.14026* (2024).

[20] Youngjoon Jang, Haran Raajesh, Liliane Momeni, Gül Varol, and Andrew Zisserman. 2025. Lost in translation, found in context: Sign language translation with contextual cues. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 8742–8752.

[21] Zifan Jiang, Anne Göhring, Amit Moryossef, Rico Sennrich, and Sarah Ebling. 2024. SwissSLi: The Multi-parallel Sign Language Corpus for Switzerland. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. 15448–15456.

[22] Ronan Johnson. 2021. Towards enhanced visual clarity of sign language avatars through recreation of fine facial detail. *Machine Translation* 35, 3 (2021), 431–445.

[23] Abhinav Joshi, Susmit Agrawal, and Ashutosh Modi. 2023. ISLTranslate: Dataset for translating Indian sign language. *arXiv preprint arXiv:2307.05440* (2023).

[24] Abhinav Joshi, Romit Mohanty, Mounika Kanakanti, Andesha Mangla, Sudeep Choudhary, Monali Barbate, and Ashutosh Modi. 2024. isign: A benchmark for indian sign language processing. *arXiv preprint arXiv:2407.05404* (2024).

[25] Sang-Ki Ko, Chang Jo Kim, Hyedong Jung, and Choongsang Cho. 2019. Neural Sign Language Translation Based on Human Keypoint Estimation. *Applied Sciences* 9, 13 (2019), 2683. https://doi.org/10.3390/app9132683

[26] Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xinchi Deng, Yingfang Zhang, Xingchao Liu, Minbin Huang, Zedong Xiao, et al. 2024. Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding. *arXiv preprint arXiv:2405.08748* (2024).

[27] Yuqi Liu, Wenqian Zhang, Sihan Ren, Chengyu Huang, Jingyi Yu, and Lan Xu. 2025. SCOPE: Sign Language Contextual Processing with Embedding from LLMs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 5739–5747.

[28] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. 2023. SMPL: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*. 851–866.

[29] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. 2019. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172* (2019).

[30] Aleix M Martínez, Ronnie B Wilbur, Robin Shay, and Avinash C Kak. 2002. Purdue RVL-SLLL ASL database for automatic recognition of American Sign Language. In *Proceedings. Fourth IEEE International Conference on Multimodal Interfaces*. IEEE, 167–172.

[31] Hamam Mokayed, Ghada Alsayed, Felicia Lodin, Olle Hagner, and Björn Backe. 2024. Enhancing object detection in snowy conditions: Evaluating yolo v9 models with augmentation techniques. In *2024 11th International Conference on Internet of Things: Systems, Management and Security (IOTSMS)*. IEEE, 198–203.

[32] Amit Moryossef, Mathias Müller, Anne Göhring, Zifan Jiang, Yoav Goldberg, and Sarah Ebling. 2023. An open-source gloss-based baseline for spoken to signed language translation. In *Proceedings of the Second International Workshop on Automatic Translation for Signed and Spoken Languages*. 22–33.

[33] B. O. Olusanya, A. C. Davis, and H. J. Hoffman. 2019. Hearing loss: rising prevalence and impact. *Bulletin of the World Health Organization* 97, 10 (2019), 646–646A. https://doi.org/10.2471/BLT.19.224683

[34] Xiankun Pei, Dan Guo, and Ye Zhao. 2019. Continuous sign language recognition based on pseudo-supervised learning. In *Proceedings of the 2nd Workshop on Multimedia for Accessible Human Computer Interfaces*. 33–39.

[35] Xuebin Qin, Hang Dai, Xiaobin Hu, Deng-Ping Fan, Ling Shao, and Luc Van Gool. 2022. Highly Accurate Dichotomous Image Segmentation. In *ECCV*.

[36] Lorna C Quandt, Athena Willis, Melody Schwenk, Kaitlyn Weeks, and Ruthie Ferster. 2022. Attitudes toward signing avatars vary depending on hearing status, age of signed language acquisition, and avatar type. *Frontiers in psychology* 13 (2022), 730917.

[37] Elakkiya R and Natarajan B. 2021. *ISL-CSLTR: Indian Sign Language Dataset for Continuous Sign Language Translation and Recognition*. https://doi.org/10.17632/kcmpdxky7p.1

[38] Charles Raude, KR Prajwal, Liliane Momeni, Hannah Bull, Samuel Albanie, Andrew Zisserman, and Gül Varol. 2024. A tale of two languages: Large-vocabulary continuous sign language recognition from spoken language supervision. *arXiv preprint arXiv:2405.10266* (2024).

[39] David M Rouse and Sheila S Hemami. 2008. Analyzing the role of visual structure in the recognition of natural image content with multi-scale SSIM. In *Human vision and electronic imaging XIII*, Vol. 6806. SPIE, 410–423.

[40] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2020. Everybody sign now: Translating spoken language to photo realistic sign language video. *arXiv preprint arXiv:2011.09846* (2020).

[41] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2022. Signing at scale: Learning to co-articulate signs for large-scale photo-realistic sign language production. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5141–5151.

