

Towards Natural Sign Language Generation: Scalable and Fluid Sign Synthesis for Inclusive Communication.

Thesis submitted in partial fulfillment
of the requirements for the degree of

Master of Science
in
Computer Science and Engineering by Research

by

Aparna Agrawal
2021121007

aparna.agrawal@research.iiit.ac.in



International Institute of Information Technology
Hyderabad - 500032, INDIA
June, 2026

Copyright © Aparna Agrawal, 2025
All Rights Reserved

International Institute of Information Technology
Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled “*Towards Natural Sign Language Generation: Scalable and Fluid Sign Synthesis for Inclusive Communication.*” by Aparna Agrawal, has been carried out under my supervision and is not submitted elsewhere for a degree.

30.06.2026

Date

Advisor:

Prof. C.V. Jawahar

Co-Advisor:

Prof. Vinay Namboodri

To those who dream boldly and act bravely.

Acknowledgments

This thesis is the product of the collective efforts and continuous support from my advisors, family and friends.

My sincere respect and profound gratitude are owed to my advisors, Prof. C.V. Jawahar and Prof. Vinay Namboodri. Their relentless support, patient guidance, and belief in my abilities have shaped this research. I am deeply thankful for their insightful discussions, constructive criticism, and the intellectual freedom they provided throughout this journey. Their mentorship continuously nurtured my curiosity, encouraged me to question assumptions, and motivated me to explore ideas with depth and purpose, shaping not just the outcomes of this thesis but also the way I think as a researcher. They have helped me appreciate the deeper rigor of research, not only developing ideas but validating them with clarity, evidence, and thoughtful execution. I will carry forward the values they have instilled in me curiosity, clarity of thought, and responsibility toward meaningful research as guiding principles throughout my academic and professional journey.

I am grateful to my brothers, who have been a constant source of strength. They played an important role in my journey to IIITH, a decision that has been one of the greatest blessings of my life, opening doors to meaningful opportunities and remarkable people. Their support has been unwavering throughout: encouraging me at every stage, listening to my endless rants, celebrating my milestones, and standing by me through every phase of this journey.

I would like to thank Anchit, Rudharabha, Surabhi and Jayshree for helping me settle into the lab and for supporting me in the early phase of my research journey. I am also grateful to Soumya, Varun, Snigdha, and Seshadri for being wonderful colleagues and for creating a collaborative and encouraging environment in the lab. Varun has been especially supportive in the initial years, often listening to me, calming my nerves before meetings, and discussing ideas whenever I needed clarity. I also Thank Sony, Cecilia, Ram for their active support.

At CVIT, I was fortunate to build friendships that shaped this journey in ways I did not expect. Astiva, Amogh, Shantika, Rahul, Ayan, Deepti, Adhiraj, Shreya, Chirag, Pranav have been there through every phase, guiding me when I needed direction, pushing me through deadlines, and inspiring me with their hard work and enthusiasm. Our conversations, whether about research or life, always left me with clarity and renewed motivation. From late-night tea at JC to our trip to Sikkim, their presence turned this journey from just a research pursuit into an experience filled with learning, laughter, and memories I will always cherish.

Each of them has supported me in their own unique way. Astiva was always there to guide me at any hour of the day. Amogh helped me through difficult times and encouraged me to believe in myself and apply for opportunities. Shantika, Deepti, Suneyna have been strong female role models whose confidence, bold decisions and dedication inspired me deeply. Rahul teased me endlessly, Ayan kept me updated with the latest research, and introduced me to more brands and coffees than I knew existed. Adhiraj stood by me through thick and thin, patiently handling my frustration and anxiety across projects and lab life. Shreya showing up for spontaneous meets and Chirag helping me whenever possible added their own warmth and support. I feel grateful to have met such kind, grounded, and genuinely good people.

Apart from my labmates, my batchmates also played a crucial role in shaping my journey at IIITH. Aditya Sehgal introduced me to robotics and helped me appreciate how computer vision and robotics intersect. We spent many hours discussing ideas and possible projects together; although we haven't worked on one yet, I look forward to the day we do. Tanmay, Maulesh, Rohan, and Chirag gave me a window into the world of hardware and neuroscience, and hearing about the challenges they tackled always inspired me. Our outings and small breaks together made life much easier and far more enjoyable.

My girl gang Khushi, Anmol, Harshita, and Divya has been with me since second year, through every high and every low. From cheering me on to grounding me when I got overwhelmed, they made sure I never felt alone in this journey.

Anjali, who has become my closest friend here, supported me throughout, understanding my deadlines, adjusting plans when I was overwhelmed, and showing up in the lab just to make things a little brighter and more fulfilling.

I also want to thank Pranav, who sat right behind me, for being both the most annoying and the kindest person at the same time. someone I could argue with, laugh with, and learn from. His honesty, liveliness, and depth in research made everyday conversations memorable.

Lastly, I would like to thank Papa, Baba, Maa, Taiji, Tajuji, Annu, and Amu Bhaiya for giving me the opportunity to come this far and for always being there for me. Your belief in me, constant encouragement, and support have been the backbone of this journey.

Abstract

Sign Language Production (SLP) is a critical technology for bridging the communication gap between the hearing and Deaf communities. However, current systems face a persistent trade-off between visual realism and semantic scalability. This thesis proposes a multi-modal framework to achieve scalable, fluid, and natural sign language synthesis across 2D and 3D representations.

First, we introduce Sanket Vaani-1K, a curated Indian Sign Language (ISL) dataset spanning over 40 high-utility domains, such as healthcare and banking. We present a 2D production pipeline that utilizes FramePack diffusion-based interpolation to stitch isolated sign units into fluid video, significantly reducing data requirements compared to end-to-end models.

Transitioning to 3D to overcome 2D limitations like occlusion and depth nuances, we utilize ExAvatar, a Gaussian-based neural renderer driven by optimized SMPL-X body models, which we modified and optimized for the specific requirements of the sign language use case.

By driving this renderer with person-specific skeletal offsets and high-fidelity facial registration, we ensure the accurate synthesis of manual and non-manual cues. This is integrated into Gloss-SLQA, a proof of concept for a sign language question-answering system that enables natural interaction through a gloss-centric 3D avatar.

Finally, we present SignWeave-3D, which treats sign production as a masked motion infilling task. By "weaving" isolated 3D motion units with a diffusion transformer, the model synthesizes natural co-articulation. Quantitative evaluations and user studies with the Deaf community demonstrate that our framework produces significantly more understandable and natural signing than existing state-of-the-art baselines.

Contents

Chapter	Page
1 Introduction	1
1.1 Sign Language Production	2
1.1.1 Motivation	2
1.1.2 Challenges	2
1.2 Scalable 2D Sign Language Production	3
1.2.1 Problem Statement	3
1.3 Scalable 3D Avatar-Based Sign Language Production	4
1.3.1 Problem Statement	4
1.4 Main Contributions of the Thesis	4
1.5 Thesis Roadmap	5
2 Background	6
2.1 Sign Language	6
2.2 Pose Estimation	8
2.2.1 2D Pose Estimation for Sign language	8
2.2.2 3D Pose Estimation for Sign Language	9
2.3 SMPL-X	10
2.4 SMPL-X Registration	11
2.5 Rendering Engines and Techniques for Avatar Synthesis	13
2.6 SMPL-X–Driven Avatar Generation and Animation	14
2.7 Sign Language Avatars with SMPL-X	15
3 Towards Scalable Sign Production: Leveraging Co-Articulated Gloss Dictionary for Fluid Sign Synthesis	17
3.1 Introduction	17
3.2 Literature Review	17
3.3 Proposed Scalable 2D Sign Language Production	18
3.3.1 Vector Database Construction	18
3.3.2 Text to Gloss Sequence	19
3.3.3 Stitching	21
3.4 Dataset: SanketVaani-1K	23
3.4.1 Data Pre-processing	24
3.4.1.1 Filtering and Synchronization	24
3.4.1.2 Signer Normalization	24
3.4.1.3 Background Replacement	25

3.4.2	Annotation	25
3.5	Experiments and Results	26
3.5.1	Data Pre-processing Evaluation	26
3.5.2	Text-to-Gloss Generation	26
3.5.3	Vector Database Retrieval and Disambiguation	27
3.5.4	Interpolation	27
3.5.5	Step-by-Step Walk-Through	30
3.6	User Study	31
3.7	Summary	32
3.8	Acknowledgment	33
4	Gloss SLQA: A Gloss-centric Approach to Sign Language Question Answering	34
4.1	Introduction	34
4.2	Dataset	35
4.3	Methodology	36
4.3.1	Gloss-Motion Video Dictionary: Leveraging Multi-Signer Data	36
4.3.2	Text to gloss	36
4.3.3	Answer to Gloss Generation	36
4.4	Experiments and Results	40
4.4.1	Quantitative Analysis	40
4.4.2	Qualitative and Human Evaluation	40
4.5	Sign Language Production	41
4.5.1	Human Evaluation	41
4.5.2	Scaled Opinion Survey	41
4.5.3	Deaf Community Feedback	42
4.5.3.0.1	Interpretation:	43
4.5.4	Qualitative Analysis	44
4.6	Application: Gloss-Centric Sign Language Question Answering (Gloss-SLQA)	44
4.6.0.0.1	1. Sign Language Retrieval (Query Understanding):	45
4.6.0.0.2	2. Question & Answer Generation (LLM Reasoning):	46
4.6.0.0.3	3. Sign Language Production (Response Synthesis):	47
4.7	Limitations	47
4.8	Discussion and Future Work	49
4.9	Summary	50
5	SignWeave-3D: Weaving Isolated Sign Units into Continuous Co-Articulated Motion	51
5.1	Introduction	51
5.2	Literature Review	52
5.3	Methodology	53
5.3.1	Preliminary	54
5.3.1.0.1	Motion Representation:	54
5.3.2	SignWeave-3D Approach	55
5.3.2.0.1	Training Strategy.	55
5.3.2.0.2	Inference Pipeline.	56
5.3.3	SignWeave-3D Architecture	56
5.3.4	Curriculum-Based Masking Strategy	56

5.3.5	Sign Stitching via Dictionary Retrieval	57
5.4	Experiments and Results	57
5.4.1	Experimental Setup	58
5.4.1.1	Dataset and Preprocessing	58
5.4.1.2	Implementation Details	58
5.4.2	Evaluation Metrics	58
5.4.3	Comparison with Text-to-Motion Baselines	59
5.4.4	Evaluating Co-Articulation and Stitching	60
5.4.4.1	Results Analysis	60
5.4.5	Ablation Study: Inference-Time Initialization	61
5.4.6	Comparing MR Precision	62
5.4.7	Human Evaluation	62
5.5	Limitations and Future Work	63
5.6	Discussion	64
6	Conclusion	65
6.0.1	Impact	65
6.0.2	Discussion and Limitations	65
6.0.3	Future Work	66
	Bibliography	69

List of Figures

Figure	Page
2.1 2D pose estimation. Example of Mediapipe Holistic Method	8
2.2 3D pose estimation. Example: SMPL-X representation from Ex-avatar SMPL-X registration method.	9
3.1 Stitching Pipeline: includes Text to gloss Translator, formation of Vector database, and video retrieval, and sign stitching with interpolation.	18
3.2 Comprehensive category-wise distribution of the ISL dataset, demonstrating coverage across over 40 high-utility domains.	23
3.3 Distribution of sentence word counts in the ISL dataset. The majority of sentences fall within the 10–30 word range.	24
3.4 Example annotation workflow showing the source sentence, video frames, and the corresponding temporally aligned glosses and sign-type tags.	25
3.5 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.	26
3.6 Qualitative comparison of DreamMover, VIFMamba, BIM-VFI, and FramePack interpolation methods.	29
3.7 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.	30
3.8 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.	31
4.1 Overview of the Gloss centric Sign Language Question Answering (Gloss SLQA) pipeline. A user inputs a question in sign language (top row), which is converted into glosses (second row), followed by text-based question and answer generation using a large language model (third and fourth rows). The generated answer is then translated back into glosses (fifth row) and synthesized into a sign language video output using a Gaussian-based avatar driven by SMPL-X models (last row).	34
4.2 Qualitative comparison of SLP methods. Rows from top to bottom: Ground Truth, SMPLer-X, Spoken2Sign, and ExAvatar* (Ours). Note the improved hand definition and clothing detail in the ExAvatar* row.	41

4.3	Education domain(ASL) SLP comparison of answer: "An angle is made by two rays meeting at a common endpoint."	45
4.4	Environment domain(ASL) SLP comparison of answer: "Chemical rain damages tree, remove nutrients from soil..."	45
4.5	Medical domain(ASL) SLP comparison of answer: "Reduce stress by practicing relaxation techniques, such as deep breathing."	46
4.6	Science Tech. domain (BSL) SLP comparison of answer: "Tsunamis are caused by underwater earthquakes or volcanic eruptions."	47
4.7	Geography (BSL) SLP comparison of answer: "This area is called an estuary, where sediment accumulates at the river mouth."	48
4.8	Chemistry(BSL) SLP comparison of answer: "Enzymes act as catalysts, speeding up chemical reactions."	48
4.9	Overview of the Gloss-SLQA pipeline. The system processes a user's sign language query through retrieval, reasons using an LLM to generate a text response, and synthesizes the final answer back into sign language using our Gloss-Centric 3D Avatar framework.	49
5.1	Overview: We propose a co-articulated sign language synthesis framework that integrates masked language modeling with a diffusion transformer conditioned on text. The pipeline begins with text input, followed by gloss retrieval from a dictionary and masked transition initialization. A diffusion-based model, SignWeave-3D, progressively refines the motion to produce smooth, contextually grounded sign sequences.	51
5.2	Overview of the SignWeave-3D architecture. The framework retrieves isolated glosses, initializes the sequence with masked transitions, and uses a Sign-PixArt- α diffusion backbone to progressively denoise and infill the co-articulated motion.	53
5.3	Qualitative comparison of text-to-3D sign language motion generation. Each row shows temporal frame sequences for T2M-GPT, MDM Curriculum, and SignWeave-3D Curriculum, with GT (Ground Truth) on top for reference.	59
5.4	Qualitative comparison on co-articulation prediction. Models predict 30 frames (blue boxes) given observed context (green boxes). MDM (middle) shows visible pose shifts at boundaries. SignWeave-3D (bottom) generates smooth, continuous transitions that naturally blend the context.	60

List of Tables

Table	Page
3.1 Example English sentences with corresponding ISL gloss sequences.	21
3.2 Comparison of sign language datasets. The <i>Controlled Env.</i> column indicates whether the dataset was recorded in a controlled environment (✓) or curated from existing sources (✗).	22
3.3 Automatic evaluation of generated gloss sequences on the RWTH-PHOENIX-Weather 2014T test set using BLEU, ROUGE-L, METEOR, CHrF, and BERTScore.	27
3.4 Evaluation of gloss sequence generation using TSOV word-order accuracy and BERTScore.	27
3.5 Top-3 sentence retrieval for five ambiguous queries with different meanings in the context.	28
3.6 Quantitative comparison of interpolation methods. Higher is better for PSNR, SSIM, MS-SSIM; lower is better for LPIPS, FID, NIQE, GMSD, and KID.	29
3.7 Subjective ranking results. Lower mean rank (MR↓) is better; higher #1 votes and Borda score (B↑) are better.	32
3.8 Mean opinion scores (MOS) from 12 raters. Scale: 1 = poor, 5 = excellent.	32
4.1 Vocabulary Statistics for Sign Language Production (SLP)	36
4.2 Comparison of motion quality and semantic intelligibility. While Spoken2Sign achieves lower keypoint error, our ExAvatar*-based approach achieves higher back-translation accuracy, indicating better preservation of semantic features.	40
4.3 Mean Opinion Scores (MOS) from user survey. Each metric was rated on a 5-point Likert scale (higher is better). Combined Score is the unweighted average across all metrics.	42
4.4 Mean Opinion Scores (MOS) from a user survey with 10 Deaf participants. Each metric was rated on a 5-point Likert scale (higher is better). Combined Score is the unweighted average across all metrics.	43
4.5 Semantic evaluation of signed video prompts (Q1–Q4). Gen. vs GT and Stitch. vs GT compare participant translations with ground truth (GT) for generated and stitched videos, respectively, rated on a 3-point Likert scale (<i>Different</i> = 1, <i>Similar</i> = 2, <i>Same</i> = 3). Sem. Rating reflects participant-perceived semantic clarity on a 5-point scale. . .	43
5.1 Quantitative comparison of text-to-sign motion generation models. SignWeave-3D demonstrates superior semantic alignment (Top-k Retrieval) compared to baselines while maintaining competitive geometric fidelity. Note the significant drop in T2M-GPT’s retrieval performance, indicating a lack of semantic precision.	59

5.2	Evaluation of curriculum-based prediction under varying observation lengths (obs). The top table reports multi-reference (MR) precision retrieval scores and AvgDTW across Hand, Face, Hand+Arm, and Overall regions for both MDM and SignWeave-3D models across different prediction lengths (pred) (10, 20, 30 frames). The bottom table presents corresponding geometric accuracy metrics, MPJPE, PA-MPJPE, and velocity error for the same setups. Interp: SLERP [87] interpolated frames initialization, Const: constant last frame initialization	61
5.3	MR-Precision on PHOENIX-2014T based on different evaluators. EPT is an evaluator trained by SignAvatars on Progressive Transformers [78] using 3D SMPL-X representations. ESA is an evaluator trained by SignAvatars on sign-avatar-fitted embeddings. .	62
5.4	Human evaluation (MOS) across four perceptual criteria (smooth trajectories, fluid transitions, naturalness) and three prediction horizons (Obs10→Pred10/Pred20/Pred30). Scores reflect the mean rating from 13 hearing participants on a 1–5 Likert scale (higher is better) for naïve stitching, MLM-based baselines, and the proposed SignWeave-3D models.	63

Chapter 1

Introduction

Language is a rich, multilayered system through which humans articulate complex ideas, construct social understanding, and negotiate shared meaning. Yet, despite its central role in human interaction, much of the world’s linguistic diversity remains marginalized within mainstream communication technologies. **Sign languages**, primary communication systems for Deaf communities, illustrate this disparity most clearly, as they remain underrepresented in digital infrastructures despite their linguistic sophistication and societal importance. Consequently, large portions of informational and cultural content circulating in society remain effectively inaccessible to Deaf and hard-of-hearing individuals. Live news broadcasts, emergency alerts, televised debates, classroom lectures, government proceedings, and a substantial portion of online video media are produced primarily for the hearing public. Live news broadcasts, emergency alerts, televised debates, classroom lectures, government proceedings, and a substantial portion of online video media are produced primarily for the hearing public.

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 [58]. In India, a WHO news report estimates that 63 million people have “significant auditory impairment” [84].¹ 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 it effectively.

Sign language communication fundamentally depends on two capacities: expressing oneself clearly and comprehending others. While substantial advances have been achieved in automatic sign language recognition, generating accurate, natural, and intelligible sign language, referred to as Sign Language Production (SLP), remains a significant challenge. Existing accessibility measures, such as captioning, inset interpreters on television, and limited sign-language translations of public announcements, are heavily reliant on manual human effort, making them resource-intensive and difficult to scale. The acute shortage of trained professionals further compounds this difficulty: the Indian Sign Language (ISL) Research and Training Centre (ISLRTC) currently lists only 304 Level-C (DISLI) certified ISL

¹Estimates 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. [84].

interpreters across six regions in India. This pronounced imbalance between communication needs and available expertise underscores the urgent necessity of developing scalable SLP systems capable of providing accessible and affordable support to Deaf and hard-of-hearing communities. This Thesis aims to address these critical gaps by developing a framework for scalable and fluid sign language synthesis.

1.1 Sign Language Production

1.1.1 Motivation

Sign languages, much like spoken languages, are fully fledged linguistic systems with their own grammar and vocabulary. They rely on a unique combination of manual and non-manual channels, used asynchronously to encode meaning [75]. Manual components encompass hand shape, movement, and placement of the hands and arms, while non-manual components include facial expressions, head motions, and overall body posture. Because of this rich multimodal structure, translating between spoken and signed languages is highly non-trivial and cannot be achieved through direct word-to-sign mapping; instead, it typically requires skilled human interpreters. Unfortunately, the limited availability of trained interpreters restricts consistent access to information for Deaf individuals. Sign Language Production (SLP), the task of generating sign language from spoken or written language, offers a promising direction toward mitigating this accessibility gap. This thesis aims to enhance Sign Language Production to support fluent and Scalable Sign Language Production.

Most sign-language technologies focus on recognition tasks [38] for domains such as HCI [89], education [107, 59], rehabilitation [72], and healthcare [52, 93]. while SLP has received comparatively limited attention. This underinvestment is partly driven by the misconception that Deaf individuals can rely entirely on written forms of spoken languages, and thus do not require sign-language translations. In reality, many Deaf users are not fully proficient in the written modality of spoken languages, reading text is challenging for DHH individuals due to the lack of association of spoken text with SL like how hearing individual associate text with sound first, a direct link between text and SL making text-only communication insufficient. Recent surveys further emphasize this need by showing a strong preference among Deaf communities for sign-language visual-modality support in language and learning technologies, including LLMs [32].

1.1.2 Challenges

Linguistic Complexity of Sign Languages Sign languages exhibit grammatical structures fundamentally different from spoken languages. Their meaning is encoded through a combination of manual (handshape, orientation, trajectory) and non-manual signals (facial expression, mouthing, gaze, body posture). Unlike linear speech or text, SL grammar incorporates simultaneity, spatial indexing, classifier constructions, and role shift, which standard NLP pipelines designed for sequential text, cannot directly model.

Limited Annotated Data

Large-scale, high-quality sign-language datasets, particularly continuous signing corpora with glosses and non-manual marker annotations, are scarce due to the cost and expertise required for annotation. This limits the training of data-hungry generative models.

Lack of Standardized Evaluation

Evaluation metrics for sign production are fragmented. Existing measures from NLP (BLEU, ROUGE), vision (MPJPE, FID), and speech do not adequately capture linguistic correctness, spatial accuracy, handshape fidelity, or non-manual alignment. The absence of standardized, SL-specific benchmarks complicates fair comparison and model development.

Regional and Dialectal Variation

Sign languages vary widely across regions, communities, and educational backgrounds. For languages ISL, variation across states creates difficulty in building a unified, scalable generative model.

Pose and Motion Generation Limitations

Off-the-shelf 2D/3D pose generators are not tailored for sign language. They often fail to capture fine-grained hand articulations, finger movements, wrist rotations, and subtle non-manual cues, leading to semantically incorrect or unintelligible output. Generating temporally coherent, linguistically valid motion sequences remains a core challenge.

Enhancing Real-Time Performance

Achieving high accuracy alongside rapid processing speeds is vital to facilitating seamless communication. Prospective research should investigate the development of compact deep-learning models and efficient inference strategies tailored for real-time execution on devices with limited resources.

1.2 Scalable 2D Sign Language Production

1.2.1 Problem Statement

Current Sign Language Production (SLP) systems face a critical trade-off between visual realism and semantic scalability. Traditional 3D avatar-based systems often suffer from the "uncanny valley" effect and robotic motion, which limits user acceptance and comprehensibility among Deaf communities. Conversely, fully generative deep learning models, such as GANs/diffusion, require massive paired training datasets (text-to-video) that simply do not exist for low-resource languages like Indian Sign Language (ISL). keypoint based methods suffer from occlusion Furthermore, end-to-end models struggle to generalize to unseen vocabulary, making them unsuitable for dynamic domains like news or education. The core problem, therefore, lies in resolving the conflict between data scarcity and visual fidelity. Rather than relying on massive, end-to-end training corpora that are unavailable for most sign languages, the challenge is to develop a modular framework that can synthesize open-vocabulary content by effectively re-purposing high-quality, isolated sign units. The goal is to achieve the fluidity

of continuous signing through advanced interpolation techniques, thereby enabling scalable production without compromising on the photorealistic detail required for sign language intelligibility.

1.3 Scalable 3D Avatar-Based Sign Language Production

1.3.1 Problem Statement

Despite their high visual fidelity, 2D interpolation techniques primarily operate in pixel space, which severely limits their ability to model the complex, non-linear kinematics of human motion. In sign language, where meaning is conveyed through precise spatial relationships, 2D methods often fail to capture the depth nuances required for accurate interpretation. Furthermore, 2D keypoint estimation methods are inherently susceptible to occlusion errors. Complex signs frequently involve hands crossing the face or torso; in a 2D projection, these articulates effectively disappear, causing tracking failures that disrupt the semantic continuity of the generated sign.

3D avatar-based representations explicitly model skeletal geometry and motion, enabling robust occlusion handling, identity anonymization, and viewpoint-independent rendering. These characteristics make 3D avatars especially suitable for multi-view signing and immersive AR/VR applications. Recent advances in Gaussian-based neural rendering enable near real-time, photorealistic 3D avatar signing by representing geometry and appearance with lightweight Gaussian primitives that can be efficiently rendered on modern GPUs. This approach bridges the gap between realism and efficiency, allowing fine-grained modeling of facial expressions and clothing dynamics while maintaining the low-latency performance required for interactive and scalable sign language production systems.

1.4 Main Contributions of the Thesis

This thesis makes the following key contributions to the field of Sign Language Production (SLP):

A Scalable Framework for 2D Sign Language Production: We propose a novel, data-efficient architecture capable of generating temporally coherent 2D sign language poses. We introduce SanketVaani-1K, a curated ISL dataset with fine-grained gloss-boundary annotations, and demonstrate that high-quality continuous sign videos can be synthesized by stitching isolated gloss clips using "FramePack" interpolation, significantly reducing data requirements compared to end-to-end models.

Gloss SLQA: A Gloss-centric Approach to Sign Language Question Answering: We propose the first scalable, end-to-end framework for Sign Language Question Answering. This pipeline integrates three modular components: (1) gloss-based sign retrieval for understanding user queries, (2) a constrained Large Language Model (LLM) for reasoning and answer formulation in glosses, and (3) a 3D avatar-based generation module for synthesizing visual responses. This thesis contributes the Sign language Production module of the collaborative Gloss SLQA framework.

SignWeave-3D Architecture: We present a novel 3D framework that integrates Masked Language Models (MLM) with diffusion-based motion generation. This system retrieves high-fidelity 3D motion units (SMPL-X) from a dictionary and uses a diffusion model to "weave" them together, effectively generating natural co-articulation and filling gaps for unseen vocabulary.

1.5 Thesis Roadmap

The remainder of this thesis is organized as follows:

Chapter 2: Background and Literature Review provides a comprehensive overview of the current state of Sign Language Production (SLP). It details the linguistic properties of sign languages that pose unique challenges for computational modeling, such as non-manual markers and spatial grammar. This chapter also categorizes existing SLP approaches ranging from rule-based avatars to modern deep generative models and critically analyzes their limitations regarding scalability, realism, and data efficiency.

