# SignSparK: Efficient Multilingual Sign Language Production via Sparse Keyframe Learning

Jianhe Low, Alexandre Symeonidis-Herzig, Maksym Ivashechkin, Ozge Mercanoglu Sincan, and Richard Bowden

CVSSP, University of Surrey, United Kingdom
{jianhe.low, a.symeonidisherzig, m.ivashechkin, o.mercanoglusincan, r.bowden}@surrey.ac.uk
https://cogvis-cvssp.github.io/papers/signspark/

**Abstract.** Sign Language Production (SLP) faces a fundamental trade-off: direct text-to-pose models suffer from regression-to-the-mean effects, while dictionary-retrieval methods produce disjointed transitions. To resolve this, we propose a novel training paradigm that leverages sparse keyframes to capture the underlying kinematic distribution of human signing. By generating dense motion from discrete anchors, our approach mitigates regression-to-the-mean while ensuring fluid articulation. To achieve this at scale, we introduce FAST, an ultra-efficient sign segmentation model that automatically mines precise temporal boundaries. We then present SignSparK, a Conditional Flow Matching (CFM) framework that utilizes these temporal anchors to synthesize 3D signing sequences. This keyframe-driven formulation also unlocks Keyframe-to-Pose (KF2P) generation, making precise spatiotemporal editing of signing sequences possible. Furthermore, SignSparK scales across four distinct sign languages, constituting the largest multilingual SLP framework to date, and integrates 3D Gaussian Splatting for photorealistic rendering. Extensive evaluations demonstrate that SignSparK achieves state-of-the-art across diverse SLP tasks and multilingual benchmarks. Our code is available at https://github.com/JianHe0628/SignSparK.

**Keywords:** Sign Language Production · Sign Segmentation · Sparse Keyframe Learning · 3D Human Motion Generation · Flow Matching

## 1 Introduction

Sign languages are linguistically rich, natural languages primarily used in Deaf communities, and are governed by precise hand shapes and fluid body dynamics [47, 58, 78]. While Sign Language Translation (SLT) from video-to-text has been extensively studied [10,26,46,51,77,101], its inverse, Sign Language Production (SLP) [3,71,80,97,104] has emerged only more recently, and carries its own unique set of structural challenges. Chief among these is the need to satisfy two demands at once: (i) each sign must be articulated with linguistic precision, and (ii) successive signs must connect with the smooth continuity of natural movement. Reconciling both characteristics within a single text-to-motion mapping has proven non-trivial, and existing methods thus sacrifice one for the other.

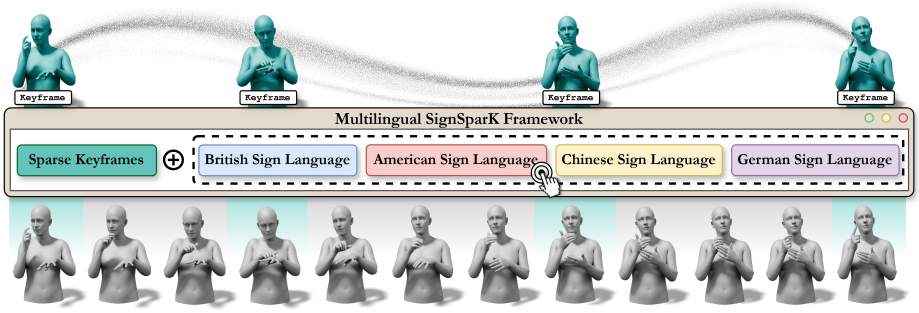


**Fig. 1: SignSparK** is a Conditional Flow Matching model trained on sparse keyframes, and generates realistic and natural 3D signing avatars given spoken text. Designed for efficiency, SignSparK scales to four distinct sign languages under a unified framework.