[42] Wonyong Seo, Jihyong Oh, and Munchurl Kim. 2025. BiM-VFI: Bidirectional Motion Field-Guided Frame Interpolation for Video with Non-uniform Motions. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 7244–7253.

[43] Abhimanyu Sharma. 2025. India's language policy for deaf and hard-of-hearing people. *Language Policy* (2025), 1–23.

[44] Liao Shen, Tianqi Liu, Huiqiang Sun, Xinyi Ye, Baopu Li, Jianming Zhang, and Zhiguo Cao. 2024. Dreammover: Leveraging the prior of diffusion models for image interpolation with large motion. In *European Conference on Computer Vision*. Springer, 336–353.

[45] Bowen Shi, Diane Brentari, Greg Shakhnarovich, and Karen Livescu. 2022. Open-domain sign language translation learned from online video. *arXiv preprint arXiv:2205.12870* (2022).

[46] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. 2019. First Order Motion Model for Image Animation. In *Conference on Neural Information Processing Systems (NeurIPS)*.

[47] Dave Uthus, Garrett Tanzer, and Manfred Georg. 2023. Youtube-asl: A large-scale, open-domain american sign language-english parallel corpus. *Advances in Neural Information Processing Systems* 36 (2023), 29029–29047.

[48] Harry Walsh, Ben Saunders, and Richard Bowden. 2024. Sign stitching: a novel approach to sign language production. *arXiv preprint arXiv:2405.07663* (2024).

[49] Xiang Wang, Shiwei Zhang, Longxiang Tang, Yingya Zhang, Changxin Gao, Yuehuan Wang, and Nong Sang. 2025. Unianimate-dit: Human image animation with large-scale video diffusion transformer. *arXiv preprint arXiv:2504.11289* (2025).

[50] Wufeng Xue, Lei Zhang, Xuanqin Mou, and Alan C Bovik. 2013. Gradient magnitude similarity deviation: A highly efficient perceptual image quality index. *IEEE transactions on image processing* 23, 2 (2013), 684–695.

[51] Peiqing Yang, Shangchen Zhou, Jixin Zhao, Qingyi Tao, and Chen Change Loy. 2025. MatAnyone: Stable Video Matting with Consistent Memory Propagation. In *CVPR*.

[52] Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. 2023. Effective whole-body pose estimation with two-stages distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4210–4220.

[53] Yu Yu, Weibin Zhang, and Yun Deng. 2021. Frechet inception distance (fid) for evaluating gans. *China University of Mining Technology Beijing Graduate School* 3, 11 (2021).

[54] Guozhen Zhang, Chuxnu Liu, Yutao Cui, Xiaotong Zhao, Kai Ma, and Limin Wang. 2024. Vfimamba: Video frame interpolation with state space models. *Advances in Neural Information Processing Systems* 37 (2024), 107225–107248.

[55] Han Zhang, Rotem Shalev-Arkushin, Vasileios Baltatzis, Connor Gillis, Gierad Laput, Raja Kushalnagar, Lorna C Quandt, Leah Findlater, Abdelkareem Bedri, and Colin Lea. 2025. Towards AI-driven Sign Language Generation with Non-manual Markers. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–26.

[56] Lvmin Zhang and Maneesh Agrawala. 2025. Packing Input Frame Contexts in Next-Frame Prediction Models for Video Generation. *Arxiv* (2025).

[57] Lin Zhang, Lei Zhang, and Alan C Bovik. 2015. A feature-enriched completely blind image quality evaluator. *IEEE Transactions on Image Processing* 24, 8 (2015), 2579–2591.

[58] Pengyu Zhang, Hao Yin, Zeren Wang, Wenyue Chen, Shengming Li, Dong Wang, Huchuan Lu, and Xu Jia. 2024. Evsign: Sign language recognition and translation with streaming events. In *European Conference on Computer Vision*. Springer, 335–351.

[59] Yanqiong Zhang and Xianwei Jiang. 2024. Recent Advances on Deep Learning for Sign Language Recognition. *Computer Modeling in Engineering & Sciences (CMES)* 139, 3 (2024).

[60] Weichao Zhao, Hezhen Hu, Wengang Zhou, Yunyao Mao, Min Wang, and Houqiang Li. 2024. Masa: Motion-aware masked autoencoder with semantic alignment for sign language recognition. *IEEE Transactions on Circuits and Systems for Video Technology* 34, 11 (2024), 10793–10804.

[61] 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*. 20871–20881.

[62] Hao Zhou, Wengang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li. 2021. Improving sign language translation with monolingual data by sign back-translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 1316–1325.

[63] Jingchen Zou, Jianqiang Li, Jing Tang, Yuning Huang, Shujie Ding, and Xi Xu. 2024. Sign Language Recognition and Translation Methods Promote Sign Language Education: A Review. In *2024 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 3479–3484.

[64] Ronglai Zuo, Fangyun Wei, Zenggui Chen, Brian Mak, Jiaolong Yang, and Xin Tong. 2024. A simple baseline for spoken language to sign language translation with 3d avatars. In *European Conference on Computer Vision*. Springer, 36–54.