Chapter 3: Scalable 2D Sign Language Production addresses the trade-off between visual realism and data scarcity. It introduces SanketVaani-1K, a curated ISL dataset with fine-grained gloss boundaries, and presents a novel modular framework for 2D SLP. We detail the proposed pipeline, which synthesizes continuous, fluid signing by stitching isolated gloss units, demonstrating how high-quality video generation can be achieved without massive end-to-end training corpora.

Chapter 4: Gloss-SLQA: A Gloss-centric Approach to Sign Language Question Answering addresses the limitations of 2D sign synthesis by transitioning to a 3D avatar-based framework. The chapter introduces a gloss-centric approach that decouples motion from appearance, leveraging a Gloss–Motion Video Dictionary built on optimized SMPL-X parametric models to ensure structurally realistic sign production across multiple sign languages (ASL, ISL, and BSL). It details enhancements to the ExAvatar framework, incorporating joint-level and facial vertex offsets to capture the fine-grained manual and non-manual articulations required for linguistic clarity. Finally, the chapter presents Gloss-SLQA, a first-of-its-kind system enabling natural sign-language-based question answering through photorealistic 3D avatar response synthesis.

Chapter 5: SignWeave-3D: A Generative Framework for Co-articulation addresses the fluidity limitations identified in the previous chapter. This chapter proposes **SignWeave-3D**, a novel architecture integrating Masked Language Models (MLM) with diffusion-based motion generation. We describe the methodology for taking the retrieved high-fidelity SMPL-X motion units and "weaving" them via diffusion, effectively introducing natural co-articulation and smoothing to the generated sequences.

Chapter 6: Conclusion and Future Work summarizes the key contributions and findings of this research. It discusses the broader implications of scalable SLP for accessibility technology and outlines potential directions for future research, including real-time rendering improvements and the extension of these frameworks to other sign languages.

Chapter 2

Background

2.1 Sign Language

Sign languages are primary communication systems for Deaf communities that possess linguistic sophistication and societal importance comparable to spoken languages. They are fully fledged linguistic systems with their own unique grammar and vocabulary. According to the World Health Organization, approximately 72 million people worldwide use sign language as their primary means of communication.

Sign language is a visually rich and complex mode of communication. Its structure is fundamentally multimodal, relying on the asynchronous use of several channels to encode meaning:

Phonology: Similar to how spoken languages use meaningless sounds (phonemes) to build words, sign languages use meaningless visual units. These are defined by five core "parameters": Handshape, Orientation, Location, Movement, and Non-manual expression.

Morphology Syntax: Sign languages have independent rules for word formation and sentence structure. For example, American Sign Language (ASL) primarily uses a Topic-Comment structure, where the main subject is introduced first, often accompanied by non-manual markers like raised eyebrows.

Non-Manual Markers (NMMs): These are essential for grammar, acting like intonation in speech. For instance, furrowed eyebrows can signal a "WH-question" (who, what, where), while a head shake can negate an entire sentence.

Spatial Grammar: Sign languages incorporate spatial indexing, simultaneous signals, classifier constructions, and role shifts, which differ from the linear nature of spoken text or speech.

Referential Indexing and Loci When a signer introduces a person or object (a referent), they assign that referent to a specific point in the space in front of them, known as a locus.

Reference Tracking: Once a locus is established (e.g., "my sister" is placed on the right), the signer can refer back to her simply by pointing to that empty space. This replaces the need for repetitive pronouns like "she" or "her."

Verb Agreement: Many verbs in sign language are "directional." For example, the sign for GIVE moves from the locus of the giver to the locus of the receiver. If the sign moves from the right (sister) to the center (me), it translates to "She gave it to me" without needing separate words for "she" or "me."

Role Shifting Signers use a technique called Role Shift (or "Referential Shift") to report dialogue or actions. Instead of saying "He said...", the signer physically shifts their body or changes their eye gaze to inhabit the "locus" of that person.

Diversity and Regional Variation

There is no "universal" sign language; there are over 300 distinct sign languages worldwide.

Independence: Sign languages are not visual versions of spoken languages. For example, British Sign Language (BSL) and American Sign Language (ASL) are mutually unintelligible, despite both countries speaking English.

Dialects: Just like spoken accents, sign languages have regional variations. In the UK, signers in the North and South may use different signs for the same concept or possess distinct "accents" in their hand movements and mouthing.

These linguistic properties fundamentally shape how sign languages must be represented and processed computationally, motivating the need for specialized representations beyond linear text.

Representations in Sign Language Processing:

Due to the absence of a standardized written form, sign languages are represented in computational systems using a variety of abstractions, each capturing different linguistic and visual aspects. These representations trade off expressiveness, annotation cost, and computational tractability.

Gloss-based representations are the most widely used symbolic abstraction. Glosses label signs with uppercase tokens that approximate their lexical meaning and provide an interpretable intermediate layer between spoken language and signing. While glosses simplify alignment and learning, they under-specify spatial structure, non-manual markers, and discourse-level phenomena.

Pose-based representations encode signing as sequences of 2D or 3D joint coordinates, including hands, arms, face, and body. These representations preserve fine-grained motion and are well-suited for synthesis and animation but lack explicit linguistic structure and are sensitive to noise and viewpoint variation.

Video-based representations operate directly on raw RGB frames. They retain full visual fidelity, including subtle non-manual cues, but are high-dimensional, computationally expensive, and difficult to interpret or linguistically constrain.

Notation systems, such as SignWriting and HamNoSys, aim to explicitly encode phonological properties of signs, including handshape, orientation, movement, and location. Although linguistically expressive, these systems require specialized expertise and are rarely available at scale, limiting their adoption in data-driven models.

Hybrid and parametric representations combine symbolic and continuous information, for example by linking glosses to pose sequences, motion primitives, or parametric body models such as SMPL-X. These approaches seek to balance linguistic interpretability with physical realism and are increasingly used in sign language production systems.

Each representation captures complementary aspects of sign language. Consequently, modern sign language processing systems increasingly adopt multi-level or hybrid representations that integrate sym-

bolic meaning with continuous motion to achieve fluent, accurate, and scalable sign language generation.

2.2 Pose Estimation

2.2.1 2D Pose Estimation for Sign language

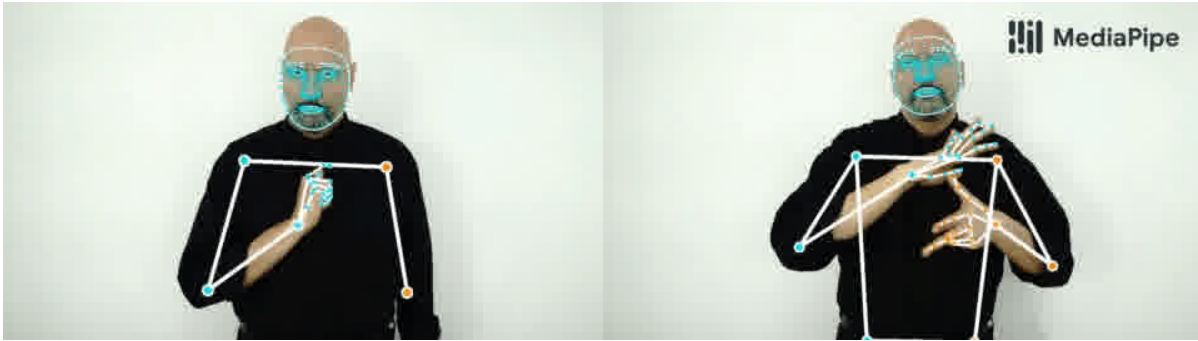


Figure 2.1 2D pose estimation. Example of Mediapipe Holistic Method

Sign language involves complex combinations of hand shapes, movements, body postures, and facial expressions. Pose estimation, the automatic detection of human skeletal keypoints, is crucial for interpreting these visual gestures. It provides a compact, invariant representation of a signer's movements, capturing essential kinematic cues from the hands, body, and face. Unlike raw video pixels, skeletal landmarks are unaffected by factors like clothing or background, making recognition models more robust [rivas.ai](#). In sign language, fine-grained pose detail is especially important: the meaning of a sign can hinge on subtle finger configurations or facial cues. For example, sign language understanding requires higher landmark granularity (including finger and facial keypoints) than general action recognition. Pose-based representations thus help isolate the what of the gesture from irrelevant visual noise, improving accuracy and even preserving user privacy in real-world systems. Overall, pose estimation serves as an information-rich intermediate, enabling sign language recognition (SLR) models or sign language production (SLP) to focus on meaningful motion patterns rather than raw video data.

OpenPose [40] is a pioneering 2D pose estimation framework known for real-time multi-person tracking. It uses a bottom-up approach with Part Affinity Fields to detect all keypoints and group them into individual skeletons.

MediaPipe Holistic [88] is another widely used 2D pose solution, particularly for real-time and mobile applications. It offers a holistic model that integrates body pose, hand, and face landmark detectors into one pipeline. In total it outputs 543 landmarks (33 for the body pose, 21 for each hand, and 468 for the face), providing an extremely detailed representation well-suited to sign language analysis. Medi-

aPipe uses a hybrid of neural network heatmaps and regression to predict keypoints, and it's optimized for speed on devices.

HRNet [99] High Resolution Net (HRNet) represents a state-of-the-art top-down pose estimation architecture that has also been adapted in sign language research. HRNet maintains high-resolution feature maps throughout the network, enabling precise localization of small keypoints (like fingers). This improved accuracy is particularly valuable for sign language, where even slight misalignments of hand pose can alter the meaning. HRNet-based models (such as the open-source WholePose from MMPose) have been evaluated on sign videos with promising results. A recent study compared an HRNet-based whole-body pose model against OpenPose and MediaPipe on multiple sign language datasets. The HRNet variant produced fewer missing or mis-detected keypoints, especially for hands, leading to better SLR performance.

2.2.2 3D Pose Estimation for Sign Language

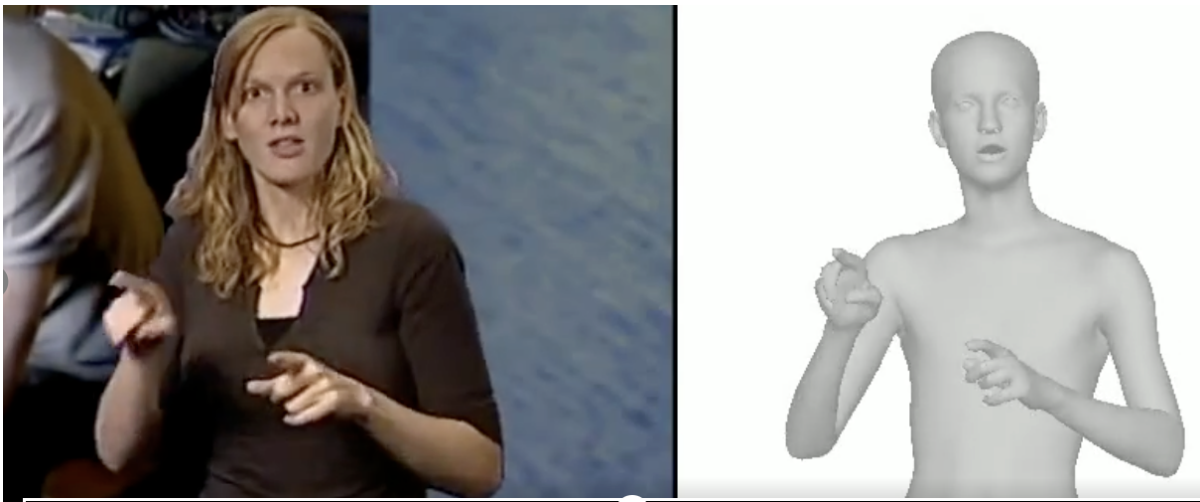


Figure 2.2 3D pose estimation. Example: SMPL-X representation from Ex-avatar SMPL-X registration method.

There is growing interest in 3D pose estimation to capture the full spatial information of signing. Three-dimensional human pose can provide depth and view-invariant features, potentially improving production of signs that involve forward/backward motions or complex hand orientations. Early sign language research often obtained 3D joint data through specialized hardware: for example, the Microsoft Kinect sensor can directly output 3D skeletons, which some works used to recognize isolated signs and alphabet gestures. These depth-camera approaches demonstrated the value of 3D data but required controlled setups and hardware not always available to end-users. In recent years, researchers have explored inferring 3D pose from regular video using advanced models. One notable approach is to fit

a parametric 3D human model to signers in video. SMPL-X [61] is a comprehensive 3D model of the human body, hands, and face.

Another advantage of 3D pose is its ability to model occlusions and view changes. In real-world settings, a signer might not always face the camera directly, 3D pose estimation can theoretically handle varying camera angles better by maintaining a consistent body-centered coordinate frame. Moreover, the inclusion of non-manual signals (facial expressions, mouth shapes) becomes more seamless in 3D models like SMPL-X, which encode face pose parameters along with body joints. This is important because certain signs are distinguished only by facial cues (e.g. a question vs. statement).

2.3 SMPL-X

Accurate modeling of upper-body kinematics, hand articulation, and facial expressions is essential for computational sign language production, as sign languages rely on the coordinated interaction of multiple expressive channels. SMPL-X (Skinned Multi-Person Linear Model – eXtended) is a parametric 3D human body model that provides a unified and expressive representation of the human body, including the hands and face. Introduced by Pavlakos et al. in 2019, SMPL-X builds upon earlier body models such as SMPL and MANO, extending their capabilities by incorporating detailed hand articulation (via 15 degrees of freedom per hand) and expressive facial modeling (via 200 blendshape coefficients) [60]. The model represents human pose using a set of parameters: body shape (via low-dimensional PCA components), skeletal pose (joint rotations), hand poses, facial expressions, and global translation. These parameters control a skinned mesh of 10,475 vertices, producing realistic full-body geometry suitable for animation, rendering, or learning-based tasks.

SMPL-X describes the human body using a compact parameter vector

$$(\beta, \theta, \psi, \mathbf{t}),$$

where β denotes PCA shape coefficients, θ encodes body, hand, and facial joint rotations, ψ represents facial expression coefficients, and $\mathbf{t}$ provides global translation. These parameters control a differentiable mesh-generation function

$$M(\beta, \theta, \psi) : \mathbb{R}^{|\beta|} \times \mathbb{R}^{|\theta|} \times \mathbb{R}^{|\psi|} \rightarrow \mathbb{R}^{3N},$$

which produces a full 3D surface suitable for synthesis of expressive human motion.

SMPL-X generates a posed template mesh through a combination of learned deformation spaces:

$$T_p(\beta, \theta, \psi) = \bar{T} + B_S(\beta) + B_P(\theta) + B_E(\psi), \quad (2.1)$$

where $\bar{T}$ is the canonical template mesh, $B_S(\beta)$ is the shape blend-shape function, $B_P(\theta)$ is the pose-corrective blend-shape function, and $B_E(\psi)$ is the expression blend-shape function.

Shape deformation is modeled as:

$$B_S(\beta) = \sum_{n=1}^{|\beta|} \beta_n S_n, \quad (2.2)$$

where S_n are PCA components learned from 3D scans.

Pose deformation uses rotation-dependent correctives:

$$B_P(\theta) = \sum_{n=1}^{9K} (R_n(\theta) - R_n(\theta^*)) P_n, \quad (2.3)$$

where $R_n(\cdot)$ extracts part-relative rotation matrices using the Rodrigues formula, and P_n are pose-corrective displacement vectors.

Expression deformation is given by:

$$B_E(\psi) = \sum_{n=1}^{|\psi|} \psi_n E_n, \quad (2.4)$$

where E_n are PCA expression components derived from the FLAME model.

The articulated mesh is produced using Linear Blend Skinning (LBS):

$$M(\beta, \theta, \psi) = W(T_p(\beta, \theta, \psi), J(\beta), \theta, \mathbf{W}), \quad (2.5)$$

where $J(\beta)$ regresses the joint locations using a sparse linear regressor applied to the deformed mesh, and $\mathbf{W}$ denotes the skinning weight matrix.

This formulation yields a mesh that is fully differentiable with respect to all parameters, enabling learning-based optimization, gradient-based fitting, and direct regression from image, video, or motion data.

2.4 SMPL-X Registration

3D human mesh recovery methods aim to estimate the parameters of a parametric human body model. Key approaches can be broadly categorized as follows:

Iterative optimization-based methods: These solve an optimization problem per input (often using 2D keypoint detections as targets) to fit the model. Example: SMPLify-X (2019) fits the SMPL-X model by minimizing reprojection error to 2D joints with regularization terms.

Regression-based methods: These use neural networks to directly regress model parameters from images (often trained on datasets with 3D ground truth or using self-supervision). Example: SMPLer-X (2023) employs a learned model to predict full SMPL-X parameters from an image (for improved speed and scalability).

Hybrid or multi-stage methods: These combine learned components with analytical steps or multiple sub-networks. For instance, some pipelines first estimate body pose then refine hand and face pose separately, or use an initial regressor followed by iterative refinement.

SMPLify-X [61] introduced one of the first practical optimization-based pipelines for estimating a full SMPL-X body mesh, including hands, face, and body, from a single RGB image. The method begins by detecting 2D keypoints for the entire body (body joints, hand landmarks, and facial keypoints) using OpenPose, and then optimizes the parameters of the SMPL-X model so that its projected 3D joints align with these 2D detections.

The optimization objective combines: Joint reprojection loss to match predicted mesh joints with 2D detections. Pose and shape regularizers, most importantly the VPoser prior, which promotes realistic full-body configurations. Interpenetration penalties, which discourage physically impossible body-part collisions.

Together, these constraints allow the algorithm to produce anatomically plausible, full-body 3D reconstructions even under noisy or partial 2D detections. Its main strengths are generalization without requiring 3D supervision and the ability to produce a complete, consistent mesh. However, being an iterative optimizer, SMPLify-X is relatively slow (seconds per image) and highly sensitive to the quality of 2D keypoints. Ambiguities in depth or occluded joints can still cause unrealistic or flipped poses.

SMPLer-X [8] advances monocular human mesh recovery by replacing per-image optimization with a large-scale, feed-forward regression framework for the SMPL-X model. The method employs a deep neural architecture that jointly predicts body pose, shape, facial expression, and hand articulation from a single RGB image, enabling expressive full-body reconstruction in a single inference step. A component-aware design allows the body, hands, and face to be encoded with part-specific attention while maintaining global coherence, thereby capturing correlations across expressive regions that are critical for fine-grained gesture understanding.

Leveraging extensive 2D and 3D datasets, including high-quality pseudo-ground-truth fits SMPLer-X demonstrates strong generalization and state-of-the-art accuracy on contemporary benchmarks such as AGORA and 3DPW. By utilizing full-image cues (texture, shading, silhouette, and context), it resolves depth and articulation ambiguities more effectively than approaches constrained only by sparse keypoints. Its feed-forward formulation also offers substantial computational advantages, supporting near real-time performance and scalable processing of long video sequences.

The primary limitation of SMPLer-X lies in its dependence on the distribution and diversity of its training data; unfamiliar clothing, extreme poses, or out-of-domain body shapes may require adaptation or fine-tuning.

ExAvatar [55] performs accurate whole-body co-registration of the SMPL-X model to short monocular videos prior to learning the Gaussian avatar representation. Initially, off-the-shelf regressors are used to estimate the SMPL-X pose, shape, and facial expression parameters for all frames. These initial estimates are then refined through a 2D keypoint-based optimization process, in which the shape parameter is shared across all frames, while the pose, global translation, and expression parameters

remain frame-specific. To further improve whole-body alignment, ExAvatar introduces two identity-dependent offsets in the SMPL-X T-pose. The first is a joint offset $\Delta\mathbf{J}$, which corrects skeletal misalignment, particularly in the hand regions. The second is a face vertex offset $\Delta\mathbf{V}_{\text{face}}$, obtained by registering a high-fidelity FLAME face model to detected 2D facial landmarks. The final optimization minimizes a weighted combination of several loss terms, including the 2D reprojection loss, face-mesh alignment losses, Laplacian smoothness, shape regularization, joint-offset regularization, and symmetry constraints. This optimization yields a tightly aligned, person-specific SMPL-X mesh that directly drives the expressive Gaussian avatar. Neural Rendering Methods for Avatar Synthesis

2.5 Rendering Engines and Techniques for Avatar Synthesis

Realistic human avatar synthesis relies critically on the choice of rendering engine and scene representation. Over the last decade, two major classes of rendering paradigms have shaped the field: traditional graphics-based rendering engines and neural rendering methods. Between 2022 and 2025, significant advances in neural representations, particularly Neural Radiance Fields (NeRFs) and 3D Gaussian Splatting, have enabled photo-realistic, animatable avatars learned directly from video data, while traditional engines remain indispensable for visualization, animation, and dataset generation.

Traditional Rendering Engines

Conventional graphics engines such as Blender, Unreal Engine, and lightweight visualization libraries like Open3D form the backbone of synthetic data generation and avatar demonstration pipelines. Blender’s physically-based Cycles renderer is extensively used for creating large-scale synthetic human datasets with accurate ground-truth annotations for pose, depth, and segmentation. Its real-time Eevee engine provides efficient preview and stylized animation capabilities. Blender is also widely adopted for visualizing reconstructed meshes and animating avatars using motion capture or parametric body models such as SMPL-X.

Unreal Engine offers real-time, high-fidelity rendering with physically-based lighting and complex scene simulation. It has been used extensively for photorealistic dataset generation (e.g., virtual human datasets with controlled lighting, clothing, and environments) and for real-time interactive avatar visualization in AR/VR and gaming contexts. Unreal supports skeletal animation pipelines, allowing parametric human models and neural avatar outputs to be deployed for real-time demonstrations.

Open3D [114], in contrast, is primarily used for lightweight visualization and geometric debugging. While it does not aim for cinematic realism, it provides GPU-accelerated rendering of point clouds, meshes, and depth maps, making it particularly useful for verifying mesh alignment, surface reconstruction, and parametric model fitting during research and development.

Although these traditional engines deliver high visual quality and real-time performance, they are inherently non-differentiable, which limits their direct use in neural optimization pipelines.

Neural Rendering for Avatars

Neural rendering has emerged as a powerful paradigm that enables learning geometry and appearance directly from images through differentiable rendering. A central representation in this class is the Neural Radiance Field (NeRF) [51], which models a scene as a continuous volumetric function mapping 3D coordinates and viewing directions to color and density. Extensions of NeRF for dynamic humans incorporate skeletal conditioning and deformation fields, enabling animatable volumetric avatars with photorealistic detail. While NeRF-based avatars achieve high fidelity, their reliance on volumetric ray marching makes real-time rendering computationally expensive.

A major breakthrough toward real-time neural avatar rendering came with 3D Gaussian Splatting (3DGS) [39]. Instead of volumetric grids, 3DGS represents a scene using anisotropic Gaussian primitives with learnable position, orientation, scale, color, and opacity. These Gaussians are rasterized using a visibility-aware splatting algorithm that bypasses empty-space computation and enables highly efficient GPU rendering. 3DGS achieves real-time novel-view synthesis at high resolution while remaining fully differentiable, making it an ideal representation for animatable avatars learned from monocular or multi-view video.

Differentiable Rasterization

Bridging traditional mesh rendering and neural training, differentiable rasterizers such as PyTorch3D and NVDiffrast allow gradients to flow from image-space losses back to 3D geometry and textures. These tools enable inverse rendering tasks such as parametric body fitting, mesh refinement, and texture optimization directly from RGB supervision. Unlike volumetric methods, rasterization operates on explicit surfaces and is highly efficient for moderate mesh resolutions, particularly when combined with parametric human models like SMPL or SMPL-X.

2.6 SMPL-X–Driven Avatar Generation and Animation

X-Avatar (CVPR 2023) introduces a unified implicit avatar model that captures full-body, facial, and hand expressiveness using SMPL-X–driven deformation fields. The method learns a neural shape and deformation field from either multi-view RGB-D frames or static 3D scans. This field is animated by SMPL-X pose inputs, enabling compatibility with standard parametric models while achieving higher realism. To effectively model the high-resolution deformations required by small and expressive regions (like fingers and facial features), X-Avatar proposes a part-aware learning framework. Each body part (e.g., torso, hands, face) has its own deformation MLP, and a part-aware skinning function maps SMPL-X joint poses to local deformations. An appearance network further refines rendering by predicting view-dependent colors, conditioned on pose, expression, and local surface features. This design allows X-Avatar to synthesize photorealistic avatars under novel poses and expressions. It is the first implicit model to jointly model body, hands, and face with high fidelity in a fully SMPL-X–driven pipeline.

3DGS-Avatar [66] introduces a deformable 3D Gaussian splatting framework for animatable avatars, built from short monocular videos. The subject is first fitted with an initial SMPL skeleton and represented as a cloud of 3D Gaussians in a canonical pose. A deformation network then maps canonical

Gaussians to posed ones, enabling animation via SMPL poses. To ensure generalization to unseen motions, the method uses near-isometric regularization on the deformation field. Though it does not model face or hands explicitly, 3DGS-Avatar achieves real-time rendering (50+ FPS) and 30-minute training on a single GPU, over 400× faster than NeRF-based avatars, with comparable or superior visual quality.

GaussianAvatar [31] proposes an explicit 3D Gaussian representation anchored to a parametric body mesh (SMPL or SMPL-X) to create photorealistic, animatable avatars from a single monocular video. It jointly optimizes the motion parameters (SMPL-X pose per frame) and a dynamic appearance network, which learns pose-conditioned color changes using an optimizable 3D feature grid. This joint optimization improves alignment even when initial poses are noisy. GaussianAvatar supports novel-pose rendering with high appearance fidelity and maintains 3D-consistent textures across body motions. Unlike 3DGS-Avatar, it supports SMPL-X skeletons and refines pose estimates during training.

ExAvatar proposes a hybrid avatar model combining a parametric SMPL-X mesh (for pose, shape, and expression control) with a 3D Gaussian splatting (3DGS)-based appearance layer. From a short monocular video, it learns a mesh-anchored Gaussian cloud that preserves spatial coherence under novel poses. The Gaussians are tightly coupled to the SMPL-X mesh via vertex bindings, maintaining geometric consistency and reducing artifacts even under unseen facial or hand motions. ExAvatar supports full-body expressiveness with photorealistic rendering, driven by arbitrary SMPL-X inputs.

2.7 Sign Language Avatars with SMPL-X

Sign language communication demands full-body expressiveness, especially detailed hand shapes and facial expressions. SMPL-X-based models are increasingly used to generate or reconstruct signing avatars, enabling translations between spoken language and sign.