For instance, *direct* Text-to-Pose (T2P) models [71, 80, 83, 97] attempt to regress this mapping directly, but often collapse under the modality gap, yielding under-articulated and unintelligible signing. Meanwhile, real-world T2P deployment instead favours sign dictionary-retrieval pipelines based on *glosses*[1] [74, 79, 105], as text is first translated to a gloss sequence, before being retrieved from a pre-recorded gloss dictionary. This pipeline preserves articulation, but sacrifices fluency: the clips must still be stitched together, and current methods do so via *naive interpolation*, producing signing motion that looks robotic. Additionally, current SLP paradigms are also bottlenecked by inaccuracies in monocular 3D estimation (e.g., HaMeR [57]), as the ambiguity of lifting 2D views into 3D meshes yields pseudo-ground truth that propagates directly into trained models.

To address these challenges, we propose SignSparK (**Sign** Language Production with **Spar**se **K**eyframes), a large-scale SLP framework driven by a novel *training paradigm*: learning to synthesize continuous signing sequences given spoken text and sparse keyframes as training inputs (Fig. 1). Since the model must explicitly hit these sparse keyframes, they act as anchors that inherently prevent regression-to-the-mean. When applied at scale, SignSparK cannot rely on naive interpolation and must thus learn the underlying distribution of fluid signing instead. This makes SignSparK a drop-in replacement for the stitching stage of retrieval-based pipelines, substituting the hand-crafted interpolation methods with a learnt motion prior. Additionally, SignSparK's training paradigm also uniquely unlocks the highly controllable task of Keyframe-to-Pose (KF2P) synthesis. Here, users can shift anchors to dictate signing speed, insert intermediates for signs with complex internal motion, or replace inaccurate keyframes with high-fidelity mocap data, bounding output quality directly to the conditioning signal rather than to imperfect pseudo-ground truth. Finally, SignSparK's Flow Matching architecture also achieves high-quality synthesis in fewer than ten sampling steps, allowing us to scale to the largest multilingual SLP framework to date, encompassing German, Chinese, American, and British sign languages.

[1] Glosses are the written representations used to denote individual signs.

However, keyframe annotations do not exist in current SLP datasets, rendering this training paradigm infeasible. To overcome this, we additionally introduce FAST, an extremely efficient sign language segmentation model designed to automatically mine linguistically meaningful temporal anchors at scale. By precisely identifying sign boundaries, FAST provides the foundational keyframe extractions required to train SignSparK. Beyond enabling SignSparK, FAST is also explicitly engineered to process large-scale corpora like BOBSL [1] with minimal computational overhead. It therefore offers immense standalone value for the broader community, as it can facilitate the rapid pseudo-annotation of emerging large-scale glossless datasets like YTSL-25 [85] and CSL-News [46], and support downstream tasks such as sign spotting [88] and tokenization [51].

By leveraging the keyframes extracted by FAST in training, SignSparK synthesizes articulate and naturalistic signing sequences directly within the 3D parametric spaces of SMPL-X [56], MANO [67], and FLAME [45]. Unlike prior 2D keypoint methods, this 3D formulation offers superior spatial depth and physical plausibility. We further show that 3D Gaussian Splatting (3DGS) [40] integrates naturally into our pipeline, overcoming the visual limitations of bare meshes by rendering the generated kinematics as photorealistic signing avatars. In short, the main contributions of our work can be summarized as follows:

- **State-of-the-art segmentation.** We introduce FAST, an ultra-efficient sign segmentation model that achieves highly accurate boundary detection at scale and unlocks the linguistic keyframes required for our training paradigm.
- **Novel training paradigm.** We propose SignSparK, a generative SLP framework driven by sparse keyframe training. This approach overcomes under-articulation and ensures fluid motion while uniquely unlocking Keyframe-to-Pose (KF2P) synthesis for precise spatial and temporal control.
- **Unprecedented efficiency and scale.** We formulate a reconstruction-based Flow Matching objective for SLP and achieve a 100× efficiency gain over prior models. We simultaneously leverage this efficiency to establish the largest multilingual SLP framework to date across four major sign languages.
- **State-of-the-art SLP performance.** Extensive evaluations demonstrate that SignSparK achieves state-of-the-art performance across a wide range of SLP regimes on multilingual sign language datasets.