SGNify: [23] introduce a method for reconstructing expressive 3D signing avatars from monocular in-the-wild videos by explicitly incorporating linguistic priors from sign language into the reconstruction pipeline. Their approach, builds upon the SMPL-X parametric human model, integrates off-the-shelf pose and face estimators with a refinement stage guided by phonological properties of sign languages, such as two-hand symmetry and hand-pose consistency over a sign unit to resolve ambiguities caused by occlusions, motion blur, and self-contact in real signing videos. By explicitly embedding linguistic structure into the optimization, the method produces significantly more accurate and perceptually realistic signing avatars than prior generic human reconstruction pipelines. This work represents one of the first attempts to bridge 3D human reconstruction with sign-language linguistics, establishing a critical foundation for realistic sign-language avatar generation and downstream sign-language production systems.

In **SMPLSign-X** [119], the authors introduce a sign-language-specific SMPL-X optimization framework to accurately recover 3D signing motion from monocular sign videos. Building upon the standard SMPLify-X fitting pipeline, they incorporate domain-aware regularization strategies tailored to signing. Since many joints—particularly the lower body—are often unobserved in sign videos, an unseen-joint

regularization term constrains such joints toward a rest pose based on 2D keypoint confidence. To reflect the characteristic posture of signers, an upright-body prior is imposed on the torso and pelvis to suppress unnatural body tilts. Furthermore, a strong temporal smoothness constraint is applied across consecutive frames to eliminate frame-wise jitter and ensure fluid sign articulation. These additions are optimized jointly with the standard 2D reprojection, pose prior, and mesh regularization losses, enabling the construction of a high-fidelity 3D sign dictionary of SMPL-X motion sequences

In **Neural Sign Actors** [6], high-quality SMPL-X fitting is used to convert continuous sign-language videos into precise 3D supervision for training the diffusion model. The authors employ a hybrid regression-and-optimization pipeline, where initial SMPL-X parameters, are first predicted using an off-the-shelf human motion estimator and then refined through 2D reprojection-based optimization. The refinement minimizes the discrepancy between projected 3D joints and detected 2D keypoints for the body, hands, and face, while enforcing strong anatomical pose priors to maintain physically plausible articulation. A temporal consistency regularizer is further applied across consecutive frames to suppress jitter and ensure smooth signing motion.

In **SignAvatars** [106], the authors develop a fully automated large-scale SMPL-X fitting pipeline to generate high-quality 3D sign-language supervision from raw videos. Each clip is initialized using existing whole-body 3D pose estimators to obtain coarse body, hand, and facial parameters, which are then refined through multi-stage optimization by minimizing 2D reprojection error of body, hand, and facial keypoints together with standard SMPL-X pose and shape priors. To ensure biomechanical plausibility and signing fidelity, the optimization enforces hand-specific kinematic constraints, joint-angle limits, and strong temporal smoothness to suppress jitter in fast hand motions. The body shape is optimized once per signer and kept identity-consistent across all frames, while pose and expression are optimized per frame. This pipeline enables the construction of a massive dataset of temporally coherent, anatomically valid SMPL-X motion sequences for sign language, which forms the foundation for large-scale 3D sign-language generation and benchmarking.

Chapter 3

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

3.1 Introduction

In this chapter, we present a novel and scalable framework for 2D Sign Language Production (SLP) that addresses the trade-off between visual realism and data scarcity. Our approach adopts a retrieval-based generation paradigm, in which high-quality, isolated gloss units are dynamically selected and composed into continuous signing sequences. To enable smooth transitions between retrieved units, we repurpose FramePack, originally proposed for long-range text-to-motion generation, as a motion interpolation module, adapting it to synthesize temporally coherent transitions between adjacent sign segments. To support this framework, we introduce **SanketVaani-1K**, a carefully curated dataset of 1,000 Indian Sign Language (ISL) sentences annotated with fine-grained gloss boundaries. By combining context-aware gloss retrieval using a vector database with diffusion-based interpolation for seamless stitching, our system generates coherent, photorealistic sign language videos without requiring large-scale end-to-end text-to-video training.

3.2 Literature Review

Initial SLP systems relied on rule-based or motion-capture-driven avatars using notations such as HamNoSys/SiGML, offering explicit linguistic control but often producing rigid or unnatural motion [35, 5, 68].

Data-driven pipelines typically decompose SLP into Text-to-Gloss, Gloss-to-Pose, and Pose-to-Sign stages [56, 96]. While modular, these approaches often suffer from poor inter-sign transitions, low-resolution synthesis, and inadequate modeling of non-manual cues. Generative Adversarial Networks (GANs) have been applied to sign language production [77, 96] to generate realistic signing videos from pose or gloss input. While models like SignGAN [77] and FSNet [80] 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. Recent

diffusion-based approaches have explored direct conversion from spoken language to sign language motion. Methods such as SignDiff [21] and text-driven 3D diffusion models [91] bypass intermediate representations to generate temporally coherent sign motions directly from text or pose conditions; however, such direct generation remains difficult to scale due to costly motion supervision and the lack of large-scale, finely aligned gloss and motion annotations.

Several datasets support sign language modeling, including RWTH-PHOENIX-Weather 2014T [9], BSL-1K [2], and How2Sign [19]. However, these datasets are primarily designed for recognition and lack fine-grained gloss-type annotations, motion segmentation, or compositional metadata needed for controlled SLP. In the ISL domain, iSign [37] and ISLTranslate [36] provide sentence-level supervision but omit temporal alignment and motion-level annotations, limiting their suitability for generation.

3.3 Proposed Scalable 2D Sign Language Production

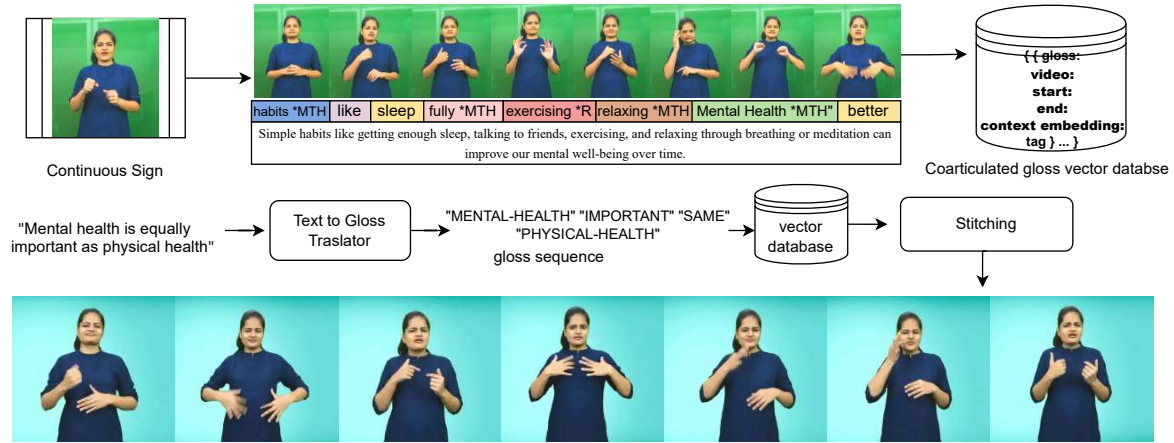


Figure 3.1 Stitching Pipeline: includes Text to gloss Translator, formation of Vector database, and video retrieval, and sign stitching with interpolation.

Our proposed framework adopts a modular, retrieval-based approach to synthesise continuous sign language videos. The pipeline consists of three main components: (1) A curated dataset and vector database construction; (2) A Text-to-Gloss translator; and (3) A stitching module utilizing FramePack interpolation.

3.3.1 Vector Database Construction

To facilitate semantically aware sign language synthesis, we establish a vector database of gloss-level embeddings designed to encapsulate both lexical identity and contextual semantics. Our approach employs a pre-trained contextual language model, specifically BERT [1], to generate robust token-level embeddings from gloss sequences.

Formally, for a given sequence of glosses representing a sentence, denoted as:

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

we calculate a contextualized embedding $\mathbf{e}_i$ for each token g_i according to:

$$\mathbf{e}_i = f(g_i \mid g_{i-k}, \dots, g_i, \dots, g_{i+k}), \quad (3.1)$$

where $f(\cdot)$ represents the encoder function and k denotes the size of the context window. These embeddings encode the meaning of a gloss relative to its neighbors, which is essential for disambiguating polysemous signs based on their intended context.

We store each resulting embedding $\mathbf{e}_i \in \mathbb{R}^d$ in a vector database, such as FAISS [17], indexing it alongside the following essential metadata:

- **Gloss Attributes:** The gloss label and specific sign type (e.g., *P for pointing, *MTH for mouthing).
- **Contextual Data:** The surrounding gloss context to ensure local coherence.
- **Motion Reference:** A pointer to the specific video segment corresponding to the sign animation.

During the inference phase, for a specific query gloss g_q , we generate its contextual embedding $\mathbf{e}_q$ and execute a nearest-neighbor search within the database:

$$\hat{s} = \arg \min_{s_j \in \mathcal{D}} \text{dist}(\mathbf{e}_q, \mathbf{e}_j), \quad (3.2)$$

where $\text{dist}(\cdot, \cdot)$ represents the cosine distance, and $\mathcal{D}$ denotes the repository of stored embeddings linked to motion segments. To further refine the selection, we utilize metadata such as sign type for ranking. This retrieval mechanism ensures that the selected motion units are not only semantically accurate but also contextually congruent for the subsequent stitching process.

By effectively capturing fine-grained, contextual gloss meanings and associating them with expressive motion data, this database functions as the semantic foundation of our retrieval-based production framework.

3.3.2 Text to Gloss Sequence

Historically, systems converting English to Indian Sign Language (ISL) relied heavily on handcrafted transfer rules to reorder source text into gloss sequences. While the scarcity of formal documentation on ISL grammar complicates this task, the fundamental limitation lies in the rigidity of rule-based architectures; static conditional logic cannot adequately encompass the stochastic variability of natural language. Consequently, novel syntactic constructions often necessitate manual, ad hoc rule adjustments, while idiomatic or polysemous expressions are frequently mistranslated. When applied to un-

constrained, real-world text, these systems typically yield gloss outputs that lack syntactic consistency and semantic completeness.

Contemporary research identifies Large Language Models (LLMs) as a robust, context-aware alternative that consistently outperforms rigid rule-based frameworks. For instance, [28] utilise few-shot prompting to generate *pseudo-gloss* labels from spoken text, thereby eliminating the need for expensive manual annotation and enhancing downstream sign language translation on benchmarks like PHOENIX14-T and CSL-Daily. Subsequent studies have further demonstrated that zero-shot prompting in dialogue systems [33], the generation of glosses alongside non-manual cues [110], and stepwise gloss induction [4] allow LLMs to model sign language grammar with greater fidelity than traditional rule sets.

LLM-based ISL Gloss Generation. In our approach, we condition the LLM using concise *ISL rule* that encapsulates the core grammatical constraints identified in [25], such as *TOPIC-COMMENT* word order, article omission, sentence-final WH-words, and clause-initial time adverbials as mentioned in [?].

ISL Gloss generation Prompt:

```
You are an Indian Sign Language (ISL) \
translation assistant.\
Your task is to convert English sentences \
into ISL gloss format following the rules below:\
=== Conversion Rules ===\
1. Grammar Transfer:\
  - Use Subject{Object{Verb (SOV) order.\
  - Nouns come before verbs.\
    - "He eats mangoes" → HE MANGO EAT\
  - Adjectives follow the noun.\
    - "blue book" → BOOK BLUE\
  - Numbers follow the noun they describe.\
    - "three sons" → SON THREE\
  - Auxiliary verbs come before the main verb.\
    - "was eating" → WAS EAT\
  - Adverbs come after the verb.\
    - "ran quickly" → RUN QUICKLY\
  - Relative clauses come after the noun.\
    - "the book which you bought" → BOOK YOU BUY\

2. Interrogatives:\
  - Convert to statement + question word at the end.\
    - "Where are you going?" → YOU GO WHERE\

3. Negatives:\
```

- Place NOT at the end.\
 - "He is not a doctor." → HE DOCTOR NOT\
- For negative questions: place NOT \ before the question word.\
 - "Who will not go with me?" → I WITH GO NOT WHO\
- 4. Multi-Word Expressions (MWEs):\
 - Detect and merge expressions like "thank you",\ "look up to" into single tokens:\
 - THANK_YOU, LOOK_UP_TO\
- 5. Stopword Removal and Lemmatization:\
 - Remove articles and unnecessary \ functional words (e.g., "is", "the").\
 - Convert all remaining words to their root form.\

#	English Sentence	ISL Gloss
1	Blood test is available tomorrow morning for thyroid check-up.	TOMORROW MORNING BLOOD_TEST AVAILABLE THYROID CHECK_UP
2	Vaccines protect children from many diseases.	VACCINE CHILDREN MANY DISEASE PROTECT
3	Doctor gives hand sanitizer to patient.	DOCTOR HAND_SANITIZER PATIENT GIVE
4	Bank approves home loan papers.	HOME_LOAN PAPER BANK APPROVE
5	Customer follows loan repayment schedule.	LOAN REPAYMENT SCHEDULE CUSTOMER FOLLOW
6	Check account balance at ATM.	ACCOUNT BALANCE ATM CHECK
7	Police put road block for safety.	POLICE ROAD_BLOCK SAFETY PUT
8	Government police Sikkim road block safety put.	GOVERNMENT POLICE SIKKIM ROAD_BLOCK SAFETY PUT
9	Hospital gives masks for safety.	MASK HOSPITAL SAFETY GIVE
10	Bus goes to tourist place in Sikkim.	BUS SIKKIM TOURIST_PLACE GO
11	Tourists see river near Sikkim.	RIVER NEAR SIKKIM TOURIST SEE
12	Hotel rooms available in Sikkim.	HOTEL ROOM SIKKIM AVAILABLE
13	Add sugar to tea in kitchen.	KITCHEN SUGAR TEA PUT
14	Put vegetables in soup pot.	VEGETABLE SOUP POT PUT
15	Add salt to curry to balance taste.	SALT CURRY BALANCE PUT

Table 3.1 Example English sentences with corresponding ISL gloss sequences.

3.3.3 Stitching

The final phase of our pipeline is the seamless integration of retrieved motion segments into a temporally coherent signing sequence. A naive concatenation of isolated gloss units inevitably introduces temporal discontinuities, abrupt transitions, and motion artifacts at segment boundaries, disrupting the semantic flow and intelligibility of the sign language output.

Conventional approaches to motion interpolation often rely on pose-space linearity, interpolating joint angles or latent vectors within frameworks like SMPL or MediaPipe. While computationally efficient, these methods fail to model the non-linear kinematics characteristic of natural human motion,

Dataset	Lang.	# Sent.	Vocab.		Domain	Gloss		Controlled
			(Text)			Annot.	Seg.	Env.
Purdue RVL [50]	ASL	2.5k	104	General	✓	✗	✓	✓
Boston 104 [18]	ASL	201	103	General	✓	✗	✓	✓
How2Sign [19]	ASL	35k	16k	Instructional	✓	✗	✓	✓
OpenASL [86]	ASL	98k	33k	Web	✗	✗	✗	✗
YT-ASL [94]	ASL	610k	60k	Web	✗	✗	✗	✗
BOBSL [3]	BSL	993k	72k	TV	✗	✗	✗	✗
BOBSL-CSLR [70]	BSL	5k	5k	TV	✓	✓	✗	✗
CSL Daily [113]	CSL	20.6k	2k	General	✗	✗	✓	✓
Phoenix14T [9]	DGS	8.2k	3k	Weather	✓	✗	✗	✗
SWISSTXT-W [34]	DSGS	811	1k	Weather	✓	✗	✗	✗
SWISSTXT-N [11]	DSGS	6k	10k	News	✓	✗	✗	✗
KETI [41]	KSL	14.6k	419	Emergency	✓	✗	✓	✓
VRT-News [11]	VGT	7.1k	7k	News	✗	✗	✗	✗
ISL-CSLRT [69]	ISL	100	–	General	✗	✗	✓	✓
ISLTranslate [36]	ISL	31k	11k	Education	✗	✗	✗	✗
iSign [37]	ISL	118k	40k	General	✓	✗	✗	✗
SANKETVAANI-1K (Ours)	ISL	1k	6k	~40 cat.	✓	✓	✓	✓

Table 3.2 Comparison of sign language datasets. The *Controlled Env.* column indicates whether the dataset was recorded in a controlled environment (✓) or curated from existing sources (✗).

frequently resulting in “robotic” or physically implausible transitions, particularly in expressive regions involving manual and non-manual cues. Furthermore, generative alternatives such as GANs or First Order Motion Models (FOMM) are prone to warping artifacts and temporal inconsistencies, while uni-directional diffusion models often suffer from “temporal drift” and attention bottlenecks that degrade long-range coherence.

To address these limitations, we employ **FramePack**, a state-of-the-art interpolation framework originally designed for long-video next-frame prediction. FramePack is distinguished by its ability to geometrically compress context, allowing the model to attend to temporally distant anchor frames without incurring quadratic computational costs. Critically, it utilizes a bi-directional “anti-drifting” sampling strategy. By locking high-fidelity start and end frames as boundary constraints, the model generates the intermediate span from both temporal directions simultaneously, thereby eliminating the drift common in autoregressive approaches.

In our implementation, we define the interpolation anchors as the final frame of gloss segment i and the initial frame of gloss segment $i + 1$. FramePack preserves these anchors at full resolution and synthesizes the missing intermediate frames, effectively bridging the semantic units. The system operates atop a frozen Hunyuan-DiT backbone, a 13-billion parameter diffusion transformer, allowing us to leverage robust text-to-motion alignment while ensuring drift-free, photorealistic transitions. The result is a fluid, continuous signing stream that maintains the visual fidelity of the original recordings while enforcing smooth kinematic transitions between disparate glosses.

3.4 Dataset: SanketVaani-1K

We introduce **SanketVaani-1K**, a rigorously curated Indian Sign Language (ISL) dataset comprising 1,000 spoken sentences. The dataset is designed to address the critical communication requirements of the Deaf community in India, spanning over 40 high-utility domains such as Banking, Healthcare, Legal Rights, Education, and Government Services. This broad categorical coverage ensures that the dataset captures a wide spectrum of everyday contexts relevant to Deaf individuals. Figure 3.2 illustrates the distribution of sentences across these domains, highlighting the dataset’s thematic diversity.

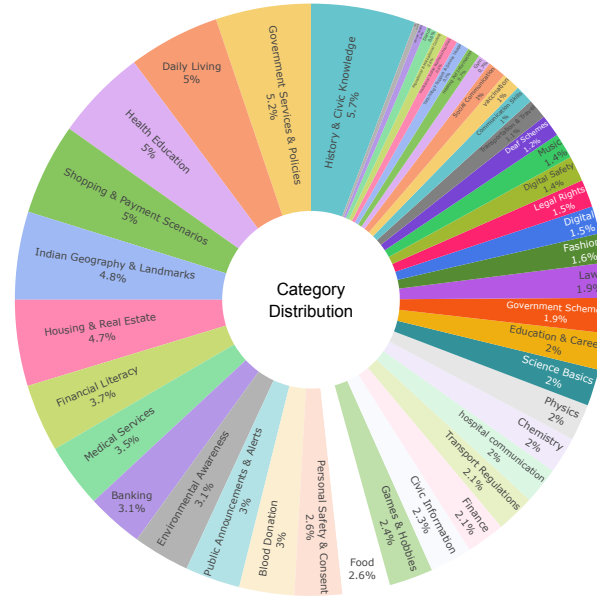


Figure 3.2 Comprehensive category-wise distribution of the ISL dataset, demonstrating coverage across over 40 high-utility domains.

To ensure the dataset’s practical utility, we selected domains based on their relevance to the Indian socio-cultural landscape. Subdomains were further defined to capture fine-grained real-world variations; for instance, the Healthcare domain encompasses specific topics like vaccination schedules and emergency protocols. We employed ChatGPT to generate sentence prompts that are linguistically accessible, culturally grounded, and contextually rich. A certified ISL interpreter and a member of deaf community translated these prompts into ISL, ensuring the preservation of grammatical structures and essential non-manual cues. Structurally, the majority of sentences in the dataset contain between 10 and 30 words, as shown in the word count distribution in Figure 3.3.

Recording was conducted using a single front-facing GoPro 10 camera operating at 1920×1080 resolution and 30 fps. The camera was tripod-mounted and centered on the signer to provide a consistent frontal view, capturing the head, torso, and hands clearly. This setup ensures that both manual signs and non-manual signals are recorded with high fidelity, maintaining visual consistency across all sessions.

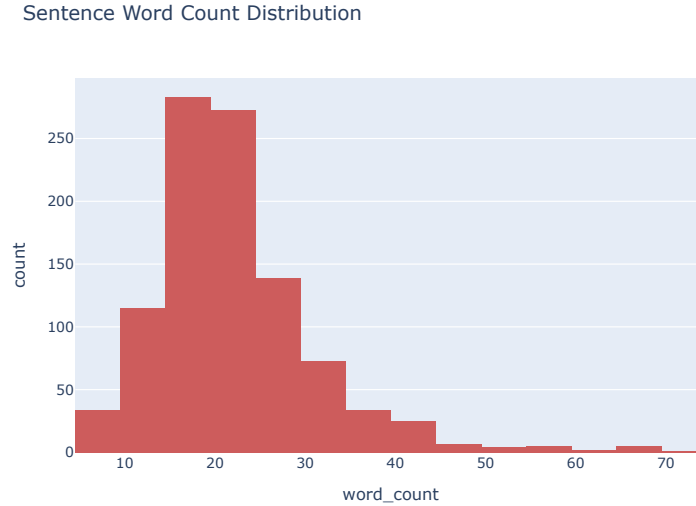


Figure 3.3 Distribution of sentence word counts in the ISL dataset. The majority of sentences fall within the 10–30 word range.

3.4.1 Data Pre-processing

3.4.1.1 Filtering and Synchronization

To facilitate precise segmentation of sign tokens, distinct acoustic cues (claps) were inserted at the beginning and end of each sentence during recording. The audio track was down-mixed to mono PCM (44.1 kHz) and normalized. We applied a Hilbert transform to extract the audio envelope; peaks exceeding 0.7 a.u. separated by at least 0.5 seconds were identified as clap events. These events served as timestamps to automatically segment the raw footage into individual clips. A subsequent manual verification step was performed to filter out noisy or invalid clips, and the retained files were renamed according to their sentence IDs (e.g., S00001, S00002).

3.4.1.2 Signer Normalization

We employed a YOLO v9 object detector [53] to localize the signer in each frame. Bounding boxes were linked temporally using ByteTrack to ensure smooth tracking. To prevent accidental cropping of articulators (hands or face), the bounding boxes were expanded by 10%. The cropped signer regions were then centered on a Full-HD canvas with a neutral green-grey background, standardizing the signer’s scale and position across the entire dataset

3.4.1.3 Background Replacement

For each normalized clip, we generated a foreground mask from the initial frame using ISNet [67]. The video was processed in 60-frame batches using MatAnyone [103] to generate high-quality per-pixel alpha maps and clean foreground RGB frames. These alpha maps allowed us to seamlessly composite the signer onto a uniform background, preserving the original resolution and frame rate while eliminating visual distractions.

3.4.2 Annotation

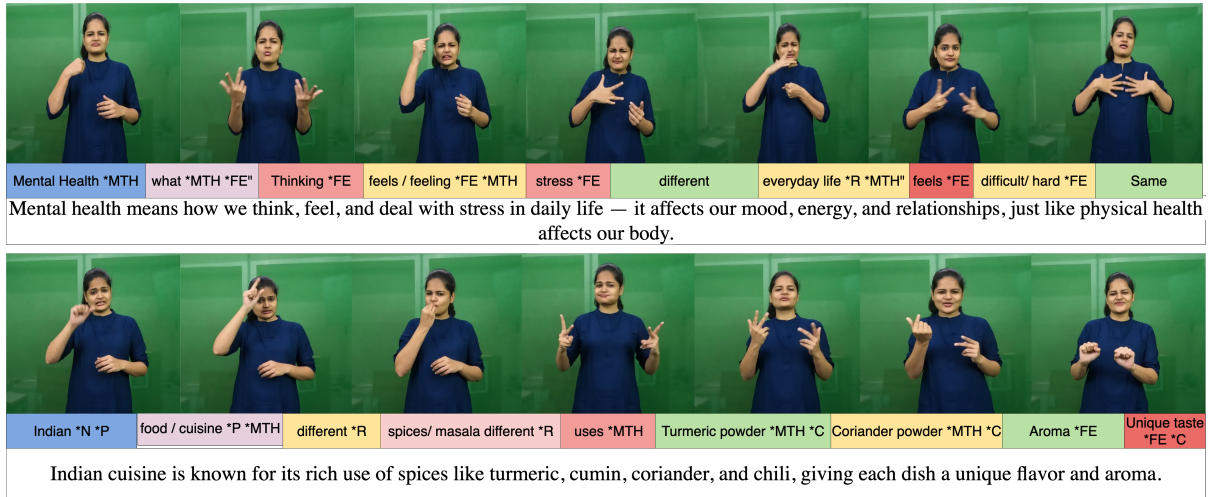


Figure 3.4 Example annotation workflow showing the source sentence, video frames, and the corresponding temporally aligned glosses and sign-type tags.

Annotations were performed using a web-based interface VIA video annotation tool [20], following protocols similar to recent CSLR studies [70]. The annotations were carried out by the certified ISL interpreter, ensuring consistency between execution and labeling. For every sign token, the interpreter provided free-form gloss labels aligned with the visual sign boundaries. The original ChatGPT-generated English text was displayed alongside the video to aid in semantic alignment, ensuring that selected glosses accurately reflected both the lexical meaning and the context. In cases where a single spoken word corresponded to multiple signs, annotations included both the compound gloss and individual components. To capture the nuances of ISL, the annotation schema included specific tags for sign types, such as *MTH for mouthing, *P for pointing, and *NS for no-sign segments. These tags, inspired by [70], enrich the dataset by distinguishing between manual signs and other communicative signals. Figure 4.1 visualizes this process, displaying the alignment of glosses and tags with the video timeline. While primarily designed for generation, the fine granularity of these annotations makes **SanketVaani-1K** a valuable resource for tasks such as sign language translation, gloss spotting, and sign retrieval.

3.5 Experiments and Results

To validate the efficacy of our proposed scalable Sign Language Production (SLP) framework, we conducted a comprehensive set of evaluations spanning data preprocessing quality, gloss generation accuracy, retrieval semantics, and motion interpolation fidelity. Furthermore, we performed a user study with Deaf participants to assess understandability and fluency of the generated sentences.

3.5.1 Data Pre-processing Evaluation

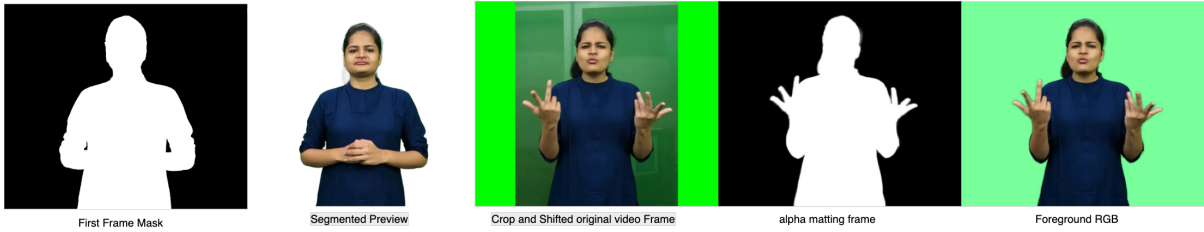


Figure 3.5 Qualitative results of our preprocessing pipeline. Each column shows key stages, including initial frame mask, initial segmented frame, signer detection and normalization, alpha matting, and final composite with background replacement.

We first qualitatively evaluated our preprocessing pipeline by examining visual consistency, signer isolation, and background replacement stability. As illustrated in Figure 3.5, the pipeline ensures precise temporal alignment and consistent framing across the dataset. The integration of YOLOv9 and ByteTrack for signer normalization successfully standardizes the signer’s position and scale, robust to varying body movements. Additionally, the matting module (MatAnyone) effectively isolates fine details such as fingers and hair, allowing for seamless compositing onto a neutral background without edge artifacts.

3.5.2 Text-to-Gloss Generation

Evaluating the Text-to-Gloss (T2G) component for Indian Sign Language (ISL) is challenging due to the absence of large-scale parallel ISL-English corpora with ground truth glosses. To rigorously assess our LLM-based approach, we utilized the RWTH-PHOENIX-Weather 2014T dataset (German Sign Language) as a proxy benchmark, given its high-quality gloss annotations.

We convert Text into a sequence of glosses following the respective linguistic rules as discussed in section (3.3.2) for ISL. Since there is no parallel text dataset for ISL, we evaluated our method on RWTH-PHOENIX-Weather 2014T [22], a German Sign Language (GSL) corpus. As shown in Table 3.3, 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

BLEU $\uparrow$				ROUGE-L $\uparrow$	METEOR $\uparrow$	CHrF $\uparrow$
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 3.3 Automatic evaluation of generated gloss sequences on the RWTH-PHOENIX-Weather 2014T test set using BLEU, ROUGE-L, METEOR, CHrF, and BERTScore.

Gloss Seq Variant	TSOV Acc. $\uparrow$	F1	Recall	Precision
RWTH-PHOENIX-Weather 2014T	0.72	0.723	0.687	0.704
ChatGPT-4o	0.76	0.791	0.740	0.764

Table 3.4 Evaluation of gloss sequence generation using TSOV word-order accuracy and BERTScore.

TSOV accuracy (0.72) compared to ChatGPT-generated glosses (0.76) in Table 3.4. 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. Certified ISL interpreter verified 88% (44/50) GPT-4o gloss outputs as correct; 6 required minor edits.

3.5.3 Vector Database Retrieval and Disambiguation

A critical challenge in SLP is handling polysemy, where a single English word maps to multiple distinct signs depending on context (e.g., "bank" as a financial institution vs. "bank" of a river). We evaluated the vector database’s ability to perform sense disambiguation by querying it with polysemous words embedded within different sentence contexts.

Table 3.5 demonstrates the system’s retrieval performance. For ambiguous query terms such as "light" (lamp vs. weight) or "fair" (festival vs. just), the contextual embeddings successfully retrieved the semantically correct sign variant (Rank 1). If the retrieval confidence score falls below a threshold (40%), the system falls back to fingerspelling to prevent semantic errors.

3.5.4 Interpolation

The interpolation of human motion within sign language videos presents unique challenges distinct from generic video frame interpolation. While minor geometric or textural inaccuracies in natural scenes often go unnoticed, sign language requires rigorous fidelity in hand configurations, palm orientation, and facial expressions, as distortions in these areas can fundamentally alter or obscure semantic meaning. The motion dynamics involve rapid, non-linear trajectories with complex co-articulation between consecutive signs. Furthermore, the necessity for synchronized bimanual coordination and the frequent occurrence of self-occlusion (e.g., hands crossing the face) render standard optical-flow-based methods susceptible to severe warping artifacts.

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 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 using a power of attorney.
bank (finance)	1	A home loan is money borrowed from a bank to buy a house and repaid monthly with interest.
	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 3.5 Top-3 sentence retrieval for five ambiguous queries with different meanings in the context.

We evaluate **FramePack** against three state-of-the-art baseline DreamMover [85], BiM-VFI [82], and VFIMamba [109] to analyze the trade-offs between perceptual quality, temporal consistency, and computational efficiency across diverse upper-body motions.

To create the evaluation set, we identify the co-articulation midpoint t_m for each annotated gloss boundary, convert it to a frame index, and extract a nine-frame window centered at this point. The two outer frames serve as anchors, while the seven inner frames constitute the ground-truth targets, yielding a total of 3,963 anchor–target pairs. All methods generate seven intermediate frames at 720×1080 resolution. Performance is assessed using PSNR, SSIM, LPIPS, FID [105], NIQE [112], MS-SSIM [73], GMSD [102], and KID [7].

Baselines: **DreamMover** is a diffusion-based approach employing feature-level warping and two-stage latent fusion. **BiM-VFI** utilizes bidirectional motion fields to interpolate frames at arbitrary timesteps, aiming for perceptually rich detail. **VFI-Mamba** integrates state-space modeling with token interleaving to manage large motions in high-resolution inputs. **FramePack** operates on a 13B Hunyuan-DiT [45] backbone, conditioned on both endpoint frames. It employs geometric compression of context frames to maintain constant memory usage and utilizes bidirectional generation with anchor sampling for interpolation.

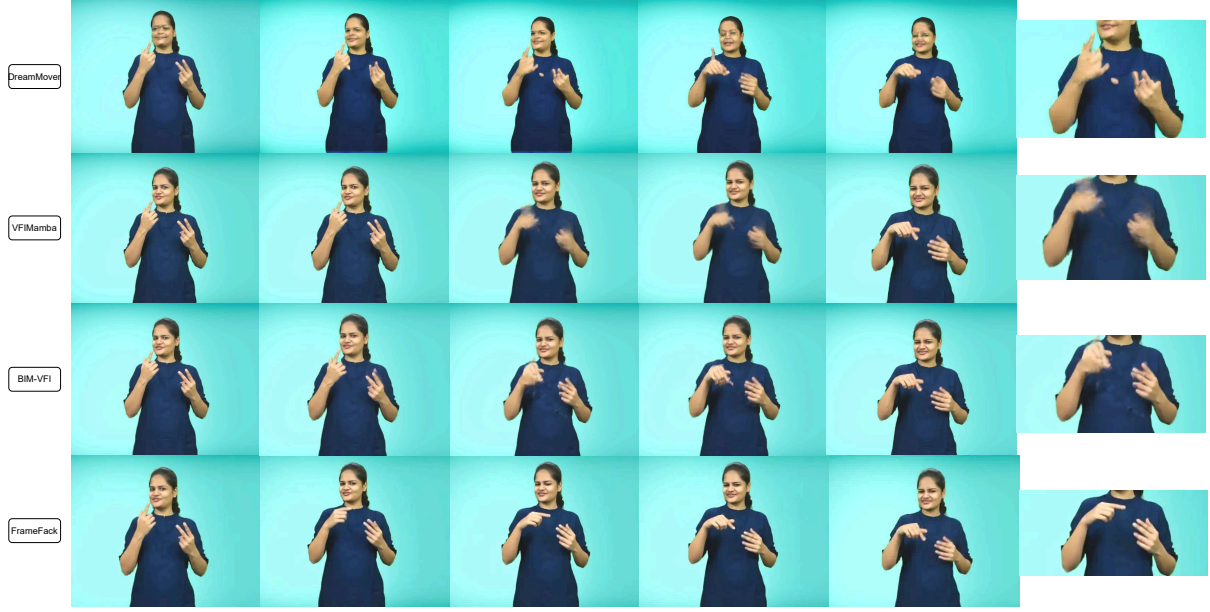


Figure 3.6 Qualitative comparison of DreamMover, VIFMamba, BiM-VFI, and FramePack interpolation methods.

Method	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	FID $\downarrow$	NIQE $\downarrow$	MS-SSIM $\uparrow$	GMSD $\downarrow$	KID $\downarrow$
DreamMover	26.02	0.915	0.117	56.82	46.85	0.947	0.103	0.065
VIFMamba	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 3.6 Quantitative comparison of interpolation methods. Higher is better for PSNR, SSIM, MS-SSIM; lower is better for LPIPS, FID, NIQE, GMSD, and KID.

Qualitative Results: Figure 3.6 presents a visual comparison of the intermediate frames generated between anchors. VIFMamba exhibits progressive blurring from the second interpolated frame onwards, particularly degrading hand and face details. DreamMover maintains the global pose effectively but introduces mild facial distortions and artifacts on the hands in the middle frames. BiM-VFI results in conspicuous artifacts, such as distortions on the subject’s left hand in the third frame. In contrast, FramePack consistently preserves facial fidelity and precise hand articulation across the entire sequence, producing the most visually coherent result.

Quantitative Results: FramePack yields lower scores on distortion-based metrics (PSNR, SSIM, MS-SSIM, GMSD) as these metrics typically penalize high-frequency details in favor of blur and strict pixel alignment. For instance, sub-pixel shifts in finger positioning incur high penalties, whereas blurred predictions minimize measured error. Conversely, perceptual metrics such as LPIPS, FID, KID, and NIQE, which utilize feature statistics from natural images, often misinterpret the high-frequency details

of hands and faces as noise. Consequently, while FramePack incurs higher numeric penalties on these specific metrics, it produces outputs that are perceptually sharper and more natural.

Figure 3.8 provides a comparative visual audit of our proposed method against three state-of-the-art Sign Language Production (SLP) pipelines:

Ground-Truth (GT): Real frames captured from the source signer provide the upper-bound reference for visual realism.

FramePack (Ours): Generates intermediate sequences by interpolating directly between real anchor frames, effectively preserving signer identity and photorealism.

Spoken2Sign: This method fits an SMPL-X mesh to each anchor frame and linearly interpolates the 3D body parameters. The resulting avatar rendering illustrates how even minor pose-estimation errors can propagate and degrade quality.

DW-Pose Interpolation [104]: This approach detects 2D keypoints on anchor frames and interpolates the skeletal structure. While the resulting motion is smooth, the final visual quality is heavily dependent on the downstream generation or rendering stage.

UniAnimate-DiT [100]: This method utilizes a single reference image alongside an interpolated DW-Pose trajectory, employing a diffusion model to synthesize the full animation. While it can produce photorealistic results, it often suffers from identity drift and loss of fine-grained hand details.

The side-by-side comparison in Figure 3.8 highlights the specific trade-offs of each approach: FramePack excels in preserving real-world texture and photorealism; Spoken2Sign and DW-Pose offer structural coherence but lack visual fidelity; and UniAnimate-DiT achieves photorealism but is prone to drifting artifacts in critical articulators.

3.5.5 Step-by-Step Walk-Through

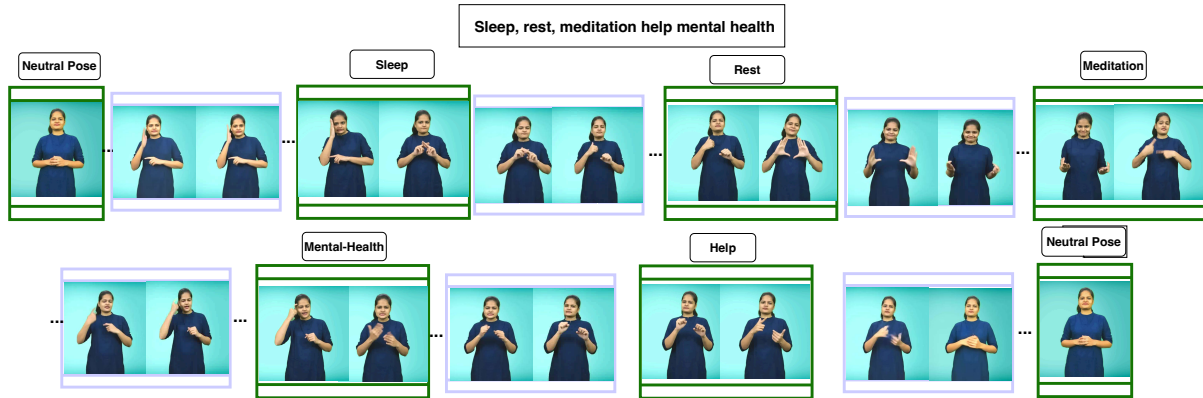


Figure 3.7 Illustrative example of the proposed SLP pipeline in a medical use case. The first row depicts the input sentence for inference. Green boxes indicate retrieved sign clips, purple boxes denote interpolated segments, and the first and last frames correspond to neutral poses.

Figure 3.7 illustrates the proposed SLP pipeline applied to the medical domain sentence: “Sleep, rest, meditation help mental health.” First, the text-to-gloss translator converts the input sentence into an ISL gloss sequence (see Section 3.3.2). For each gloss, the system queries the vector database to retrieve contextually appropriate sign clips. If the retrieval match score falls below 40%, the system defaults to fingerspelling the word using pre-recorded alphabet clips. Finally, the retrieved segments are stitched together using interpolation. We generate eight intermediate frames between each pair of original segments, with neutral poses inserted at the boundaries, resulting in a continuous, fluid signing sequence.

3.6 User Study

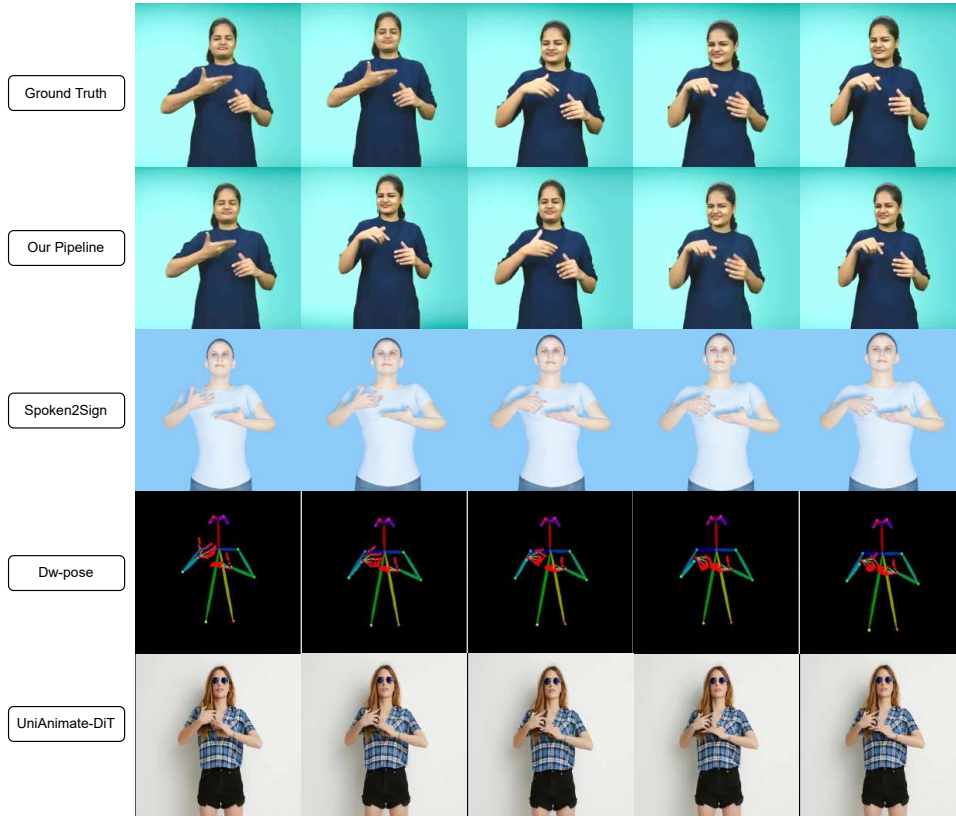


Figure 3.8 Qualitative comparison of SLP approaches: (1) FramePack – interpolation based; (2) Spoken2Sign – avatar-based synthesis using an SMPL-X body model; (3) UniAnimate-DiT – diffusion-based, 2D keypoint-driven.

As the ultimate objective of this system is to bridge the communication gap between the Deaf and hearing communities, direct evaluation by Deaf signers is indispensable. We conducted a user study to assess the naturalness and accuracy of the generated signing, specifically aiming to identify usability challenges that automated metrics may fail to capture.

Method	Only Interpolation			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 3.7 Subjective ranking results. Lower mean rank (MR↓) is better; higher #1 votes and Borda score (B↑) are better.

Method	Understandability ↑	Fluency ↑	Naturalness ↑	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 3.8 Mean opinion scores (MOS) from 12 raters. Scale: 1 = poor, 5 = excellent.

The study involved 12 participants and evaluated the methods using two protocols. Table 3.7 details the mean rank (MR), average Borda score (B), and the number of first-place rankings (#1) for each method. The first protocol assessed pure interpolation quality, displaying clips containing only the two anchor frames and the synthesized in-betweens at half speed, allowing raters to focus on motion smoothness and artifacts. The second protocol evaluated full-sentence videos, where seven synthetic frames were integrated at every gloss boundary (as described in Section 3.5.4). This dual approach distinguishes raw interpolation fidelity from its impact on continuous signing flow. The results suggests that among the interpolation methods, FramePack introduces the least perceptible distortion.

The results of the Mean Opinion Score (MOS) evaluation are summarized in Table 3.8. The study consistently favors FramePack, which achieved the highest scores across all dimensions, with an overall MOS of 3.15/5. Spoken2Sign ranked second (2.47), while UniAnimate received the lowest ratings (2.06). These findings confirm that Deaf raters perceive FramePack’s output as significantly more understandable, fluent, and natural compared to the baseline methods.

3.7 Summary

In this chapter, we presented a comprehensive framework for scalable 2D Sign Language Production. We introduced SanketVaani-1K, a domain-diverse, gloss-annotated dataset specifically tailored for generation tasks. We proposed a modular pipeline that leverages a context-aware vector database to retrieve high-quality video segments and employs FramePack, a diffusion-based interpolation model, to stitch these segments into fluid, continuous sign language videos. Our approach addresses the critical lack of data for low-resource sign languages like ISL. Extensive evaluations, including user studies with the Deaf community, confirm that our method produces sign language videos that are more understandable, fluent, and natural than existing state-of-the-art avatar and generative baselines. This work lays the

foundation for accessible, automated sign language translation systems that can be easily updated and scaled to new domains.

In the following chapter, we will address the limitations regarding the lack of complex co-articulation, depth information, and the need for signer anonymisation by exploring 3D motion representations and generative frameworks.

3.8 Acknowledgment

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.

Chapter 4

Gloss SLQA: A Gloss-centric Approach to Sign Language Question Answering

4.1 Introduction

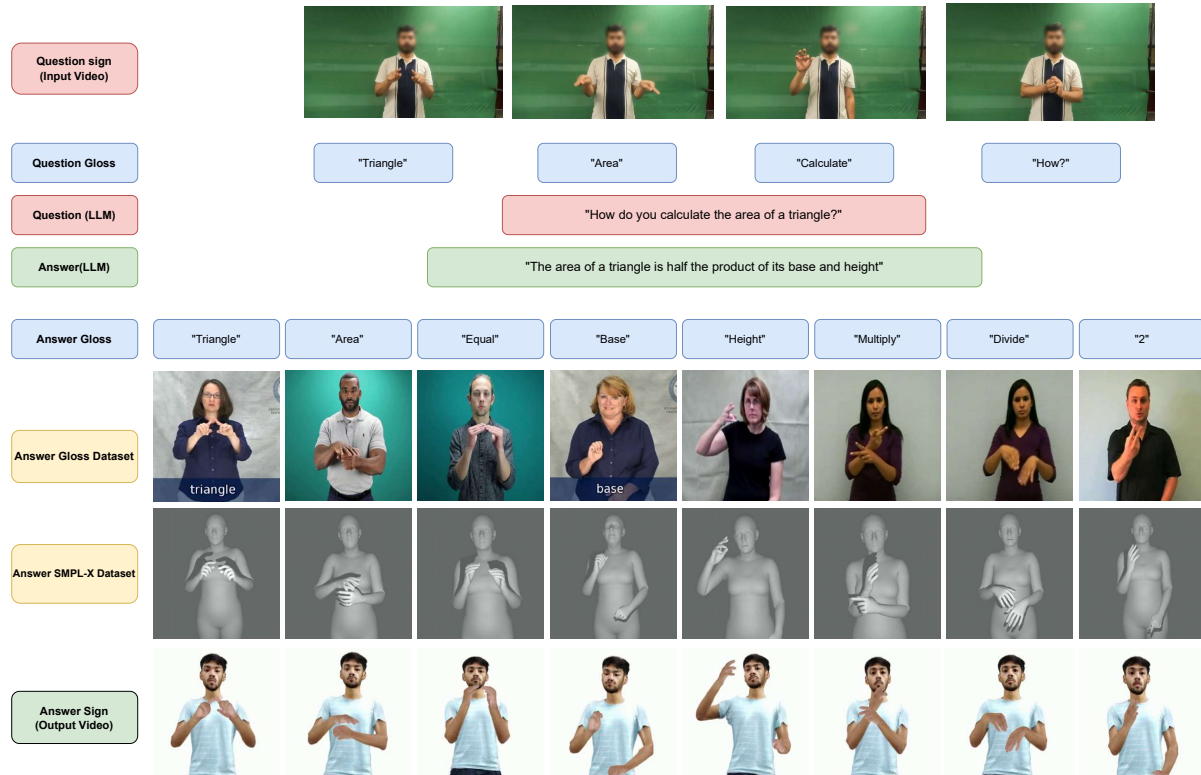


Figure 4.1 Overview of the Gloss centric Sign Language Question Answering (Gloss SLQA) pipeline. A user inputs a question in sign language (top row), which is converted into glosses (second row), followed by text-based question and answer generation using a large language model (third and fourth rows). The generated answer is then translated back into glosses (fifth row) and synthesized into a sign language video output using a Gaussian-based avatar driven by SMPL-X models (last row).

Current conversational AI technologies have revolutionized information access for hearing populations. Sophisticated speech-driven systems such as Siri, Alexa, and the recent GPT-4o enable users to ask questions and receive answers seamlessly in natural spoken language. However, this technological leap has largely overlooked the Deaf and Hard-of-Hearing (DHH) community, for whom sign language is the primary mode of communication. The text-based interfaces often offered as "accessible" alternatives are insufficient, as they do not align with the linguistic structure or modality of sign language.

While the 2D interpolation framework presented in Chapter 3 offers high visual fidelity, it is inherently limited by its reliance on a specific signer’s footage. Scaling such a system requires re-recording a single actor for thousands of vocabulary words, a resource-intensive bottleneck that restricts vocabulary expansion. Furthermore, pixel-level manipulation limits the ability to anonymize the signer or render the signing from novel viewpoints, capabilities essential for privacy-preserving applications and immersive AR/VR environments.

To address these limitations and bridge the accessibility gap, we propose a transition to a Gloss-Centric 3D Sign Language Production framework. In this chapter, we introduce a method that decouples motion from appearance. By leveraging existing large-scale, multi-signer datasets and retargeting their motion onto a unified parametric representation, we can synthesize consistent, photorealistic sign language videos without the need for extensive new recordings. This approach serves as the visual backbone for Gloss-SLQA, effectively creating a "Siri for Sign Language."

4.2 Dataset

To support the *Sign Language Production* (SLP) component of the Gloss SLQA framework, a robust, multi-lingual dataset was constructed to facilitate high-fidelity video synthesis in both American Sign Language (ASL) and Indian Sign Language (ISL). The dataset serves as a **Gloss-Motion Video Dictionary**, where each sign is mapped to optimized SMPL-X parametric body models to ensure structural motion and realism.

Data Curation and Scaling The repository was built by aggregating and cleaning existing sign language resources to provide a diverse foundation for synthesis: **Source Repositories:** Data for ASL was gathered from the American Sign Language Dictionary and the WLASL dataset. ISL data was sourced from the ISLRT Dictionary and CISLR. **Vocabulary Size:** This integration yielded a unified vocabulary of 13,357 unique glosses for ASL and 4,148 unique glosses for ISL. For British Sign Language (BSL) we combine BSL Dictionary¹ and BSL dict [54].

¹<https://www.signbsl.org/>

Dataset	Vocab Size	Avg. Freq.	Modal Freq.
Unified ASL-SLP	13,357	3.57	1
Unified BSL-SLP	13313	4.08	1
Unified ISL-SLP	4,148	1.14	1

Table 4.1 Vocabulary Statistics for Sign Language Production (SLP)

4.3 Methodology

4.3.1 Gloss-Motion Video Dictionary: Leveraging Multi-Signer Data

The foundation of our avatar is a dictionary of 3D motion primitives derived from word-level sign videos. We utilize the ASL-SLP and ISL-SLP datasets, which contain over 13,000 and 4,000 unique glosses, respectively, to build this repository. Capturing the nuances of sign language requires precise estimation of manual (hand) and non-manual (face) markers, which standard pose estimators often miss. We employ a multi-stage fitting process:

1. **Body and Face Estimation:** We use SMPLer-X for robust whole-body 3D pose and shape estimation. For facial expressivity, critical for grammatical markers in sign language, we utilize the SMIRK model to regress accurate FLAME parameters.
2. **Refined Fitting (ExAvatar):*** Standard parametric models often fail to capture the fine-grained contact and articulation of fingers. To address this, we apply the ExAvatar fitting method, which introduces joint offsets ΔJ and ΔV to optimize the alignment of hand and face joints against 2D keypoints detected by HRNet.

The resulting optimized SMPL-X parameters, along with camera extrinsics and joint offsets, are stored in the dictionary.

4.3.2 Text to gloss

We use the same strategy as in the earlier chapter. We convert it into a sequence of glosses following the respective linguistic rules. we prompt the LLM using curated linguistic guidelines—drawn from Lifepri² for ASL and from [26] for ISL. Full prompt instructions are provided below:

4.3.3 Answer to Gloss Generation

Here are the Prompts used with Rules defined in the prompt:

- **ASL Gloss generation Prompt:**

²<https://www.lifepriint.com/>

You are an American Sign Language (ASL) \
translation assistant.\

use following the unique grammar \
and structure of American \
Sign Language (ASL).\

Here are the guidelines you \
should follow for each translation:

Grammar and Structure: \

Use ASL's "topic-comment" or \
"subject-verb-object" structure, \
placing the main idea or topic \
at the beginning of each sentence.
Omit unnecessary English articles \
like "a" and "the."