## 2 Related Work

***Sign Language Segmentation.*** In continuous signing, the fluid transitional movements between consecutive signs, known as *coarticulations*, create smooth motion continuity. Sign language segmentation seeks to temporally localize individual signs within these sequences, while disregarding coarticulatory transitions. Early approaches relied on statistical techniques [6, 17, 68] and machine learning approaches [25], but struggled with the multimodal, continuous nature of signing. Deep learning approaches later adopted sequential architectures such as the BiLSTM [8] and spatiotemporal models like the I3D [13] to achieve notable improvements [7, 65, 66]. More recent studies, inspired by linguistics, then

reframed sign segmentation as a Begin-In-Out (BIO) tagging task [53]. With Hands-On [52], advancing this task through strong 3D body [34] and hand [57] priors, achieving state-of-the-art on the DGS Corpus [43]. Our method, FAST, further improves this via a unimodal two-stream design and interpolated training, providing superior accuracy and significant efficiency gains.

***Sign Language Production.*** Generating naturalistic signing videos or avatars from language or motion-based inputs defines the field of SLP. Early research was filled with graphics-based avatars driven by handcrafted linguistic rules and annotations [4, 19, 22, 23, 106]. Although linguistically structured, these pipelines produced unnatural motions, limiting acceptance within the Deaf community [42].

The advent of deep learning shifted SLP towards data-driven approaches. Owing to the non-monotonic mapping between sign and spoken languages, many systems introduced glosses as intermediate supervision, forming text-to-gloss-to-pose (T2G2P) frameworks [70–73, 80, 83, 97]. In contrast, direct regression methods typically predict 2D/3D poses straight from text or linguistic inputs such as HamNoSys [2] or glosses [14, 33, 84]; however, they frequently produce under-articulated motion due to regression-to-the-mean effects. Alternatives like interpolating isolated signs [74, 89] improve said articulations, but rely on complex pipelines and smoothing filters that fail to capture natural human signing.

To enhance realism, methods employing GANs [74, 79, 80], diffusion models [24, 90], and recently 3DGS [36] have also been explored. However, a growing trend now centres on 3D avatar generation, where models predict parametric human body models such as SMPL-X [56] to enable full-body mesh reconstructions [3, 5, 20, 82, 96, 104]. These frameworks span diffusion [3, 5], VAE [20], and VQ-VAE [96, 104] paradigms, but they remain largely monolingual and computationally inefficient. To overcome this, Conditional Flow Matching (CFM) has recently emerged as a highly efficient generative alternative, with SignFlow [41] pioneering its use in SLP. However, its resulting synthesis still exhibits clear articulatory inaccuracies and its evaluation remains restricted to a single dataset.

Ultimately, progress in 3D production, multilingual scaling, and inference efficiency has remained highly fragmented. SignSparK instead presents a unified framework to bridge these domains. By integrating a keyframe-based paradigm for fluid articulation with a highly efficient reconstruction-based CFM formulation, as well as realistic rendering via 3DGS [40], we scale to a massive multilingual setting to provide a practical, high-fidelity system for the Deaf community.

***Human Body Motion Generation.*** Human motion generation is a long-standing problem in computer vision, conditioned on diverse modalities such as action labels [9, 59, 92, 95], audio [44, 76], and increasingly, text [15, 27–29, 60, 86, 87, 100], enabled by large-scale datasets like KIT-ML [62], BABEL [64], and HumanML3D [28]. Recent approaches typically represent motion as latent tokens within autoencoder frameworks for autoregressive Transformer-based generation [27, 60, 86, 99], or leverage diffusion-based models [31] to produce temporally coherent sequences [15, 39, 75, 87, 100]. However, despite significant progress, human motion generation approaches emphasize full-body motion, and thus neglect the fine-grained hand and finger dynamics essential for sign language production.

***Keyframe In-betweening Generation.*** In motion/video synthesis, *keyframe in-betweening* aims to generate intermediate frames given sparse keyframe constraints. Diffusion-based approaches have recently advanced this across domains, including video [30, 37, 91, 93], human motion [16, 39, 87], and hand interactions [48]. However, existing human motion frameworks remain limited: MDM [87] suffers from foot-sliding and unrealistic transitions under keyframe conditioning; GMD [39] handles sparse interpolation but focuses on pelvis trajectories; Cond-MDI [16] is closest to our setting, but targets general locomotion without explicit hand modelling and relies on a computationally expensive diffusion process. In contrast, our approach enables efficient few-step sampling while explicitly capturing the fine-grained articulations essential for natural signing motion.