Conceptual Translation:\

Focus on conveying the meaning \
of the sentence rather than \
a word-for-word translation.\

Some English words or expressions \
may need to be simplified or \
replaced with ASL concepts that \
express the same idea.

Fingerspelling:\

When a word needs to be fingerspelled \
in ASL (such as names, \
abbreviations, or uncommon words), \
insert hyphens between each \
letter (e.g., "J-O-H-N" for the \
name John). This accurately represents \
the ASL fingerspelling convention.

- **BSL Gloss generation Prompt:**

Convert the following \
English sentences into \

British Sign Language (BSL) grammar.
Use the topic-comment structure,\
omit unnecessary linking words \
(like "is," "are," "to"),\
and if there is a time element,
place it at the beginning.\

Follow these rules:

1. Use ****topic-comment**** structure:\
start with the main topic, \
followed by details or actions.
2. Place ****time indicators****\
(e.g., "yesterday," "tomorrow") \
at the beginning of the sentence.
3. ****Omit linking words****\
(e.g., "is," "are," "a," "the") \
and keep the sentence direct.
4. ****Facial expressions****
\are implied for questions and emphasis:
 - Raise eyebrows for yes/no questions.
 - Furrow eyebrows for WH-questions \
(who, what, where, etc.).
5. ****Negate**** with \NOT" if needed, \
and omit unnecessary pronouns\
if the context is clear.

Example conversion:

- English: "I will meet you \
tomorrow at the park."
- BSL: "TOMORROW PARK MEET YOU."

Now, please convert these\
sentences into BSL grammar:

Once the target gloss sequence is determined by the upstream SLQA module, the synthesis pipeline retrieves the corresponding optimized SMPL-X parameter sequences from the Gloss-Motion Dictionary. Each entry in the dictionary represents a canonical realization of a single gloss, typically captured in

isolation. While this design enables scalable reuse across sentences, naïvely concatenating such segments leads to visually implausible motion artifacts, most notably the *rest-pose collapse* phenomenon, where the signer’s hands return to a neutral lowered position between successive signs.

To address this challenge, we introduce a structured multi-stage synthesis procedure that enforces kinematic consistency, temporal smoothness, and perceptual continuity across the generated sign sequence.

Motion Processing and Stitching: To synthesize a fluid sequence, we apply the following processing steps: **Consistency Enforcement:** To ensure the avatar maintains a consistent identity throughout the sentence, we retain the body shape parameters (β) and the global orientation from the first frame of the sequence, applying them to all subsequent retrieved segments, ensuring consistent body proportions, camera alignment, and viewpoint throughout the sequence. **Kinematic Stabilization:** Since leg and spine movements are minimal and often noisy in sign language datasets, we set the corresponding joint angles in the kinematic tree to zero to prevent unnatural lower-body jitter. **Hand Trimming:** To mitigate discontinuities introduced by isolated gloss recordings, we apply a hand activity–based trimming heuristic. Using wrist and finger joint velocities, we detect frames where both hands fall outside the signing space (i.e., lowered or resting positions). These frames are removed from the beginning and end of each gloss segment, retaining only the linguistically salient signing portion.

Co-articulation via Interpolation: To model natural transitions between glosses (co-articulation), we employ Spherical Linear Interpolation (SLERP) in the quaternion space. We generate five intermediate frames between the final pose of gloss g_i and the initial pose of gloss g_{i+1} :

$$q_{interp}(t) = \text{Slerp}(q_1, q_2, t) \quad (4.1)$$

where q_1 and q_2 represent the quaternions of the boundary poses. SLERP ensures constant angular velocity in quaternion space, avoiding artifacts such as joint flipping or sudden rotational acceleration that commonly arise from linear interpolation. The interpolated poses are inserted between gloss segments, producing visually continuous transitions while preserving the original articulation of each sign.

To further suppress high-frequency noise introduced during fitting and interpolation, we apply a Savitzky–Golay temporal smoothing filter across the entire sequence. This filter locally fits a low-degree polynomial to joint trajectories, effectively reducing jitter while preserving sharp motion boundaries essential for sign articulation, such as handshape changes and directional movement.

Neural Rendering: For the final visual output, we utilize ExAvatar, an expressive 3D human avatar model based on 3D Gaussian Splatting (3DGS). Unlike traditional mesh-based rendering, this approach allows for whole-body control via SMPL-X while capturing high-frequency visual details such as clothing dynamics and lighting, significantly enhancing photorealism.

4.4 Experiments and Results

We evaluated the performance of our 3D sign-lanugage generation pipeline against state-of-the-art baselines.

4.4.1 Quantitative Analysis

We compare with SMPLer-X (SMPLer-X-L32), which we use to initialize the optimized SMPL-X parameters. For qualitative comparison, we replace SMPL-X expressions with Flame parameters regressed by the Metrical Tracker [118]

We also evaluate against Spoken2Sign[119], obtaining human pose estimations from their framework, interpolating motions, and rendering with Blender. While Spoken2Sign uses a separate model for co-articulation, we apply SLERP interpolation.

- **Keypoint Loss (MPJPE):** we use HRNet 2D keypoint estimation as pseudo ground truth, re-projecting SMPL-X 3D joints into 2D in the OpenPose format with 135 keypoints. We calculate the Mean Per Joint Position Error (MPJPE) for visible keypoints in each frame. Then, we average the MPJPE across all frames within each video and further average these scores across 250 randomly sampled videos from the WLASL test set to obtain the final MPJPE score.
- **Back-Translation Accuracy:** As a proxy for semantic intelligibility, we fed the generated videos back into a pre-trained Sign Language Recognition (SLR) model (NLA-SLR) to verify if the synthesized signs could be correctly recognized.

Model	Keypoint Loss ↓	Retrieval Acc. (Top-1) ↑	Retrieval Acc. (Top-5) ↑
HRNet (GT)	0	0.58	0.80
SMPLer-X	5.66	0.46	0.66
Spoken2Sign	3.19	0.53	0.75
ExAvatar* (Ours)	3.86	0.55	0.76

Table 4.2 Comparison of motion quality and semantic intelligibility. While Spoken2Sign achieves lower keypoint error, our ExAvatar*-based approach achieves higher back-translation accuracy, indicating better preservation of semantic features.

As shown in Table 4.2, our approach (ExAvatar*) achieves the highest back-translation accuracy among the generation methods. This indicates that although our keypoint loss is slightly higher than Spoken2Sign, our method better captures the critical hand and finger configurations necessary for accurate sign recognition.

4.4.2 Qualitative and Human Evaluation

To assess the perceptual quality and semantic fidelity of the generated sign language videos, we conducted a two-part human evaluation: (1) a scaled opinion survey with general users to evaluate

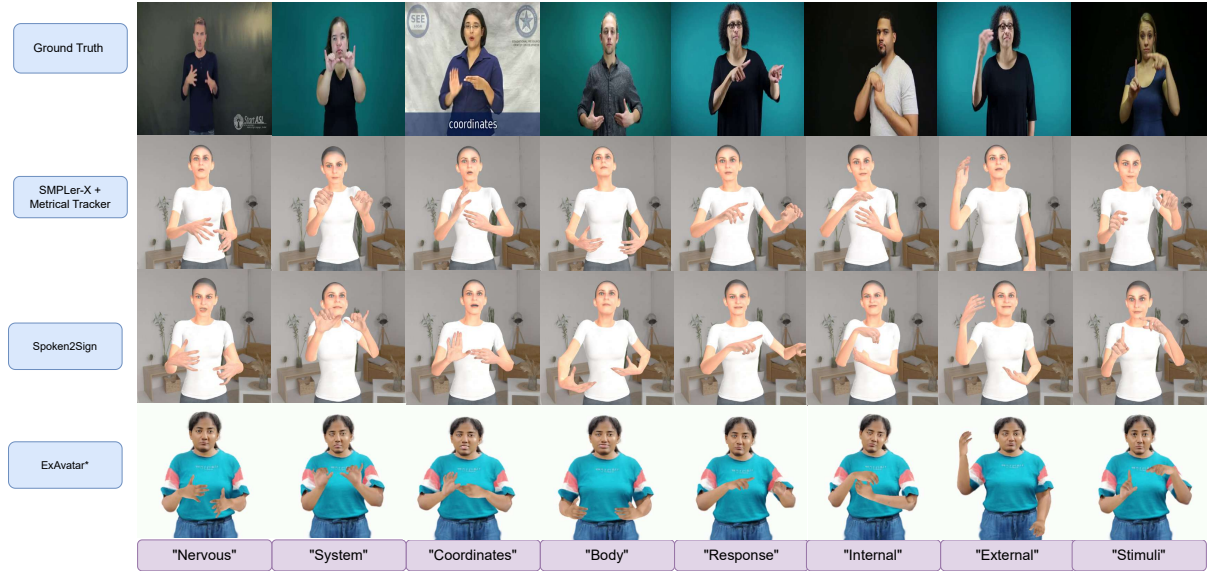


Figure 4.2 Qualitative comparison of SLP methods. Rows from top to bottom: Ground Truth, SMPLer-X, Spoken2Sign, and ExAvatar* (Ours). Note the improved hand definition and clothing detail in the ExAvatar* row.

visual realism, and (2) targeted feedback from Deaf ISL users to assess comprehensibility and linguistic accuracy.

4.5 Sign Language Production

4.5.1 Human Evaluation

To assess both perceptual quality and semantic fidelity of the generated sign language videos, we conducted a two-part human evaluation: (1) a *Scaled Opinion Survey* with a general participant pool focused on visual and expressive quality, and (2) *targeted feedback from members of the Deaf community* evaluating the semantic accuracy of translations.

4.5.2 Scaled Opinion Survey

We designed a survey featuring 18 anonymized sign language videos 6 each from **ExAvatar***, **Spoken2Sign**, and **SMPLer-X**. A total of 27 participants rated the videos on a 5-point Likert scale across the following perceptual dimensions:

Overall Naturalness: How natural the video appears (1 = Very unnatural, 5 = Very natural)

- "How natural does this sign language video appear to you on a scale of 1 to 5?" (1 - Very unnatural, 5 - Very natural)

Movement Fluency: Smoothness of hand, arm, and facial movements

- "How smooth and fluid are the hand, arm, and facial movements in this video?" (1 - Very rigid, 5 - Very fluid)

Facial Realism: Realism of facial expressions and body language

- "How realistic are the facial expressions and body language in this video?" (1 - Very unrealistic, 5 - Very realistic)

Photo-Realism: Likeness to a photo-realistic human signer.

- "To what extent does this video resemble signing done by a photo-realistic person?" (1 - Not at all, 5 - Very much so)

Metric	ExAvatar*	Spoken2Sign	SMPLer-X
Overall Naturalness ↑	3.19 ± 0.29	3.17 ± 0.30	3.10 ± 0.29
Fluency of Movement ↑	3.46 ± 0.27	3.33 ± 0.30	3.25 ± 0.27
Realism of Facial Expressions ↑	3.11 ± 0.33	3.05 ± 0.35	2.97 ± 0.36
Photo-Realism ↑	3.24 ± 0.33	2.96 ± 0.32	2.93 ± 0.34
Combined Score ↑	3.27 ± 0.27	3.19 ± 0.28	3.11 ± 0.28

Table 4.3 Mean Opinion Scores (MOS) from user survey. Each metric was rated on a 5-point Likert scale (higher is better). Combined Score is the unweighted average across all metrics.

ExAvatar* consistently outperformed the other systems across all dimensions, especially in movement fluency and perceived realism. This suggests that its hand and body dynamics are more coherent and its rendering more visually engaging.

4.5.3 Deaf Community Feedback

To evaluate semantic accuracy and understandability, we conducted a focused study with 10 Deaf participants (one interpreter and nine daily ISL users), all aged 20–29 with good to very good fluency in both ISL and English.

Participants were first shown a brief demo of our *Sign Language Question Answering (SLQA)* system, featuring *ExAvatar**-generated responses. They could select a question and view its signed video answer. After viewing, they rated each video on four dimensions using a 5-point Likert scale:

Comprehensibility: "How easy is it to understand the video?"

Semantic Alignment: "How well does the sign language video match the text answer?"

Movement Fluency: "How natural and fluid are the signing movements?"

Overall Understandability: "How well did you understand the complete answer?"

Each used a 5-point Likert scale from "Very Poor" to "Very Good."

Metric	ExAvatar*
Comprehensibility ↑	2.4 ± 1.15
Semantic Alignment ↑	2.8 ± 1.03
Movement Fluency ↑	3.1 ± 0.98
Overall Understandability ↑	2.9 ± 1.19
Combined Score ↑	2.8 ± 1.09

Table 4.4 Mean Opinion Scores (MOS) from a user survey with 10 Deaf participants. Each metric was rated on a 5-point Likert scale (higher is better). Combined Score is the unweighted average across all metrics.

Out of 10 participants, 7 found the idea of the Sign Language Question Answering System very useful, while 3 slightly agreed. As shown in Table 4.4, scores were slightly lower than in the general survey, expected due to Deaf participants’ heightened sensitivity to sign nuances. Movement Fluency rated highest (3.1), reaffirming ExAvatar*’s strength in gesture articulation. Lower scores in Semantic Alignment (2.8) and Comprehensibility (2.4) suggest that certain signs may lack clarity or linguistic precision.

To further assess semantic accuracy, each participant was randomly assigned 2 out of 4 signed video prompts to translate into text. For each, they translated the **generated video first**, followed by the **stitched (reference)** version. These translations were then compared to the original ground truth text to assess how well meaning was preserved.

Q#	Gen. vs GT	Stitch. vs GT	Sem. Rating
Q1	2.8	2.8	3.0
Q2	1.2	1.6	2.6
Q3	2.0	2.5	3.0
Q4	2.5	2.5	3.5

Table 4.5 Semantic evaluation of signed video prompts (Q1–Q4). **Gen. vs GT** and **Stitch. vs GT** compare participant translations with ground truth (GT) for generated and stitched videos, respectively, rated on a 3-point Likert scale (*Different* = 1, *Similar* = 2, *Same* = 3). **Sem. Rating** reflects participant-perceived semantic clarity on a 5-point scale.

Signed Video Prompts: **Q1:** A helmet protects the head from injury. **Q2:** The minimum time is usually 6 months. **Q3:** A tax refund is money returned by the government if extra tax was paid. **Q4:** Exercise and rest help reduce stress.

The semantic ratings generally align with the translation scores videos rated higher in clarity also showed better translation accuracy (e.g., Q1 and Q4). Generated videos preserved meaning reasonably well compared to stitched references(sign video GT) (Gen avg: 2.12, Stitch avg: 2.35).

4.5.3.0.1 Interpretation: These evaluations collectively suggest that ExAvatar* shows improved perceptual realism and movement fluency, consistently outperforming baselines in general opinion ratings. While Deaf participants rated semantic and linguistic fidelity slightly lower, their feedback pro-

vides actionable insights for improving sign precision and grammar. The variance across individual scores also reflects the subjective nature of signing preferences, regional ISL variations, and the inherent complexity of evaluating expressive signing across gloss styles. Importantly, the generated videos already show competent baseline comprehensibility, reinforcing their potential for real-world SLQA deployment.

4.5.4 Qualitative Analysis

Figures 4.3, 4.4, 4.5, 4.6, 4.7 illustrate visual comparisons of one frame per gloss in a sentence, organized by domain.

In **Figure 4.3**, for the gloss ‘MEET’, occlusion causes Spoken2Sign to fail, while ExAvatar* successfully tracks which hand is above the other. For glosses such as ‘SAME’ and ‘TWO’, when a hand is not visible, both methods attempt to optimize the hand’s position, albeit differently.

In **Figure 4.4**, the gloss ‘NUTRIENTS’ shows a crossed finger gesture that Spoken2Sign misses, whereas ExAvatar* tracks it accurately.

In **Figure 4.6**, for the gloss ‘CAUSE’, both systems struggle when the hand is not visible and attempt to optimize the position elsewhere. However, the gloss ‘VOLCANIC ERUPTION’ demonstrates a clear failure in Spoken2Sign, while ExAvatar* handles the tracking correctly.

In **Figure 4.7**, in ‘ACCUMULATE’ the ExAvatar* makes the correction action while the Spoken2Sign fails.

Regarding expressions, Spoken2Sign tends to exaggerate facial expressions, leading to a less natural appearance. ExAvatar* generally performs better but occasionally misinterprets eye textures as part of the eyebrows, causing unnatural movements in certain scenarios.

4.6 Application: Gloss-Centric Sign Language Question Answering (Gloss-SLQA)

To establish the practical utility of the proposed 3D generation framework, we integrated it into a larger end-to-end application: **Gloss-Centric Sign Language Question Answering (Gloss-SLQA)**. This pipeline represents a collaborative effort where our text-to-sign generation module serves as the final response synthesis pipeline.

The Gloss SLQA pipeline is designed as a modular, scalable framework to facilitate natural interaction between sign language users and information systems. By utilizing a gloss-constrained approach, the system can be adapted to several critical real-world domains where information access is currently limited for the Deaf community. The Gloss-SLQA system is designed to bridge the information gap for Deaf individuals by enabling them to ask questions in sign language and receive answers in the same modality. As illustrated in Figure 4.9, the pipeline operates in three modular stages, utilizing glosses as the universal intermediate representation to ensure scalability across domains.

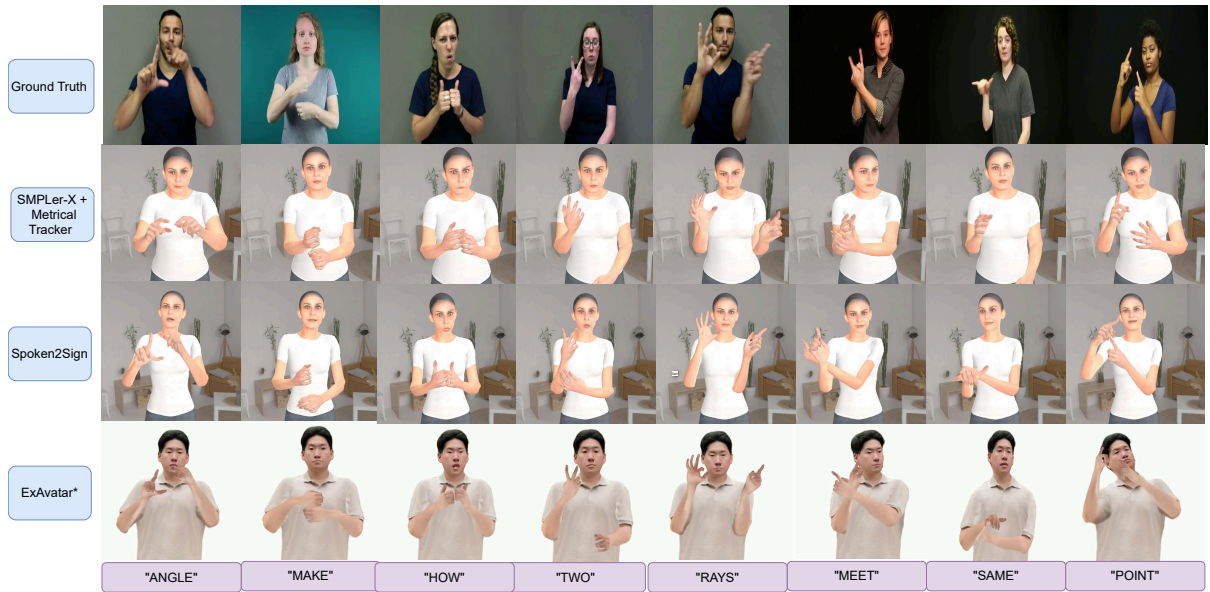


Figure 4.3 Education domain(ASL) SLP comparison of answer: "An angle is made by two rays meeting at a common endpoint."

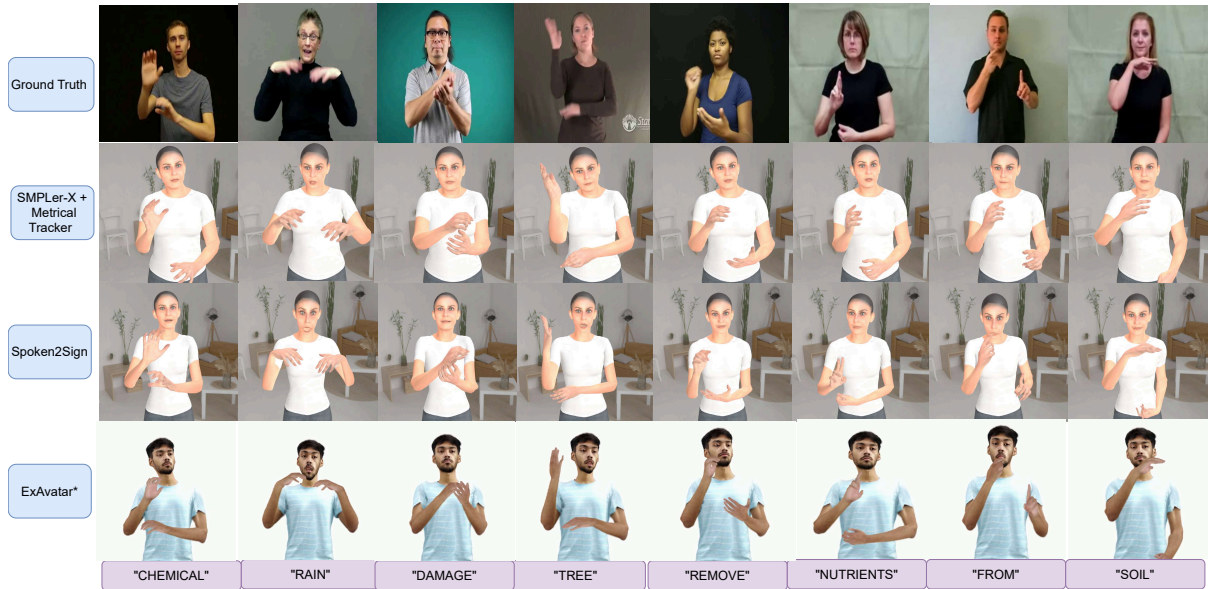


Figure 4.4 Environment domain(ASL) SLP comparison of answer: "Chemical rain damages tree, remove nutrients from soil..."

4.6.0.0.1 1. Sign Language Retrieval (Query Understanding): The user provides a query in sign language (e.g., "High blood pressure cause what?"). The system takes a "semi-continuous" video where the signer pauses briefly between signs. A recognition module segments the video and retrieves the corresponding sequence of glosses using a pre-trained Sign Language Retrieval (SLR) model.



Figure 4.5 Medical domain(ASL) SLP comparison of answer: "Reduce stress by practicing relaxation techniques, such as deep breathing."

4.6.0.0.2 2. Question & Answer Generation (LLM Reasoning): The retrieved gloss sequence is fed into a Large Language Model (LLM) to perform the reasoning tasks:

- **Gloss-to-Question:** The LLM first converts the disjointed glosses into a coherent natural language question (e.g., "What are the causes of high blood pressure?").

Question Generation from Identified Glosses Because sign language retrieval can sometimes be noisy or ambiguous, the system does not simply translate the Top-1 gloss.

Top-K Gloss Input: The LLM (such as GPT-4o or LLaMA-3.1) receives the Top-3 nearest candidates for each signed word to resolve potential recognition errors.

Semantic Grounding: A specific prompt guides the LLM to reconstruct these candidates into a coherent, contextually grounded natural language question.

Performance: Evaluation shows high BERT similarity scores (>0.95), confirming that the generated text accurately preserves the intended meaning of the original signs

- **Question-to-Answer:** The system generates a text answer restricted to a predefined vocabulary to ensure that every word in the response can be visually rendered.
- **Answer-to-Gloss:** Finally, the text answer is translated back into a sequence of ISL glosses (e.g., "HIGH BLOOD PRESSURE CAUSE STRESS SMOKING...") using a grammar-constrained prompt.

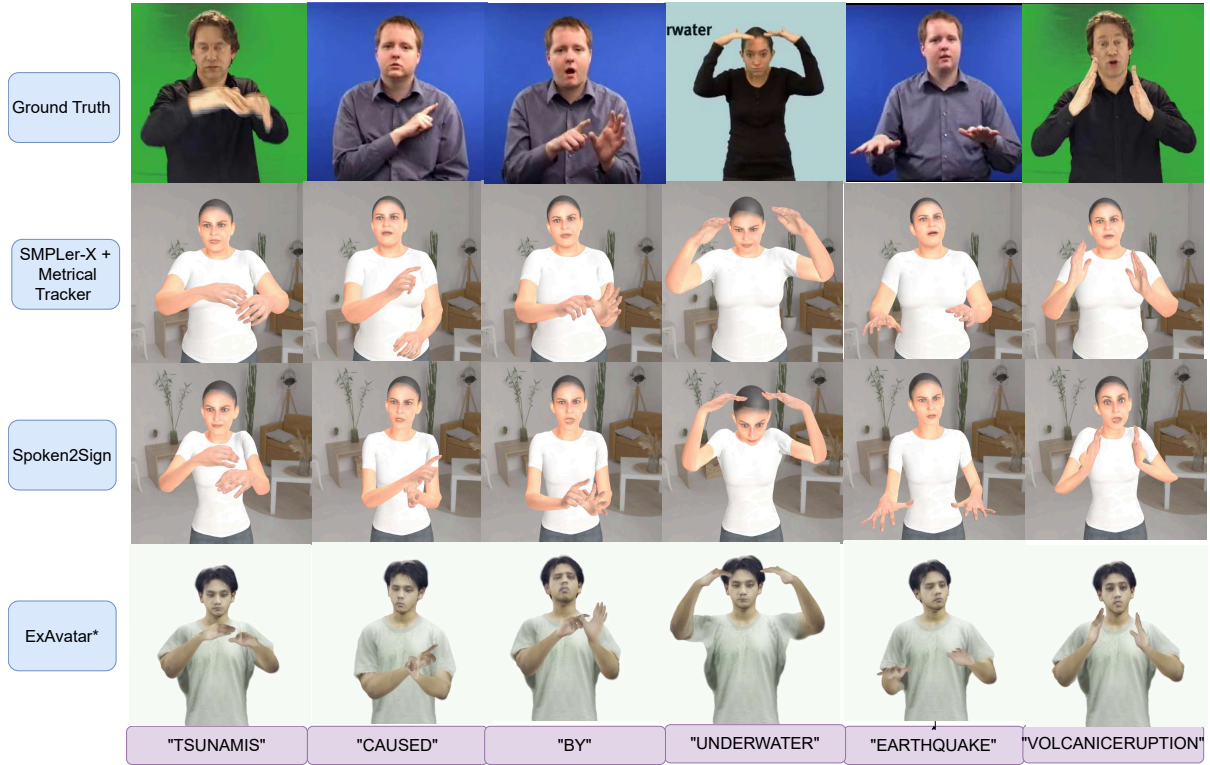


Figure 4.6 Science Tech. domain (BSL) SLP comparison of answer: "Tsunamis are caused by underwater earthquakes or volcanic eruptions."

4.6.0.0.3 3. Sign Language Production (Response Synthesis): This component represents the core contribution of this chapter. The generated answer gloss sequence is passed to our **3D Avatar Generation module**. The system retrieves the optimized SMPL-X motion parameters for each gloss from our dictionary, stitches them using the proposed SLERP interpolation and Savitzky-Golay smoothing , and renders the final photorealistic video using the **ExAvatar*** neural renderer.

This application demonstrates the robustness of our generation framework. Unlike previous methods that rely on pre-recorded sentences, our gloss-based stitching approach allows the QA system to dynamically synthesize answers for diverse domains—such as healthcare, finance, and education without requiring new motion capture data for every possible sentence.

4.7 Limitations

Despite the promising results of the Gloss-SLQA framework, several limitations remain that impact its generalizability.

Vocabulary Depth: While the curated gloss dictionaries are extensive, they may lack low-frequency or highly domain-specific terms, which can lead to retrieval failures in specialized scenarios.

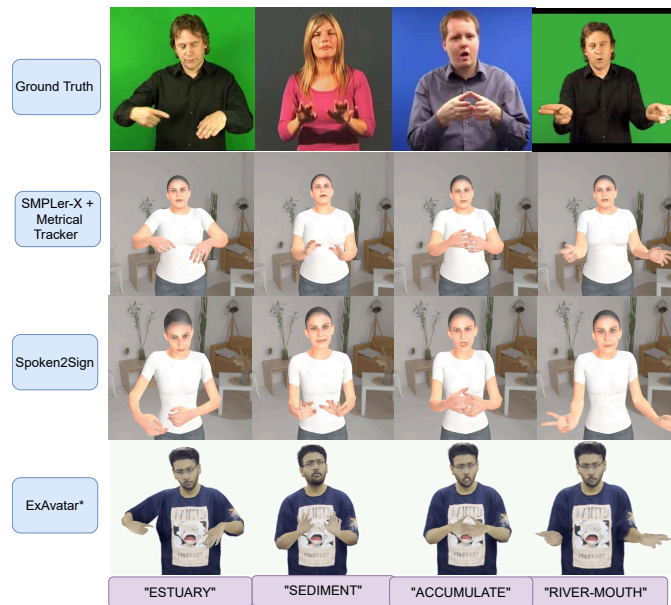


Figure 4.7 Geography (BSL) SLP comparison of answer: "This area is called an estuary, where sediment accumulates at the river mouth."

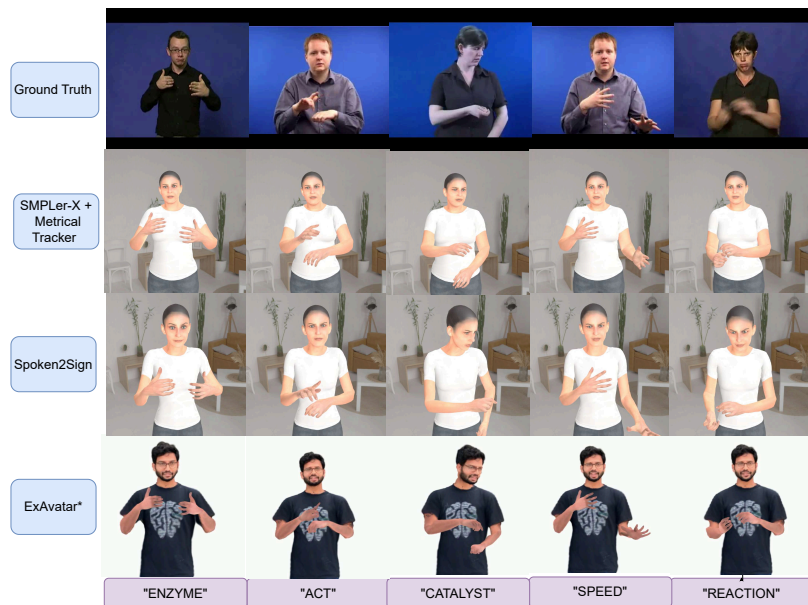


Figure 4.8 Chemistry (BSL) SLP comparison of answer: "Enzymes act as catalysts, speeding up chemical reactions."

Sign Style Constraints: The system performs best with semi-continuous signing (pauses between signs); however, performance degrades with natural, continuous signing due to the complexities of co-articulation and sign segmentation. Continuous Sign language recognition remains a challenge today.

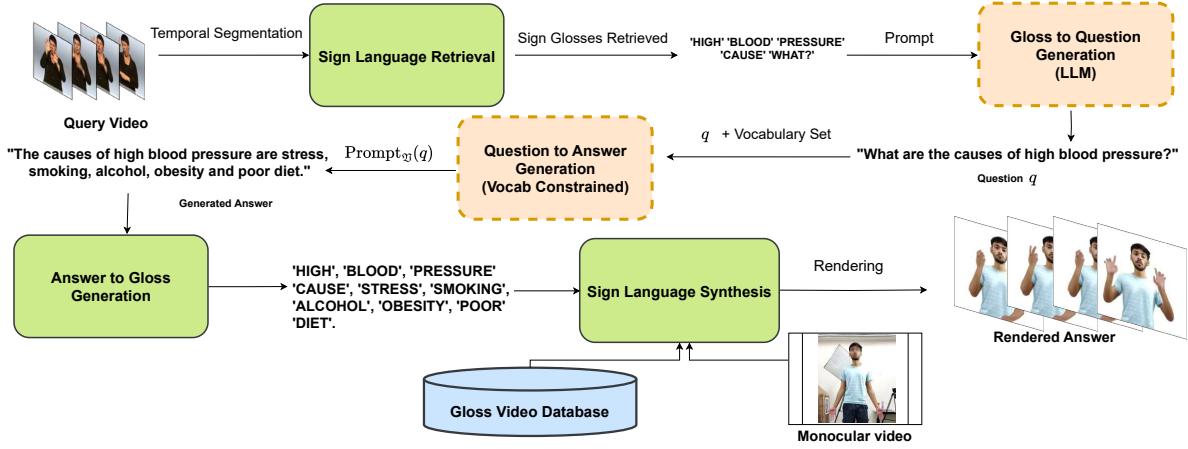


Figure 4.9 Overview of the Gloss-SLQA pipeline. The system processes a user’s sign language query through retrieval, reasons using an LLM to generate a text response, and synthesizes the final answer back into sign language using our Gloss-Centric 3D Avatar framework.

Non-Manual Markers: The current reliance on FLAME-based tracking for facial expressions fails to capture subtle non-manual cues, which are essential for full linguistic comprehension in sign language.

Regional Variation: Feedback from the Deaf community indicated slightly lower scores in Semantic Alignment, likely due to regional dialect variations and the precision of specific gloss usages that the model does not yet account for.

4.8 Discussion and Future Work

The development of the Gloss-SLQA pipeline establishes a robust baseline for inclusive digital communication, yet it opens several avenues for future research.

Vocabulary Expansion: Future iterations will focus on scaling the gloss-motion dictionary to include a wider array of low-frequency signs and diverse domain-specific and regional terminologies.

Continuous Sign Recognition: Researchers aim to move beyond semi-continuous video by developing a more robust feature embedding space capable of accurately segmenting and recognizing natural, fluid signing.

Enhanced Realism: There is a critical need to refine sign video synthesis by integrating more expressive facial and body markers to improve signing accuracy and comprehensibility for native users.

Real-World Integration: Ultimately, the goal is to explore the deployment of this pipeline in diverse applications, such as real-time public service kiosks or automated healthcare assistants, to provide the Deaf community with equitable access to information.

4.9 Summary

In this chapter, we presented a 3D avatar based framework for sign language production and demonstrated its effectiveness within a real-world application scenario. By leveraging a gloss-motion dictionary enriched with optimized SMPL-X parameters and combining it with Gaussian-based neural rendering, the proposed system achieves a strong balance between photorealism and semantic fidelity. Furthermore, the gloss-centric design overcomes the identity, viewpoint, and scalability limitations of 2D video-based approaches, offering a practical and extensible pathway toward immersive, anonymized, and domain adaptive sign language interfaces.

Chapter 5

SignWeave-3D: Weaving Isolated Sign Units into Continuous Co-Articulated Motion

5.1 Introduction

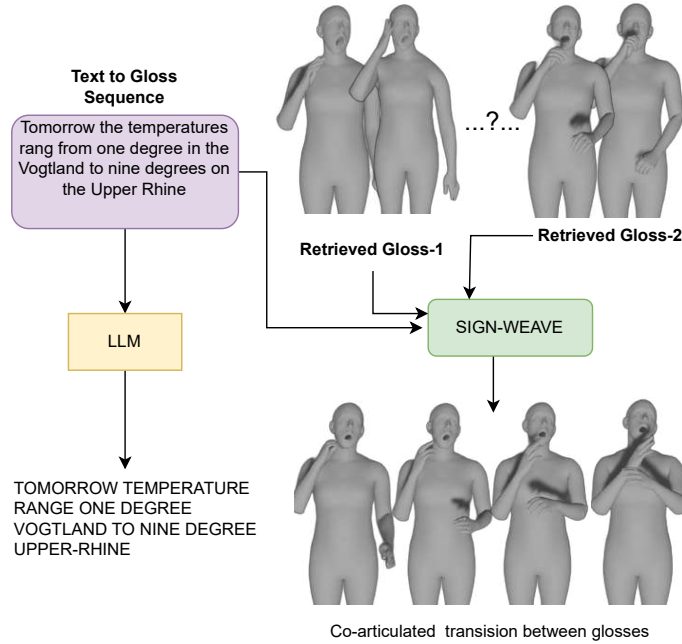


Figure 5.1 Overview: We propose a co-articulated sign language synthesis framework that integrates masked language modeling with a diffusion transformer conditioned on text. The pipeline begins with text input, followed by gloss retrieval from a dictionary and masked transition initialization. A diffusion-based model, SignWeave-3D, progressively refines the motion to produce smooth, contextually grounded sign sequences.

In the previous chapters, we established a scalable framework for Sign Language Production (SLP) by retrieving isolated motion units from a dictionary and stitching them using geometric techniques like Spherical Linear Interpolation (SLERP). While this dictionary-based approach ensures high fidelity for

individual signs, it fundamentally lacks co-articulation the complex, non-linear blending of movements, facial expressions, and body dynamics that occurs between signs in fluent speech. Naïve stitching often yields kinematic discontinuities and desynchronized non-manual markers, resulting in "robotic" output.

Conversely, end-to-end neural generation models promise fluidity but struggle with vocabulary sparsity; they often fail to generate rare or domain-specific signs accurately due to the lack of large-scale parallel training data.

To bridge this gap, we introduce SignWeave-3D, a novel hybrid framework that integrates the precision of dictionary retrieval with the contextual fluidity of generative modeling. Our core insight is to formulate sign language production as a masked motion infilling task. By retrieving accurate motion units for known glosses and using a text-conditioned diffusion model to "weave" (infill) the transitions between them, we achieve both semantic accuracy and natural temporal continuity.

5.2 Literature Review

Human Motion Synthesis. Modeling realistic human motion is a long-standing problem in computer vision and graphics, with applications in animation, robotics, embodied communication, and virtual humans [42, 117, 57, 101]. Recent advances increasingly rely on intermediate 3D pose representations such as SMPL-X [60], which provide strong structural and kinematic priors. Works including EVA3D [30], Robot Motion Diffusion Model [83], and CO3GESTURE [64] demonstrate that explicit pose-level modeling improves synthesis fidelity and controllability. These representations have enabled progress across motion prediction [49], text- and action-conditioned generation [63, 111, 14], co-speech gesture synthesis [47, 115, 65], dance generation [43, 44], and sign language synthesis [10, 96].

Text-to-Human Motion Generation. Early text-to-motion models explored latent and discrete representations to improve linguistic alignment and motion diversity. TEMOS [63] introduced a VAE-based framework for bidirectional text–motion mapping, while T2M-GPT [111] combined VQ-VAE tokenization with GPT-style transformers to produce temporally coherent motions. Diffusion probabilistic models [29] have since emerged as a dominant paradigm. The Human Motion Diffusion Model (MDM) [92] demonstrated that adapting diffusion to motion data enables diverse, temporally smooth, and semantically aligned generation, outperforming GANs [27] and VAEs. MDM has been extended to multiple domains [108, 83, 74]. Recent Diffusion Transformer (DiT) architectures [62] replace convolutional backbones with attention-based denoising. MotionDiT [98] and HumanDiT [24] further adapt DiTs for temporal motion modeling, improving long-range coherence and controllability. In parallel, masked modeling techniques, originating from NLP (e.g., BERT [15]) and vision (e.g., MAE [16]), have been applied to motion. Models such as MotionBERT [116] and MMDM [13] leverage temporal masking to enhance contextual reasoning and long-horizon motion synthesis.

Sign Language Motion Generation. Sign Language Production (SLP) has evolved from rule-based and avatar-driven systems to neural, data-driven approaches. Early pipelines decomposed the task into text-to-gloss, gloss-to-pose, and pose-to-video stages [90, 97]. While modular and scalable, these meth-

ods often suffer from limited co-articulation and temporal smoothness. Transformer-based models, such as Progressive Transformers [78], enabled direct text-to-pose generation, with further improvements through adversarial training and Mixture Density Networks [79]. More recently, diffusion-based SLP models [6, 106] have shown improved expressiveness and motion continuity. However, direct text-to-motion approaches remain constrained by limited large-scale, linguistically annotated datasets. In contrast, gloss-driven and dictionary-based methods achieve higher lexical accuracy for known vocabulary but struggle with smooth transitions [95]. Advances in parametric human modeling, particularly SMPL-X [60, 48], now enable continuous, full-body, and linguistically grounded 3D sign language generation.

5.3 Methodology

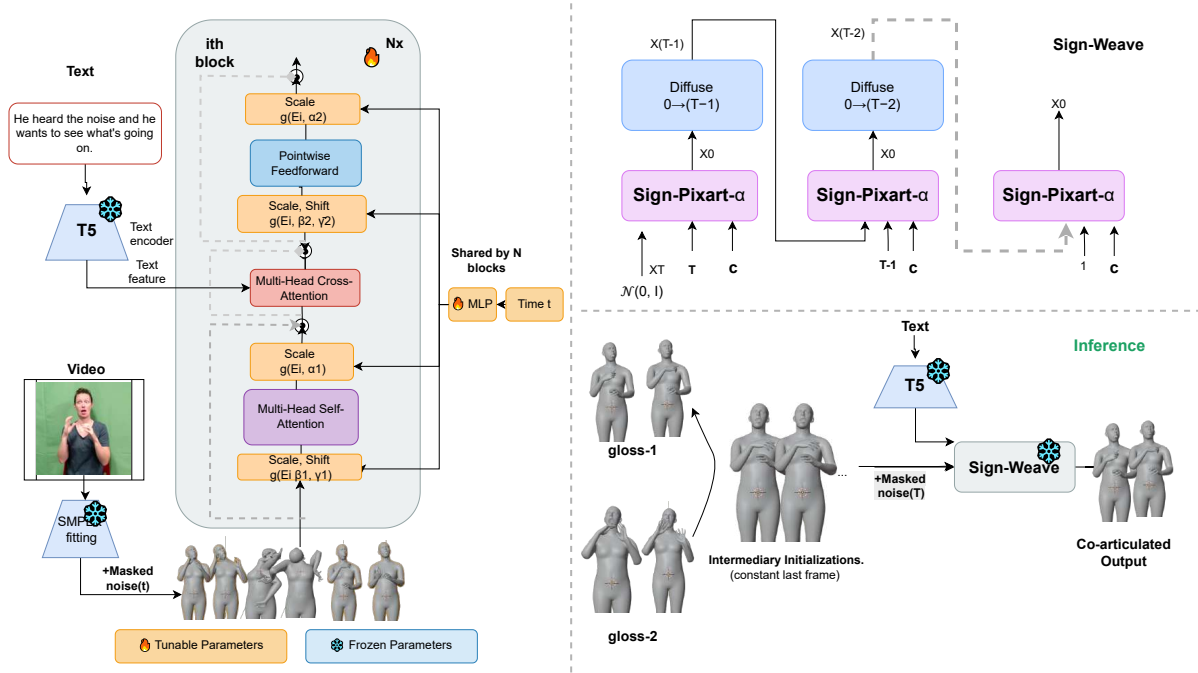


Figure 5.2 Overview of the SignWeave-3D architecture. The framework retrieves isolated glosses, initializes the sequence with masked transitions, and uses a Sign-PixArt- α diffusion backbone to progressively denoise and infill the co-articulated motion.

The SignWeave-3D framework operates by taking a text sequence, retrieving corresponding motion units, and using a specialized diffusion model to synthesize the missing transitional segments.

5.3.1 Preliminary

Diffusion-based Motion Generation: Diffusion models synthesize data by learning a denoising process that gradually reconstructs clean samples from Gaussian noise. Let x_0 denote the original motion sequence and $x_T \sim \mathcal{N}(0, I)$ represent the fully-noised latent. A Markov forward process corrupts motion as:

$$q(x_t | x_{t-1}) = \mathcal{N}(\alpha_t x_{t-1}, (1 - \alpha_t)I), \quad (5.1)$$

where $\alpha_t \in (0, 1)$ controls the noise schedule. A neural model learns the reverse process and predicts either noise or the clean signal. Following MDM, we predict the clean motion x_0 instead of noise, which allows the use of motion-domain geometric constraints such as position or velocity consistency.

The learning objective minimizes the reconstruction error over a randomly sampled noise step t :

$$\mathcal{L} = \mathbb{E}_{\mathbf{x}_0, t} \left[\|\mathbf{x}_0 - G(\mathbf{x}_t, t, \mathbf{c})\|_2^2 + \lambda_{\text{geo}} \mathcal{L}_{\text{geo}} \right].$$

where $\mathbf{x}_0$ is the ground-truth motion, $\mathbf{x}_t$ is its noisy version at timestep t , and G denotes the denoising network. the vector $\mathbf{c}$ is the conditioning signal and $\mathcal{L}_{\text{geo}}$ denotes geometric regularization encouraging smooth, physically-plausible joint motion.

Following DiT [62], MDM[13], Unlike U-Net diffusion models we adopt a Transformer denoiser, which better captures long-range temporal dependencies across joints and gloss segments. At each diffusion step, the noised motion tokens are projected into a latent space and combined with timestep encodings and text conditioning. The Transformer encoder then performs self-attention over time to model motion. For a noisy motion sequence x_t , we form token embeddings as

$$z_0 = P(x_t) + \text{PE}(t) + E(c),$$

where $P(\cdot)$ projects motion tokens, $\text{PE}(t)$ is a diffusion timestep encoding, and $E(c)$ encodes text conditioning. A Transformer $\mathcal{T}$ models temporal dependencies and co-articulation:

$$z = \mathcal{T}(z_0), \quad x_0 = \psi(z).$$

5.3.1.0.1 Motion Representation: Given the need for both expressive detail and parameter efficiency, we adopt the SMPL-X framework [60], an extension of SMPL [48] that augments the base body mesh with articulated hand and facial modeling. We use its native axis-angle representation to encode full-body poses. Specifically, each motion frame m_i is represented as a tuple:

$$m_i = \{\dot{r}_r, \dot{r}_t, \theta_r, f_s, f_e, \theta_j, \theta_b\}$$

where:

- $\dot{r}_r \in \mathbb{R}^3$: global root orientation (axis-angle),
- $\dot{r}_t \in \mathbb{R}^3$: global root translation,
- $\theta_r \in \mathbb{R}^{3N}$: local joint rotations for body and hands, where N is the number of joints,
- $f_s \in \mathbb{R}^{100}$: FLAME face shape parameters,
- $f_e \in \mathbb{R}^{50}$: FLAME face expression parameters,
- $\theta_j \in \mathbb{R}^3$: jaw rotation (axis-angle),
- $\theta_b \in \mathbb{R}^{10}$: body shape parameters (betas).

This 322-dimensional vectorized format captures rich and spatially expressive motion cues across the entire body, including facial nuance and finger articulation. For facial modeling, we follow the Motion-X representation [46] and adopt the FLAME face model [76]. This comprehensive formulation supports end-to-end modeling of sign-language motion. To support stable and identity-consistent sign language motion generation, we fix the global root orientation $\dot{r}_r$ and translation $\dot{r}_t$ to zero, and keep the body shape parameters θ_b constant across the sequence.

5.3.2 SignWeave-3D Approach

The SignWeave-3D framework is designed to learn co-articulation and synthesize continuous, context-aware sign sequences by integrating three core components: masked motion infilling, PixArt- α -based diffusion modeling, and dictionary-driven interpolation. Figure 5.2 provides a comprehensive overview of the architecture, detailing both the training and inference pipelines. Our fundamental insight is that co-articulation, the natural kinematic blending between consecutive signs can be effectively formulated as an implicit infilling task. By training the model to reconstruct masked motion spans within valid sign sequences, it learns the underlying temporal dynamics required for fluid transitions.

5.3.2.0.1 Training Strategy. During training, the input text sequence is encoded into semantic features using a pre-trained T5 encoder. Simultaneously, the corresponding ground-truth motion sequence $x_0 = \{m_1, m_2, \dots, m_T\}$, where each $m_i \in \mathbb{R}^{322}$ represents a high-fidelity SMPL-X motion vector (see Section 5.3.1.0.1), is subjected to partial masking and corrupted with Gaussian noise. A diffusion backbone, adapted from the PixArt- α [12] architecture, iteratively denoises and reconstructs the complete motion sequence from this noisy input.

To enhance temporal reasoning, we employ a *curriculum-based circular masking strategy*. The training schedule begins with a low masking ratio (10%) and progressively scales to 100%. This curriculum ensures that the model first masters the recovery of short, local spans before advancing to reconstruct longer or entirely missing segments, thereby facilitating the learning of smoother and more contextually grounded transitions between glosses.

5.3.2.0.2 Inference Pipeline. At inference, isolated gloss motion units are first retrieved from the sign dictionary. These units are temporally aligned with a predefined gap of X frames between consecutive glosses to accommodate transitions. Through empirical analysis, we determined that initializing these transition gaps using a **constant last-frame pose** strategy yields superior stability compared to other initialization methods (such as SLERP). The diffusion model then refines this coarse, initialized sequence via masked denoising steps conditioned on the text embeddings. This process effectively infills the motion gaps, enhancing temporal coherence and producing a semantically consistent, 3D co-articulated sign sequence with natural continuity across both manual and non-manual components.

5.3.3 SignWeave-3D Architecture

Our architecture advances the foundational Motion Diffusion Model (MDM) by integrating conditional generation mechanisms derived from the PixArt- α framework, specifically tailored for sign language synthesis. We substitute the standard Transformer encoder of MDM with a PixArt-style backbone comprising self-modulating Transformer blocks. These blocks utilize **Adaptive Layer Normalization (AdaLN)** to inject conditioning signals directly into the network depth.

Each block is conditioned on a fused embedding of the diffusion timestep t and the textual prompt c , which is projected via a Multi-Layer Perceptron (MLP). The AdaLN transformation applied to a hidden state h is formulated as:

$$\text{AdaLN}(h) = \gamma \cdot \text{LayerNorm}(h) + \beta,$$

where the modulation parameters (γ, β) are learned functions of the condition embedding: $(\gamma, \beta) = \text{MLP}(\text{Embed}(t, c))$. This mechanism facilitates fine-grained control over temporal dynamics and semantic alignment at every layer of the network.

To further enforce linguistic consistency, we interleave **Multi-Head Cross-Attention** layers with self-attention blocks. These layers align the latent motion tokens with the encoded text features. Given motion queries Q and textual key-value pairs (K, V) from the T5 encoder, the attention mechanism is computed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V.$$

This explicit alignment supports the generation of smooth, co-articulated transitions between manual and non-manual components, grounded in the semantic gloss input. Finally, to prioritize the fidelity of critical articulators, we incorporate a **weighted geometric loss** into the diffusion objective, assigning higher penalties to hand-joint errors. These architectural enhancements collectively improve the expressivity, smoothness, and semantic fidelity of the generated 3D sign sequences.

5.3.4 Curriculum-Based Masking Strategy

To robustly learn temporal transitions and co-articulation, we employ a **curriculum-based circular masking strategy** during training. Let $\mathbf{x}_0$ denote a batch of ground-truth motion sequences. At each

training step $s \in [1, S]$, we sample a masking ratio $r_s \in [0, 1]$ from a monotonic curriculum schedule $\{r_1, \dots, r_S\}$, where the ratio increases progressively over the course of training.

Given a ratio r_s , the number of frames to mask is determined by $L_s = \lfloor r_s \cdot T \rfloor$, where T is the sequence length. We apply circular masking by randomly sampling temporal spans $\{(t_i, \ell_i)\}$ such that their total length equals L_s . A binary mask tensor $\mathbf{M}$ is constructed (where 1 indicates a masked region), and the input sequence is corrupted with Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$:

$$\tilde{\mathbf{x}}_0 = (1 - \mathbf{M}) \odot \mathbf{x}_0 + \mathbf{M} \odot \epsilon.$$

This curriculum ensures that the model first masters the reconstruction of short, unambiguous segments before progressing to long-range dependencies and fully missing glosses, thereby fostering the ability to infer context-dependent transitions.

5.3.5 Sign Stitching via Dictionary Retrieval

The inference pipeline leverages our dictionary of isolated motion units to construct continuous signing. Given an input text sequence t and its corresponding gloss sequence $\mathbf{g} = \{g_1, g_2, \dots, g_M\}$, we retrieve the high-fidelity SMPL-X motion vectors $\{m_i\}_{i=1}^M$ associated with each gloss.

To ensure clean boundaries, each retrieved clip is trimmed based on a hand-activity heuristic derived from MediaPipe landmarks: frames are retained only if the vertical distance between the hands and the torso center exceeds a threshold τ_h . The refined clips are then concatenated to form a coarse, pseudo-continuous sequence.

To facilitate co-articulation, we insert a fixed transition gap of L frames between consecutive signs. In cases where a gloss g_i is out-of-vocabulary (missing), we allocate a larger interpolation span $L' > L$ to preserve the rhythmic pacing of the sentence. We initialize these gaps using a **constant last-frame pose** strategy (repeating the final pose of the preceding sign). A binary mask $\mathbf{M}$ is generated to identify these transition gaps and missing regions.