## 3 Methodology

This work proposes a sparse keyframe-based SLP training paradigm consisting of two primary components. Firstly, a sign language segmentation model (Sec. 3.1) that localizes sign boundaries and extracts linguistically meaningful keyframes; and secondly, a CFM model (Sec. 3.2) that synthesizes smooth and realistic sign language motion conditioned on these sparse keyframes and spoken text input.

### 3.1 Sign Language Segmentation

Sign language segmentation can be formulated as a *BIO-tagging* problem, where each frame is classified as either `B` (beginning of a sign), `I` (inside a sign), or `O` (outside a sign), corresponding to the class indices $\{2, 1, 0\}$. In this setting, given an input video $\mathbf{X} \in \mathbb{R}^{T\times H\times W\times C}$, with $T$ frames, $H \times W$ spatial resolution, and $C$ channel dimensions, the task is to assign each frame with a one-hot label $\mathbf{Y} \in \mathbb{R}^{T\times 3}$, treating segmentation as a per-frame classification problem.

***Feature Representation.*** To capture signing dynamics, previous work relied on HaMeR-extracted MANO hand parameters [57,67] and 3D pose skeletons [34], achieving state-of-the-art segmentation [52] but at high computational costs impractical for large-scale multilingual sign datasets. We instead adopt WiLoR [63], a MANO regression model that recently attained superior performance, while featuring a 45× faster and 32× more compact hand detector. In addition, we also introduce our segmentation model as a unimodal approach to reduce complexity and feature extraction time. This design choice was to specifically improve model efficiency, but interestingly yields no drop in performance (see supplementary).

***Architecture Design.*** Our **FAST** (**F**ast and **A**ccurate **S**ign segmen**T**ation) framework, is a transformer-based per-frame classification model (Fig. 2a) designed to capture hand shape transitions vital to segmentation. Specifically, given a video sequence $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$, we first extract per-frame MANO parameters for the left and right hands, $\mathbf{H}_L, \mathbf{H}_R \in \mathbb{R}^{T\times(J\times d_r)}$, where $J$ is the number of joints and $d_r$ the rotation dimensions. Here, we adopt 6D rotations, as it has demonstrated better suitability for deep learning [103]. To capture the independent semantic contributions of each hand, FAST encodes left and right

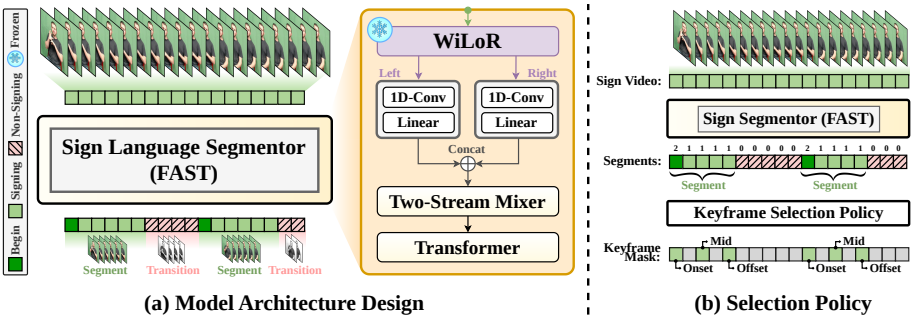


**Fig. 2: Overview of FAST. (a) Architecture:** WiLoR first extracts the left and right hand representations from input frames. These are then encoded via parallel spatio-temporal streams, concatenated, and refined by a two-stream mixer before a Transformer generates dense per-frame BIO segmentation labels. **(b) Selection Policy:** Leveraging the predicted BIO segments, we explicitly isolate the onset, midpoint, and offset frames of each sign to construct a semantically rich keyframe mask.

MANO features separately via parallel streams before fusing them to predict frame-wise BIO labels. We train the network using cross-entropy (frame-level) and CTC (sign-level) alongside temporal augmentations to yield a final segmentation model that is highly scalable, accurate, and computationally efficient. Further implementation details are provided in the supplementary material.