Finally, the SignWeave-3D diffusion model performs **masked denoising** conditioned on the text embeddings and the partial observation $\mathbf{M}$. This process effectively "inpaints" the missing transitions, synthesizing smooth, context-aware co-articulation that naturally bridges the retrieved dictionary units while preserving their high-fidelity articulation.

5.4 Experiments and Results

To validate the efficacy of the **SignWeave-3D** framework, we conducted a comprehensive evaluation focusing on two primary objectives: (1) assessing the quality of text-driven sign language generation against state-of-the-art motion synthesis baselines, and (2) evaluating the specific capability of our masked infilling approach to generate smooth, realistic co-articulation between isolated signs.

5.4.1 Experimental Setup

5.4.1.1 Dataset and Preprocessing

We utilized the **PHOENIX-2014T** dataset [9], a standard benchmark for continuous German Sign Language (DGS). It comprises a vocabulary of approximately 1,066 glosses aligned with 8,257 German text sentences. To ensure compatibility with standard pre-trained text encoders (e.g., T5, CLIP), we translated the German text prompts into English.

Since PHOENIX-2014T lacks ground-truth 3D annotations, we generated a unified 3D representation using the **SMPL-X** parametric model. Our pipeline involved:

1. **Initialization:** Predicting per-frame SMPL-X parameters using SMPLer-X [8].
2. **Refinement:** Optimizing the body alignment against 2D keypoints extracted via MMPose [81], utilizing an objective function inspired by Ex-Avatar [55] to ensure spatial consistency.
3. **Facial Enhancement:** Replacing coarse facial parameters with high-fidelity codes derived from the FLAME expression basis [71], capturing the nuanced non-manual markers critical for sign language.

5.4.1.2 Implementation Details

To rigorously evaluate the co-articulation capabilities, we designed controlled settings that simulate the transition between isolated signs. We employed a masked-motion prediction protocol with three difficulty levels:

- **Short-Range:** Predict 10 frames given 10 observation frames (Obs 10 $\rightarrow$ Pred 10).
- **Medium-Range:** Predict 20 frames given 10 observation frames (Obs 10 $\rightarrow$ Pred 20).
- **Long-Range:** Predict 30 frames given 10 observation frames (Obs 10 $\rightarrow$ Pred 30).

These settings progressively challenge the model to infer longer temporal dependencies and recover naturalistic transitions from limited context.

5.4.2 Evaluation Metrics

We utilized a multi-dimensional evaluation framework to assess the system’s ability to maintain structural motion integrity while ensuring high-fidelity semantic mapping from text to sign.

- **Geometric Error:** We report **Mean Per-Joint Position Error (MPJPE)** and **Procrustes-Aligned MPJPE (PA-MPJPE)** to measure spatial accuracy. Additionally, **Velocity Error (Vel)** quantifies the temporal smoothness and dynamic realism of the motion.

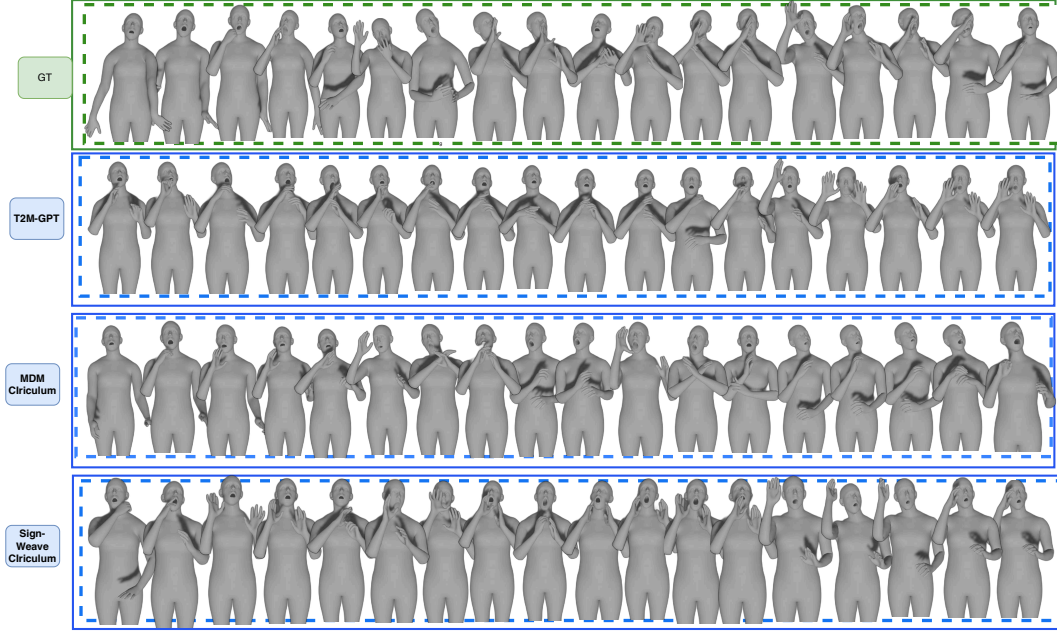


Figure 5.3 Qualitative comparison of text-to-3D sign language motion generation. Each row shows temporal frame sequences for T2M-GPT, MDM Curriculum, and SignWeave-3D Curriculum, with GT (Ground Truth) on top for reference.

- **Semantic Retrieval (MR-Precision):** Following SignAvatar [106], we evaluate semantic consistency via retrieval. Generated sequences are matched against a gallery containing the ground truth and 15 distractor sequences. We report Top-1, Top-3, and Top-5 retrieval accuracy.
- **Temporal Alignment (DTW):** We compute the **Dynamic Time Warping (DTW)** distance on joint trajectories to assess how well the generated temporal dynamics align with the ground truth, independent of sequence length.

5.4.3 Comparison with Text-to-Motion Baselines

Since no open-source frameworks currently exist for text-to-3D sign language generation, we adapted two state-of-the-art general motion generation models **T2M-GPT** [111] and **MDM** [92] to the sign language domain.

Method (GSL)	Hand							Face							HandArm							Overall						
	MPJPE	PA	Vel	Top-1	Top-3	Top-5	DTW	MPJPE	PA	Vel	Top-1	Top-3	Top-5	DTW	MPJPE	PA	Vel	Top-1	Top-3	Top-5	DTW	MPJPE	PA	Vel	Top-1	Top-3	Top-5	DTW
SignWeave-3D-T2M (Ours)	0.241	0.206	2.019	0.21	0.44	0.60	0.086	0.038	0.038	0.284	0.19	0.43	0.58	0.203	0.223	0.206	1.869	0.20	0.40	0.56	0.100	0.092	0.122	0.751	0.21	0.44	0.60	0.086
MDM-T2M	0.259	0.207	2.236	0.20	0.40	0.57	0.079	0.041	0.041	0.309	0.21	0.40	0.56	0.180	0.240	0.211	2.069	0.20	0.44	0.61	0.093	0.099	0.130	0.829	0.20	0.40	0.57	0.079
T2M-GPT	0.239	0.189	2.083	0.10	0.29	0.45	0.1426	0.036	0.035	0.283	0.07	0.23	0.39	0.0194	0.221	0.192	1.923	0.10	0.28	0.45	0.1327	0.090	0.118	0.768	0.11	0.27	0.43	0.0573

Table 5.1 Quantitative comparison of text-to-sign motion generation models. SignWeave-3D demonstrates superior semantic alignment (Top-k Retrieval) compared to baselines while maintaining competitive geometric fidelity. Note the significant drop in T2M-GPT’s retrieval performance, indicating a lack of semantic precision.

Table 5.1 presents the quantitative comparison. **SignWeave-3D** demonstrates a superior balance between motion quality and semantic precision. Notably:

- **Semantic Alignment:** Our method achieves the highest retrieval accuracy (Top-1: 0.21, Top-5: 0.60), significantly outperforming T2M-GPT (Top-1: 0.11). This indicates that our dictionary-guided approach preserves the distinct linguistic meaning of signs, whereas T2M-GPT tends to generate generic, over-smoothed motion.
- **Temporal Consistency:** While T2M-GPT achieves a lower DTW (0.057), this is often an artifact of collapsing to mean poses. Our slightly higher DTW (0.086) reflects the preservation of realistic, high-frequency temporal variations found in natural signing.

Qualitatively, Figure 5.3 shows that SignWeave-3D produces smoother temporal transitions and more coherent hand-face coordination compared to MDM, which exhibits visible jitter at segment boundaries.

5.4.4 Evaluating Co-Articulation and Stitching

To isolate the model’s ability to generate realistic transitions (co-articulation), we compared our approach against MDM under identical curriculum training strategies.

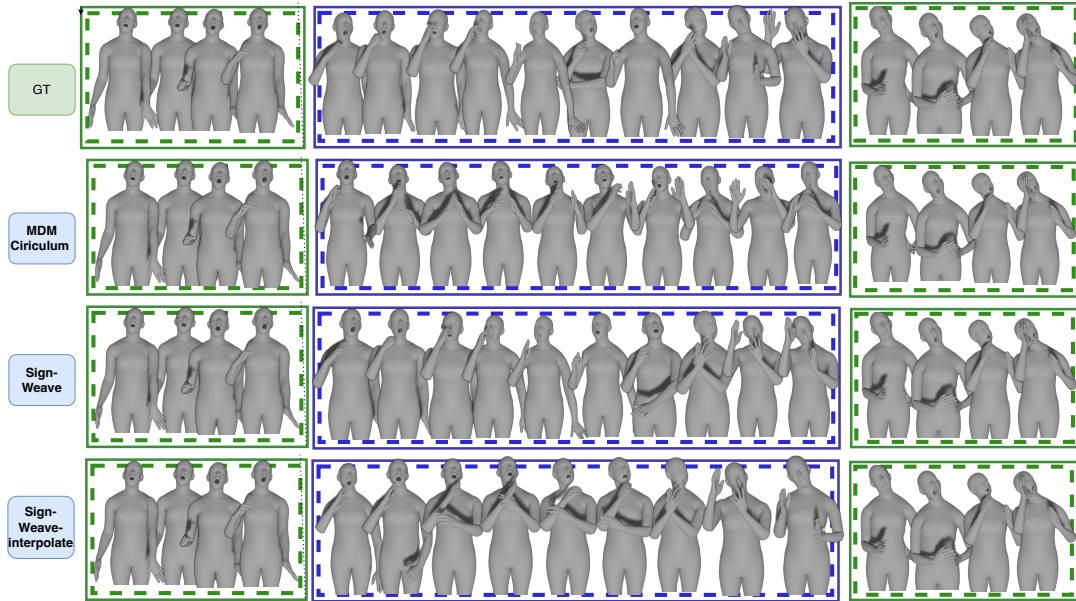


Figure 5.4 Qualitative comparison on co-articulation prediction. Models predict 30 frames (blue boxes) given observed context (green boxes). MDM (middle) shows visible pose shifts at boundaries. SignWeave-3D (bottom) generates smooth, continuous transitions that naturally blend the context.

5.4.4.1 Results Analysis

Table 5.2 details the performance in the most challenging long-horizon setting (Obs 10 $\rightarrow$ Pred 30).

Method (GSL) w/ Curriculum	Setting	Hand (MR- Precision)				Face (MR- Precision)				HandArm (MR- Precision)				Overall (MR- Precision)			
		Top-1	Top-3	Top-5	AvgDTW	Top-1	Top-3	Top-5	AvgDTW	Top-1	Top-3	Top-5	AvgDTW	Top-1	Top-3	Top-5	AvgDTW
MDM [Const]	Obs 10 → Pred 10	0.90	0.96	0.99	0.1166	0.75	0.91	0.95	0.0325	0.91	0.96	0.99	0.1094	0.95	0.99	1.00	0.0596
MDM [Const]	Obs 10 → Pred 20	0.85	0.95	0.97	0.1366	0.76	0.86	0.92	0.0188	0.84	0.94	0.97	0.1273	0.89	0.96	0.98	0.0544
MDM [Const]	Obs 10 → Pred 30	0.71	0.88	0.93	0.1505	0.55	0.74	0.85	0.0216	0.72	0.89	0.94	0.1403	0.77	0.92	0.96	0.0611
SignWeave-3D [Interp]	Obs 10 → Pred 10	0.87	0.94	0.96	0.1163	0.86	0.93	0.97	0.0141	0.88	0.95	0.97	0.1081	0.92	0.96	0.97	0.0445
SignWeave-3D [Interp]	Obs 10 → Pred 20	0.63	0.17	0.07	0.1498	0.64	0.17	0.08	0.0187	0.64	0.17	0.08	0.1391	0.72	0.14	0.05	0.0686
SignWeave-3D [Interp]	Obs 10 → Pred 30	0.39	0.23	0.11	0.1665	0.39	0.23	0.11	0.0212	0.40	0.22	0.12	0.1549	0.47	0.23	0.10	0.0769
SignWeave-3D (Const)(ours)	Obs 10 → Pred 10	0.89	0.95	0.98	0.1152	0.74	0.92	0.97	0.0322	0.89	0.95	0.98	0.1081	0.95	0.98	1.00	0.0590
SignWeave-3D (Const)(ours)	Obs 10 → Pred 20	0.84	0.95	0.96	0.1327	0.76	0.89	0.93	0.0178	0.85	0.95	0.97	0.1235	0.89	0.94	0.97	0.0527
SignWeave-3D (Const)(ours)	Obs 10 → Pred 30	0.72	0.90	0.93	0.1469	0.59	0.79	0.86	0.0203	0.75	0.91	0.94	0.1370	0.80	0.93	0.96	0.0591
Method (GSL)	Setting	Hand			Face			HandArm			Overall						
		MPJPE	PA-MPJPE	Vel	MPJPE	PA-MPJPE	Vel	MPJPE	PA-MPJPE	Vel	MPJPE	PA-MPJPE	Vel				
MDM [Const]	Obs 10 → Pred 10	0.091	0.079	1.173	0.013	0.013	0.163	0.084	0.079	1.085	0.040	0.054	0.514				
MDM [Const]	Obs 10 → Pred 20	0.145	0.101	1.206	0.017	0.016	0.163	0.134	0.105	1.117	0.052	0.068	0.445				
MDM [Const]	Obs 10 → Pred 30	0.173	0.144	1.582	0.027	0.026	0.224	0.160	0.145	1.464	0.066	0.086	0.589				
SignWeave-3D [Interp]	Obs 10 → Pred 10	0.107	0.090	1.310	0.014	0.014	0.174	0.099	0.091	1.210	0.046	0.063	0.568				
SignWeave-3D [Interp]	Obs 10 → Pred 20	0.169	0.140	1.628	0.024	0.024	0.219	0.156	0.141	1.504	0.074	0.099	0.707				
SignWeave-3D [Interp]	Obs 10 → Pred 30	0.202	0.165	1.779	0.029	0.029	0.239	0.186	0.167	1.644	0.089	0.117	0.773				
SignWeave-3D (Const)(Ours)	Obs 10 → Pred 10	0.088	0.078	1.153	0.012	0.012	0.159	0.081	0.077	1.066	0.039	0.053	0.504				
SignWeave-3D (Const)(Ours)	Obs 10 → Pred 20	0.114	0.092	1.362	0.015	0.015	0.183	0.105	0.094	1.259	0.042	0.056	0.501				
SignWeave-3D (Const)(Ours)	Obs 10 → Pred 30	0.168	0.139	1.599	0.026	0.026	0.224	0.156	0.141	1.479	0.064	0.084	0.594				

Table 5.2 Evaluation of curriculum-based prediction under varying observation lengths (obs). The top table reports multi-reference (MR) precision retrieval scores and AvgDTW across Hand, Face, Hand+Arm, and Overall regions for both MDM and SignWeave-3D models across different prediction lengths (pred) (10, 20, 30 frames). The bottom table presents corresponding geometric accuracy metrics, MPJPE, PA-MPJPE, and velocity error for the same setups. Interp: SLERP [87] interpolated frames initialization, Const: constant last frame initialization

- **Performance Gain:** SignWeave-3D (Const) consistently outperforms MDM, improving Overall Top-1 accuracy from 0.77 to **0.80** and reducing MPJPE from 0.066 to **0.064**. Similar gains are observed in hand and face subsets, highlighting the model’s ability to maintain fine-grained control over long sequences.
- **Visual Fidelity:** Qualitative analysis in Figure 5.4 corroborates these metrics. MDM often introduces discontinuities or “drift” at transition boundaries. In contrast, SignWeave-3D produces seamless connections, effectively “weaving” the retrieved glosses into a unified sequence.

5.4.5 Ablation Study: Inference-Time Initialization

We investigated the impact of the initialization strategy for the transition gaps. We compared standard **SLERP Interpolation** against our proposed **Constant Last-Frame** strategy.

- **Short-Range (Pred 10):** SLERP performs competitively, leveraging its smooth transition prior (Top-1: 0.92 vs 0.95).
- **Long-Range (Pred 30):** As the prediction window expands, SLERP suffers catastrophic degradation (Top-1 drops to 0.47) due to accumulated interpolation drift. Conversely, Constant Initialization remains robust, maintaining a Top-1 accuracy of 0.80.

This ablation confirms that for generative infilling of long gaps, a stable anchor (constant frame) is decisively superior to a drifting geometric prior (SLERP).

5.4.6 Comparing MR Precision

Method	Evaluator	MR-P@1	MR-P@3	MR-P@5
PT	EPT (PT-fitted keypoints)	0.088	0.145	0.212
SignAvatars	EPT (PT-fitted keypoints)	0.219	0.325	0.443
SignWeave-3D (ours)	DTW-based	0.210	0.440	0.600
SignAvatars	ESA (SignAvatars-fitted evaluator)	0.497	0.691	0.753
SignWeave-3D (ours, gloss-based)	DTW-based	0.950	0.980	1.000

Table 5.3 MR-Precision on PHOENIX-2014T based on different evaluators. EPT is an evaluator trained by SignAvatars on Progressive Transformers [78] using 3D SMPL-X representations. ESA is an evaluator trained by SignAvatars on sign-avatar-fitted embeddings.

MR-Precision on PHOENIX-2014T (Text-to-Sign). On PHOENIX-2014T, SignAvatars reports MR-Precision using embedding distances under the Progressive Transformer evaluator (EPT), trained on PT-fitted keypoints. Under this protocol, PT achieves 0.088 / 0.145 / 0.212, while SignAvatars (EPT) reports 0.219 / 0.325 / 0.443. In contrast, our method achieves 0.21 / 0.44 / 0.60 using DTW directly on joint trajectories. Although the distance functions differ, the consistent improvement over PT and SignAvatars under the EPT Protocols indicate stronger retrieval consistency and temporal coherence in pose space. SignAvatars further reports higher MR-Precision under an improved evaluator (ESA: 0.497 / 0.691 / 0.753), reflecting evaluator–representation alignment, whereas our DTW-based evaluation measures motion similarity using a model-agnostic, joint-level distance. We note that the evaluators used in SignAvatars (CPT and ESA) are not publicly released, preventing direct re-evaluation under the same embedding-based protocols. Finally, SignWeave-3D highlights the effectiveness of explicit 3D gloss-level stitching, achieving MR-Precision of 0.95/0.98/1.00.

5.4.7 Human Evaluation

To complement quantitative metrics, we conducted a human evaluation to assess the perceived naturalness of motion in automatically generated Sign Language Production. Participants were shown three sentence-level clips, corresponding to the Obs 10 → Pred 10, Obs 10 → Pred 20, and Obs 10 → Pred 30 co-articulation settings (as mentioned in paper). For each setting, we presented one example from every method in our comparison suite: direct stitching, repeated last-frame initialization, SLERP interpolation, MDM (baseline), SignWeave-3D with interpolation initialization, and SignWeave-3D with constant-frame initialization (our final model). All clips were generated from the same gloss sequence to ensure a controlled, sentence-matched comparison.

Participants rated each clip independently using a 5-point Likert scale with the labels *Very Poor*, *Poor*, *Acceptable*, *Good*, and *Excellent*. Each video was evaluated on the following four perceptual questions: (1) *Hand movement trajectories are smooth and non-jerky*. (2) *Transitions between signs are fluid without abrupt jumps*. (3) *The signing looks natural and human-like*. The videos were in German

Method (GSL)	Smooth Hand Trajectory			Fluid Sign Transitions			Natural / Human-like Signing		
	Obs10→Pred10	Obs10→Pred20	Obs10→Pred30	Obs10→Pred10	Obs10→Pred20	Obs10→Pred30	Obs10→Pred10	Obs10→Pred20	Obs10→Pred30
Naïve Stitching Baselines									
Direct Stitching	2.23	2.25	1.61	2.00	2.00	1.54	2.23	2.17	1.62
Repeated Last-Frame Init	1.92	2.15	1.69	1.54	1.38	1.46	1.77	2.00	1.62
SLERP Interpolation	3.38	3.08	2.62	3.00	2.67	2.15	2.92	2.25	1.85
MLM-Based Methods									
MDM (MLM Baseline)	3.38	3.08	3.25	3.15	2.85	2.83	3.38	2.92	2.58
SignWeave-3D (Interp Init)	4.08	3.38	3.31	3.69	3.38	2.69	3.54	3.31	2.69
SignWeave-3D (Const Init) (Ours)	3.31	3.69	3.31	3.15	3.54	3.46	3.23	3.23	3.08

Table 5.4 Human evaluation (MOS) across four perceptual criteria (smooth trajectories, fluid transitions, naturalness) and three prediction horizons (Obs10→Pred10/Pred20/Pred30). Scores reflect the mean rating from 13 hearing participants on a 1–5 Likert scale (higher is better) for naïve stitching, MLM-based baselines, and the proposed SignWeave-3D models.

Sign Language (GSL), but participants were informed that knowledge of GSL was not required, as the evaluation targets motion fluency rather than language comprehension. All responses were collected anonymously.

Table 5.4 reports the Mean Opinion Scores from our perceptual study, in which 13 hearing participants rated generated videos using a 1–5 Likert scale across three prediction horizons (Obs10→Pred10, Obs10→Pred20, Obs10→Pred30) and three criteria relevant to signing quality: smooth hand trajectories, fluid sign transitions, and natural/human-like motion. The naïve stitching baselines obtain the lowest ratings and deteriorate rapidly as the prediction span increases, indicating that geometric concatenation alone cannot sustain fluency. SLERP interpolation performs better for short spans but drops substantially for longer ones, showing that local smoothing does not recover linguistic co-articulation. Among MLM-based approaches, MDM improves over stitching but still loses perceptual coherence as transitions accumulate. SignWeave-3D with interpolation initialization performs competitively for short horizons but declines as prediction length grows, consistent with its reliance on geometric seeding. Our full SignWeave-3D with constant initialization receives the highest and most stable MOS across all horizons and criteria, demonstrating that the model synthesizes transitions based on global sentence-level context rather than short-range pose smoothing.

A critical finding was the superiority of constant last-frame initialization during inference. While standard SLERP interpolation provides a smooth geometric baseline for short spans, it introduces “interpolation drift” in complex signing sequences. In contrast, our constant initialization allows the diffusion backbone to rely purely on the learned co-articulatory dynamics and text conditioning to “weave” signs together, ensuring transitions remain linguistically grounded rather than just mathematically smoothed.

5.5 Limitations and Future Work

Despite the advancements in motion continuity, several constraints persist that define our roadmap for future work:

Vocabulary Dependency: While the system generalizes to out-of-distribution words, its highest fidelity is still contingent on the availability of high-quality isolated signs within the dictionary. Future work will focus on expanding these symbolic resources to cover broader linguistic domains.

Non-Manual Marker Synchronization: Although the use of SMPL-X and FLAME models captures facial nuances, the precise temporal synchronization of subtle non-manual markers—such as eyebrow articulation or mouthings with rapid hand gestures remains a challenge in fully continuous scenarios.

Temporal Variability: The framework currently utilizes pre-defined frame gaps for interpolation. Natural signing exhibits significant rhythmic variations and prosodic emphases that are not yet fully captured by fixed-length transitions. Integrating prosody-aware temporal modeling is a key area for future development.

Computational Complexity: The iterative denoising process of the diffusion transformer requires higher computational overhead than traditional non-autoregressive or GAN-based models. Optimizing the inference speed for real-time applications is a priority for subsequent iterations of the framework.

5.6 Discussion

The results of this study validate the effectiveness of the SignWeave-3D framework in bridging the gap between isolated sign retrieval and continuous motion synthesis. Our approach confirms that co-articulation can be successfully modeled as an implicit motion infilling task when combined with a curriculum-based masking strategy. By training the model to reconstruct increasingly larger masked spans, it learns to prioritize local joint constraints before mastering the global temporal dependencies required for long-horizon transitions. Our extensive experiments demonstrate that this hybrid approach effectively resolves the trade-off between vocabulary coverage and motion fluency. By significantly outperforming baselines in both semantic retrieval and geometric fidelity, SignWeave- 3D establishes a scalable and robust pathway for naturalistic 3D sign language generation.

Chapter 6

Conclusion

This thesis advances the field of Sign Language Production by addressing the fundamental challenges of data scarcity, motion continuity, and visual fidelity.

6.0.1 Impact

This thesis aims to advance the field of Inclusive Communication Technologies by providing scalable and efficient approaches for digital sign language synthesis. The proposed frameworks can be of great use for:

Public Accessibility: potential for automated sign language support for educational material, emergency alerts, and government proceedings.

Interactive Systems: Powering sign-language-based question-answering systems (Gloss-SLQA) that function as a "Siri for Sign Language".

Low-Resource Languages: Enabling the development of SLP for languages such as Indian Sign Language (ISL) where massive training datasets do not exist.

Immersive Environments: Utilization of 3D avatars for privacy-preserving telepresence and AR/VR applications where identity anonymization and viewpoint independence are required.

By releasing the Sanket Vaani-1K dataset, we provide a foundational resource comprising 1,000 ISL sentences across 40 domains to aid the sign-language research community.

Quantitative evaluations and user studies with the Deaf community confirmed that the proposed framework produces signing that is significantly more understandable, fluent, and natural than current state-of-the-art baselines.

6.0.2 Discussion and Limitations

Evaluating sign language generation is challenging due to its multimodal and linguistic complexities. Existing metrics mainly rely on frame-wise pose errors (e.g., MPJPE) or adapted surface-level similarity scores (e.g., BLEU), which fail to capture linguistic accuracy, grammatical structure, and semantic fidelity. They are largely insensitive to non-manual markers, coarticulation, and temporal alignment

between manual and non-manual channels, and do not account for high variability across signers, regions, and contexts. Consequently, models may achieve strong quantitative scores while producing signs that appear unnatural, ambiguous, or linguistically incorrect to deaf signers, underscoring the need for linguistically grounded, perception-aware, and human-in-the-loop evaluation protocols tailored to sign language generation.

Lack of Data: Existing datasets are often small, domain-specific, or lack detailed linguistic annotations, particularly for non-manual markers and discourse-level structures. This scarcity restricts robust training, limits generalization across signers and regions, and hampers the development of reliable and linguistically informed evaluation metrics.

Limitations of 2D and Need for More Expressive 3D Parametric Models: 2D-based sign language representations are inherently constrained by occlusions, depth ambiguity, and viewpoint dependence, limiting accurate modeling of complex hand-body interactions. Although 3D parametric models such as SMPL-X address some of these issues, they remain insufficiently expressive to capture fine-grained hand articulations, detailed facial expressions, and subtle coarticulation effects critical for sign language.

Moreover, existing models typically generate signs in a locally conditioned manner, with limited ability to propagate linguistic context across time or maintain discourse-level references. Consequently, while individual signs or short segments may seem plausible, longer sign sequences often fail to maintain semantic coherence, consistent referent tracking, and grammatical structure.

AI-driven accessibility technologies have the potential to empower deaf communities by improving access to information, education, and services through sign language generation and translation. However, ethical concerns arise when such systems are positioned as replacements rather than supports for human interpreters, risking job displacement, loss of cultural nuance, and overreliance on imperfect models. Responsible deployment requires framing AI as an assistive tool that augments human expertise, respects linguistic diversity, ensures transparency and consent, and actively involves deaf communities in design, evaluation, and governance.

6.0.3 Future Work

Disambiguating Self-Occlusion: In sign language, hands often cross the face or torso. Multiview data mitigate these "invisible" regions by capturing the articulation from alternative angles, preventing the "face flipping" or "finger twists" common in under-constrained monocular fits.

Richer and More Expressive Parametric Representations: While SMPL-X provides structural realism and viewpoint independence, it remains limited in capturing fine-grained finger articulation, facial dynamics, and non-manual cues essential for sign language. Future work will explore richer or hybrid parametric models that explicitly model hands, face, and upper-body dynamics to improve linguistic expressiveness.

Context-Aware and Discourse-Level Sign Generation: Current systems primarily rely on local conditioning, which limits discourse-level coherence. Future work will incorporate long-range context

modeling to support consistent reference tracking, spatial indexing, and grammatical continuity across extended sign sequences.

Scalable Data Expansion and Annotation: Although SanketVaani-1K offers a strong foundation, broader vocabulary coverage, signer diversity, and discourse complexity are required. Future work will explore semi-automatic and weakly supervised annotation strategies for efficient data collection.

Publications

1. **Aparna Agrawal**, seshadri Mazumdar, C.V.Jawahar, Vinay Namboodri “**Towards Scalable Sign Production: Leveraging Co-Articulated Gloss Dictionary for Fluid Sign Synthesis**”, in proceedings of *Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP)*, 2025.
2. **Aparna Agrawal***, seshadri Mazumdar*, C.V.Jawahar, Vinay Namboodri “**Gloss SLQA: A Gloss-centric Approach to Sign Language Question Answering**”, under Review at *Pattern Recognition*, 2026.
3. **Aparna Agrawal**, C.V.Jawahar, Vinay Namboodri “**SignWeave-3D: Weaving Isolated Sign Units into Continuous Co-Articulated Motion**”, under review at *ACM International Conference on Multimedia* 2026.

Other Publications

- Jayasree Saha, Rudrabha Mukhopadhyay, **Aparna Agrawal**, Surabhi Jain, C. V. Jawahar; *An Approach for Speech Enhancement in Low SNR Environments using Granular Speaker Embedding to Enhance ASR*; in proceedings of *CODS-COMAD '24*.

Bibliography

- [1] F. A. Acheampong, H. Nunoo-Mensah, and W. Chen. Transformer models for text-based emotion detection: a review of bert-based approaches. *Artificial Intelligence Review*, 54(8):5789–5829, 2021.
- [2] S. Albanie, G. Varol, L. Momeni, T. Afouras, J. S. Chung, N. Fox, and A. Zisserman. Bsl-1k: Scaling up co-articulated sign language recognition using mouthing cues. In *European conference on computer vision*, pages 35–53. Springer, 2020.
- [3] S. Albanie, G. Varol, L. Momeni, H. Bull, T. Afouras, H. Chowdhury, N. Fox, B. Woll, R. Cooper, A. McParland, et al. Bbc-oxford british sign language dataset. *arXiv preprint arXiv:2111.03635*, 2021.
- [4] Z. An and R. Kawakami. Teach me sign: Stepwise prompting llm for sign language production. *arXiv preprint arXiv:2507.10972*, 2025.
- [5] M. Aziz and A. Othman. Evolution and trends in sign language avatar systems: Unveiling a 40-year journey via systematic review. *Multimodal Technologies and Interaction*, 7(10):97, 2023.
- [6] V. Baltatzis, R. A. Potamias, E. Ververas, G. Sun, J. Deng, and S. Zafeiriou. Neural sign actors: A diffusion model for 3d sign language production from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1985–1995, 2024.
- [7] E. Betzalel, C. Penso, and E. Fetaya. Evaluation metrics for generative models: An empirical study. *Machine Learning and Knowledge Extraction*, 6(3):1531–1544, 2024.
- [8] Z. Cai, W. Yin, A. Zeng, C. Wei, Q. Sun, W. Yanjun, H. E. Pang, H. Mei, M. Zhang, L. Zhang, et al. Smpler-x: Scaling up expressive human pose and shape estimation. *Advances in Neural Information Processing Systems*, 36:11454–11468, 2023.
- [9] N. C. Camgoz, S. Hadfield, O. Koller, H. Ney, and R. Bowden. Neural sign language translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7784–7793, 2018.
- [10] N. C. Camgoz, O. Koller, S. Hadfield, and R. Bowden. Sign language transformers: Joint end-to-end sign language recognition and translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10023–10033, 2020.
- [11] N. C. Camgöz, B. Saunders, G. Rochette, M. Giovanelli, G. Inches, R. Nachtrab-Ribback, and R. Bowden. Content4all open research sign language translation datasets. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–5. IEEE, 2021.
- [12] J. Chen, J. Yu, C. Ge, L. Yao, E. Xie, Y. Wu, Z. Wang, J. Kwok, P. Luo, H. Lu, et al. Pixart-

\

alpha : Fasttrainingofdiffusiontransformerforphotorealistictext – to – imagesynthesis.*arXivpreprintarXiv* : 2310.00426, 2023.

- [13] X. Chen. Text-driven human motion generation with motion masked diffusion model. *arXiv preprint arXiv:2409.19686*, 2024.

- [14] W. Dai, L.-H. Chen, J. Wang, J. Liu, B. Dai, and Y. Tang. Motionlcm: Real-time controllable motion generation via latent consistency model. In *European Conference on Computer Vision*, pages 390–408. Springer, 2024.
- [15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.
- [16] X. Dong, J. Bao, T. Zhang, D. Chen, W. Zhang, L. Yuan, D. Chen, F. Wen, and N. Yu. Bootstrapped masked autoencoders for vision bert pretraining. In *European Conference on Computer Vision*, pages 247–264. Springer, 2022.
- [17] M. Douze, A. Guzhva, C. Deng, J. Johnson, G. Szilvasy, P.-E. Mazaré, M. Lomeli, L. Hosseini, and H. Jégou. The faiss library. 2024.
- [18] P. Dreuw, C. Neidle, V. Athitsos, S. Sclaroff, and H. Ney. Benchmark databases for video-based automatic sign language recognition. In *Proceedings of Language Resources and Evaluation Conference (LREC) 2008*. EUROPEAN LANGUAGE RESOURCES ASSOC-ELRA, 2008.
- [19] A. Duarte, S. Palaskar, L. Ventura, D. Ghadiyaram, K. DeHaan, F. Metze, J. Torres, and X. Giro-i Nieto. How2sign: a large-scale multimodal dataset for continuous american sign language. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2735–2744, 2021.
- [20] A. Dutta and A. Zisserman. The via annotation software for images, audio and video. In *Proceedings of the 27th ACM international conference on multimedia*, pages 2276–2279, 2019.
- [21] S. Fang, C. Sui, X. Zhang, and Y. Tian. Signdiff: Learning diffusion models for american sign language production. *arXiv preprint arXiv:2308.16082*, 2023.
- [22] J. Forster, C. Schmidt, O. Koller, M. Bellgardt, and H. Ney. Extensions of the sign language recognition and translation corpus rwth-phoenix-weather. In *LREC*, pages 1911–1916, 2014.
- [23] M.-P. Forte, P. Kulits, C.-H. P. Huang, V. Choutas, D. Tzionas, K. J. Kuchenbecker, and M. J. Black. Reconstructing signing avatars from video using linguistic priors. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 12791–12801, June 2023.
- [24] Q. Gan, Y. Ren, C. Zhang, Z. Ye, P. Xie, X. Yin, Z. Yuan, B. Peng, and J. Zhu. Humandit: Pose-guided diffusion transformer for long-form human motion video generation. *arXiv preprint arXiv:2502.04847*, 2025.
- [25] A. Ghosh and R. Mamidi. 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, 2022.
- [26] A. Ghosh and R. Mamidi. English to Indian Sign Language:rule-based translation system along with multi-word expressions and synonym substitution. In M. S. Akhtar and T. Chakraborty, editors, *Proceedings of the 19th International Conference on Natural Language Processing (ICON)*, pages 123–127, Dec. 2022.
- [27] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.
- [28] J. Guo, P. Li, and T. Cohn. Bridging sign and spoken languages: Pseudo gloss generation for sign language translation. *arXiv preprint arXiv:2505.15438*, 2025.
- [29] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.
- [30] F. Hong, Z. Chen, Y. Lan, L. Pan, and Z. Liu. Eva3d: Compositional 3d human generation from 2d image collections. *arXiv preprint arXiv:2210.04888*, 2022.
- [31] L. Hu, H. Zhang, Y. Zhang, B. Zhou, B. Liu, S. Zhang, and L. Nie. Gaussianavatar: Towards realistic human avatar modeling from a single video via animatable 3d gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 634–644, 2024.

- [32] S. Huffman, S. Chen, K. A. Mack, H. Su, Q. Wang, and R. Kushalnagar. "we do use it, but not how hearing people think": How the deaf and hard of hearing community uses large language model tools. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–9, 2025.
- [33] M. Inan, K. Atwell, A. Sicilia, L. Quandt, and M. Alikhani. Generating signed language instructions in large-scale dialogue systems. *arXiv preprint arXiv:2410.14026*, 2024.
- [34] Z. Jiang, A. Göhring, A. Moryossef, R. Sennrich, and S. Ebling. Swissli: 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)*, pages 15448–15456, 2024.
- [35] R. Johnson. Towards enhanced visual clarity of sign language avatars through recreation of fine facial detail. *Machine Translation*, 35(3):431–445, 2021.
- [36] A. Joshi, S. Agrawal, and A. Modi. Isltranslate: Dataset for translating indian sign language. *arXiv preprint arXiv:2307.05440*, 2023.
- [37] A. Joshi, R. Mohanty, M. Kanakanti, A. Mangla, S. Choudhary, M. Barbate, and A. Modi. isign: A benchmark for indian sign language processing. *arXiv preprint arXiv:2407.05404*, 2024.
- [38] S. Kausar and M. Y. Javed. A survey on sign language recognition. In *2011 frontiers of information technology*, pages 95–98. IEEE, 2011.
- [39] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.
- [40] W. Kim, J. Sung, D. Saakes, C. Huang, and S. Xiong. Ergonomic postural assessment using a new open-source human pose estimation technology (openpose). *International Journal of Industrial Ergonomics*, 84:103164, 2021.
- [41] S.-K. Ko, C. J. Kim, H. Jung, and C. Cho. Neural sign language translation based on human keypoint estimation. *Applied Sciences*, 9(13):2683, 2019.
- [42] T. Komura, I. Habibie, D. Holden, J. Schwarz, and J. Yearsley. A recurrent variational autoencoder for human motion synthesis. In *The 28th British Machine Vision Conference*, 2017.
- [43] B. Li, Y. Zhao, S. Zhelun, and L. Sheng. Danceformer: Music conditioned 3d dance generation with parametric motion transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1272–1279, 2022.
- [44] R. Li, S. Yang, D. A. Ross, and A. Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13401–13412, 2021.
- [45] Z. Li, J. Zhang, Q. Lin, J. Xiong, Y. Long, X. Deng, Y. Zhang, X. Liu, M. Huang, Z. Xiao, et al. Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding. *arXiv preprint arXiv:2405.08748*, 2024.
- [46] J. Lin, A. Zeng, S. Lu, Y. Cai, R. Zhang, H. Wang, and L. Zhang. Motion-x: A large-scale 3d expressive whole-body human motion dataset. *Advances in Neural Information Processing Systems*, 36:25268–25280, 2023.
- [47] H. Liu, Z. Zhu, N. Iwamoto, Y. Peng, Z. Li, Y. Zhou, E. Bozkurt, and B. Zheng. Beat: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis. In *European conference on computer vision*, pages 612–630. Springer, 2022.
- [48] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866. 2023.
- [49] K. Lyu, H. Chen, Z. Liu, B. Zhang, and R. Wang. 3d human motion prediction: A survey. *Neurocomputing*, 489:345–365, 2022.
- [50] A. M. Martínez, R. B. Wilbur, R. Shay, and A. C. Kak. Purdue rvl-slll asl database for automatic recognition of american sign language. In *Proceedings. Fourth IEEE International Conference on Multimodal Interfaces*, pages 167–172. IEEE, 2002.

- [51] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.
- [52] R. Mineo, A. Sorrenti, G. Caligiore, F. P. Salanitri, G. Bellitto, S. Polikovsky, S. Fontana, E. Ragonese, C. Spampinato, and S. Palazzo. Text-aligned radar-based sign language recognition for healthcare communication. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4894–4902, 2025.
- [53] H. Mokayed, G. Alsayed, F. Lodin, O. Hagner, and B. Backe. 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)*, pages 198–203. IEEE, 2024.
- [54] L. Momeni, G. Varol, S. Albanie, T. Afouras, and A. Zisserman. Watch, read and lookup: learning to spot signs from multiple supervisors. In *ACCV*, 2020.
- [55] G. Moon, T. Shiratori, and S. Saito. Expressive whole-body 3d gaussian avatar. In *ECCV*, 2024.
- [56] A. Moryossef, M. Müller, A. Göhring, Z. Jiang, Y. Goldberg, and S. Ebling. 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*, pages 22–33, 2023.
- [57] E. Ng, S. Ginosar, T. Darrell, and H. Joo. Body2hands: Learning to infer 3d hands from conversational gesture body dynamics. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11865–11874, 2021.
- [58] B. O. Olusanya, A. C. Davis, and H. J. Hoffman. Hearing loss: rising prevalence and impact. *Bulletin of the World Health Organization*, 97(10):646–646A, 2019.
- [59] K. Papadimitriou, G. Potamianos, G. Sapountzaki, T. Goulas, E. Efthimiou, S.-E. Fotinea, and P. Maragos. Greek sign language recognition for an education platform. *Universal Access in the Information Society*, 24(1):51–68, 2025.
- [60] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. Osman, D. Tzionas, and M. J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985, 2019.
- [61] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. A. Osman, D. Tzionas, and M. J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019.
- [62] W. Peebles and S. Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023.
- [63] M. Petrovich, M. J. Black, and G. Varol. Temos: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision*, pages 480–497. Springer, 2022.
- [64] X. Qi, Y. Wang, H. Zhang, J. Pan, W. Xue, S. Zhang, W. Luo, Q. Liu, and Y. Guo. Co³ gesture: Towards coherent concurrent co-speech 3d gesture generation with interactive diffusion. *arXiv preprint arXiv:2505.01746*, 2025.
- [65] X. Qi, H. Zhang, Y. Wang, J. Pan, C. Liu, M. Sun, W. Xue, S. Zhang, S. Han, Q. Liu, et al. Cocogesture: Towards coherent co-speech 3d gesture generation in the wild. *Information Fusion*, page 103613, 2025.
- [66] Z. Qian, S. Wang, M. Mihajlovic, A. Geiger, and S. Tang. 3dgs-avatar: Animatable avatars via deformable 3d gaussian splatting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5020–5030, 2024.
- [67] X. Qin, H. Dai, X. Hu, D.-P. Fan, L. Shao, and L. V. Gool. Highly accurate dichotomous image segmentation. In *ECCV*, 2022.
- [68] L. C. Quandt, A. Willis, M. Schwenk, K. Weeks, and R. Ferster. Attitudes toward signing avatars vary depending on hearing status, age of signed language acquisition, and avatar type. *Frontiers in psychology*, 13:730917, 2022.

- [69] E. R and N. B. ISL-CSLTR: Indian Sign Language Dataset for Continuous Sign Language Translation and Recognition, 2021.
- [70] C. Raude, K. Prajwal, L. Momeni, H. Bull, S. Albanie, A. Zisserman, and G. Varol. A tale of two languages: Large-vocabulary continuous sign language recognition from spoken language supervision. *arXiv preprint arXiv:2405.10266*, 2024.
- [71] G. Retsinas, P. P. Filntisis, R. Danecek, V. F. Abrevaya, A. Roussos, T. Bolkart, and P. Maragos. Smirk: 3d facial expressions through analysis-by-neural-synthesis. *arXiv preprint arXiv:2404.04104*, 2024.
- [72] P. A. Rodríguez-Correa, A. Valencia-Arias, O. N. Patiño-Toro, Y. Oblitas Díaz, and R. Teodori De la Puente. Benefits and development of assistive technologies for deaf people’s communication: A systematic review. In *Frontiers in Education*, volume 8, page 1121597. Frontiers Media SA, 2023.
- [73] D. M. Rouse and S. S. Hemami. Analyzing the role of visual structure in the recognition of natural image content with multi-scale ssim. In *Human vision and electronic imaging XIII*, volume 6806, pages 410–423. SPIE, 2008.
- [74] P. Ruiz-Ponce, G. Barquero, C. Palmero, S. Escalera, and J. García-Rodríguez. Mixermdm: Learnable composition of human motion diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12380–12390, 2025.
- [75] W. Sandler and D. C. Lillo-Martin. *Sign language and linguistic universals*. Cambridge University Press, 2006.
- [76] S. Sanyal, T. Bolkart, H. Feng, and M. J. Black. Learning to regress 3d face shape and expression from an image without 3d supervision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7763–7772, 2019.
- [77] B. Saunders, N. C. Camgoz, and R. Bowden. Everybody sign now: Translating spoken language to photo realistic sign language video. *arXiv preprint arXiv:2011.09846*, 2020.
- [78] B. Saunders, N. C. Camgoz, and R. Bowden. Progressive transformers for end-to-end sign language production. In *European Conference on Computer Vision*, pages 687–705. Springer, 2020.
- [79] B. Saunders, N. C. Camgoz, and R. Bowden. Continuous 3d multi-channel sign language production via progressive transformers and mixture density networks. *International journal of computer vision*, 129(7):2113–2135, 2021.
- [80] B. Saunders, N. C. Camgoz, and R. Bowden. 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*, pages 5141–5151, 2022.
- [81] A. Sengupta, F. Jin, R. Zhang, and S. Cao. mm-pose: Real-time human skeletal posture estimation using mmwave radars and cnns. *IEEE sensors journal*, 20(17):10032–10044, 2020.
- [82] W. Seo, J. Oh, and M. Kim. Bim-vfi: Bidirectional motion field-guided frame interpolation for video with non-uniform motions. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7244–7253, 2025.
- [83] A. Serifi, R. Grandia, E. Knoop, M. Gross, and M. Bächer. Robot motion diffusion model: Motion generation for robotic characters. In *SIGGRAPH asia 2024 conference papers*, pages 1–9, 2024.
- [84] A. Sharma. India’s language policy for deaf and hard-of-hearing people. *Language Policy*, pages 1–23, 2025.
- [85] L. Shen, T. Liu, H. Sun, X. Ye, B. Li, J. Zhang, and Z. Cao. Dreammover: Leveraging the prior of diffusion models for image interpolation with large motion. In *European Conference on Computer Vision*, pages 336–353. Springer, 2024.
- [86] B. Shi, D. Brentari, G. Shakhnarovich, and K. Livescu. Open-domain sign language translation learned from online video. *arXiv preprint arXiv:2205.12870*, 2022.
- [87] K. Shoemake. Animating rotation with quaternion curves. In *Proceedings of the 12th annual conference on Computer graphics and interactive techniques*, pages 245–254, 1985.

- [88] A. K. Singh, V. A. Kumbhare, and K. Arthi. Real-time human pose detection and recognition using mediapipe. In *International conference on soft computing and signal processing*, pages 145–154. Springer, 2021.
- [89] J. Singha, A. Roy, and R. H. Laskar. Dynamic hand gesture recognition using vision-based approach for human-computer interaction. *Neural Computing and Applications*, 29(4):1129–1141, 2018.
- [90] S. Stoll, N. C. Camgoz, S. Hadfield, and R. Bowden. Text2sign: Towards sign language production using neural machine translation and generative adversarial networks. *International Journal of Computer Vision*, 128(4):891–908, 2020.
- [91] S. Tang, J. He, D. Guo, Y. Wei, F. Li, and R. Hong. Sign-idd: Iconicity disentangled diffusion for sign language production. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 7266–7274, 2025.
- [92] G. Tevet, S. Raab, B. Gordon, Y. Shafir, D. Cohen-Or, and A. H. Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022.
- [93] A. P. Uchil, S. Jha, and B. Sudha. Vision based deep learning approach for dynamic indian sign language recognition in healthcare. In *International conference on computational vision and bio inspired computing*, pages 371–383. Springer, 2019.
- [94] D. Uthus, G. Tanzer, and M. Georg. Youtube-asl: A large-scale, open-domain american sign language-english parallel corpus. *Advances in Neural Information Processing Systems*, 36:29029–29047, 2023.
- [95] H. Walsh, E. Fish, O. M. Sincan, M. I. Lakhall, R. Bowden, N. Fox, B. Woll, K. Wu, Z. Li, W. Zhao, et al. Slrtp2025 sign language production challenge: Methodology, results and future work. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4109–4119, 2025.
- [96] H. Walsh, B. Saunders, and R. Bowden. Sign stitching: a novel approach to sign language production. *arXiv preprint arXiv:2405.07663*, 2024.
- [97] H. Walsh, B. Saunders, and R. Bowden. Sign stitching: A novel approach to sign language production, 2024.
- [98] B. Wang, X. Wang, C. Ni, G. Zhao, Z. Yang, Z. Zhu, M. Zhang, Y. Zhou, X. Chen, G. Huang, et al. Humandreamer: Generating controllable human-motion videos via decoupled generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12391–12401, 2025.
- [99] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364, 2020.
- [100] X. Wang, S. Zhang, L. Tang, Y. Zhang, C. Gao, Y. Wang, and N. Sang. Unianimate-dit: Human image animation with large-scale video diffusion transformer. *arXiv preprint arXiv:2504.11289*, 2025.
- [101] P. Wolfert, N. Robinson, and T. Belpaeme. A review of evaluation practices of gesture generation in embodied conversational agents. *IEEE Transactions on Human-Machine Systems*, 52(3):379–389, 2022.
- [102] W. Xue, L. Zhang, X. Mou, and A. C. Bovik. Gradient magnitude similarity deviation: A highly efficient perceptual image quality index. *IEEE transactions on image processing*, 23(2):684–695, 2013.
- [103] P. Yang, S. Zhou, J. Zhao, Q. Tao, and C. C. Loy. MatAnyone: Stable video matting with consistent memory propagation. In *CVPR*, 2025.
- [104] Z. Yang, A. Zeng, C. Yuan, and Y. Li. Effective whole-body pose estimation with two-stages distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4210–4220, 2023.
- [105] Y. Yu, W. Zhang, and Y. Deng. Frechet inception distance (fid) for evaluating gans. *China University of Mining Technology Beijing Graduate School*, 3(11), 2021.
- [106] Z. Yu, S. Huang, Y. Cheng, and T. Birdal. Signavatars: A large-scale 3d sign language holistic motion dataset and benchmark. In *European Conference on Computer Vision*, pages 1–19. Springer, 2024.
- [107] Z. Zafrulla, H. Brashear, P. Yin, P. Presti, T. Starner, and H. Hamilton. American sign language phrase verification in an educational game for deaf children. In *2010 20th International Conference on Pattern Recognition*, pages 3846–3849. IEEE, 2010.

- [108] C. Zhang, Y. Tang, N. Zhang, R.-S. Lin, M. Han, J. Xiao, and S. Wang. Bidirectional autoregressive diffusion model for dance generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 687–696, 2024.
- [109] G. Zhang, C. Liu, Y. Cui, X. Zhao, K. Ma, and L. Wang. Vfimamba: Video frame interpolation with state space models. *Advances in Neural Information Processing Systems*, 37:107225–107248, 2024.
- [110] H. Zhang, R. Shalev-Arkushin, V. Baltatzis, C. Gillis, G. Laput, R. Kushalnagar, L. C. Quandt, L. Findlater, A. Bedri, and C. Lea. Towards ai-driven sign language generation with non-manual markers. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–26, 2025.
- [111] J. Zhang, Y. Zhang, X. Cun, S. Huang, Y. Zhang, H. Zhao, H. Lu, and X. Shen. T2m-gpt: Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.
- [112] L. Zhang, L. Zhang, and A. C. Bovik. A feature-enriched completely blind image quality evaluator. *IEEE Transactions on Image Processing*, 24(8):2579–2591, 2015.
- [113] H. Zhou, W. Zhou, W. Qi, J. Pu, and H. Li. Improving sign language translation with monolingual data by sign back-translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1316–1325, 2021.
- [114] Q.-Y. Zhou, J. Park, and V. Koltun. Open3d: A modern library for 3d data processing. *arXiv preprint arXiv:1801.09847*, 2018.
- [115] L. Zhu, X. Liu, X. Liu, R. Qian, Z. Liu, and L. Yu. Taming diffusion models for audio-driven co-speech gesture generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10544–10553, 2023.
- [116] W. Zhu, X. Ma, Z. Liu, L. Liu, W. Wu, and Y. Wang. Motionbert: A unified perspective on learning human motion representations. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15085–15099, 2023.
- [117] W. Zhu, X. Ma, D. Ro, H. Ci, J. Zhang, J. Shi, F. Gao, Q. Tian, and Y. Wang. Human motion generation: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(4):2430–2449, 2023.
- [118] W. Zielonka, T. Bolkart, and J. Thies. Towards metrical reconstruction of human faces. In *European conference on computer vision*, pages 250–269. Springer, 2022.
- [119] R. Zuo, F. Wei, Z. Chen, B. Mak, J. Yang, and X. Tong. A simple baseline for spoken language to sign language translation with 3d avatars. In *European Conference on Computer Vision*, pages 36–54. Springer, 2024.