***Keyframe Selection Policy.*** Our selection policy aims to extract a sparse and semantically rich subset of frames that distils the motion and content of each sign. Leveraging the frame-wise predictions $y_t \in \{0, 1, 2\}$ from FAST, we ground our sampling in linguistic structure by explicitly isolating the exact temporal boundaries of each gesture: the onset ($y_t = 2$) and the offset (defined as $y_t = 1$ given $y_{t+1} = 0$). To capture the core content of the gesture, we further incorporate the segment midpoint. This resulting policy of onset→mid→offset (Fig. 2b) outperforms both random and heuristically augmented baselines (Sec. 4.3), and also provides the optimal sparsity-performance trade-off (Supplementary).

### 3.2 Sparse Keyframe-conditioned Training via Flow Matching

Unlike prior SLP training approaches that condition sign generation on spoken text [3], tokenized motion sequences [104], or textual motion descriptions [5], we adopt the conditioning signal of sparse keyframes and spoken text as it enables the model to learn the underlying articulatory patterns of human signing, while also respecting its linguistic reference (Fig. 3). Formally, given a frame control signal $\mathbf{C} \in \mathbb{R}^{T \times D}$ and text condition $\mathcal{T}$, the objective is to synthesize a signing sequence $\mathbf{S} \in \mathbb{R}^{T \times D}$. The control signal $\mathbf{C}$ comprises of only $k$ keyframes ($k \ll T$), with the remaining $(T - k)$ frames corrupted with Gaussian noise $\mathcal{N}(\mathbf{0}, \mathbf{I})$.

To represent signing sequence $\mathbf{S} \in \mathbb{R}^{T \times D}$, we regress SMPL-X upper-body parameters $\mathbf{B} \in \mathbb{R}^{T \times (10 \times d_r)}$ using NLF [69], MANO hand parameters $\mathbf{H}_L, \mathbf{H}_R \in$

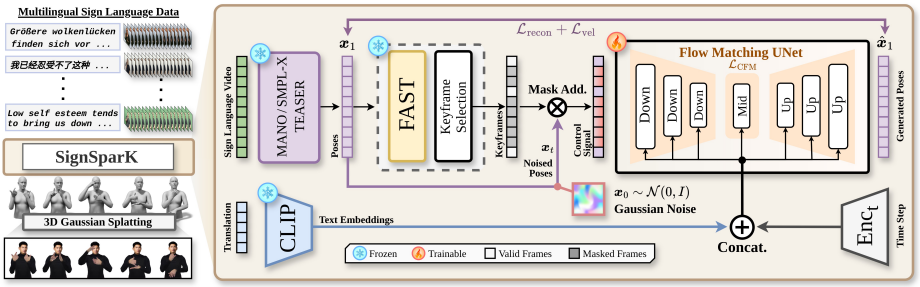


**Fig. 3: Architecture of SignSparK**. (i) A sign language video is first processed by WiLoR, NLF, and TEASER to extract 3D parametric representations, while its text translation is embedded via Multilingual-CLIP [12]. (ii) FAST subsequently localizes sign segments, and the selection policy pinpoints the keyframes needed to form the control signal. (iii) A UNet, conditioned on timestep, control signal, and text, then reconstructs clean poses. (iv) These poses can then be rendered from meshes into realistic signing avatars via 3DGS. Further 3DGS implementation details in the supplementary

$\mathbb{R}^{T\times(15\times d_r)}$ via WiLoR [63], and FLAME facial parameters $\mathbf{F} \in \mathbb{R}^{T\times 56}$ (50 expression coefficients and a 6D jaw rotation) with TEASER [50]. We follow [104] by modeling the signing avatar as 10 upper-body joints and 15 joints per hand, but adopt 6D rotation parameterization [103]. Consistent with prior work, we also maintain $\mathbf{B}$, $(\mathbf{H}_L, \mathbf{H}_R)$, and additionally $\mathbf{F}$ as separate channels to preserve their distinct motion dynamics. Keyframes are then obtained using the FAST model and selection policy (Sec. 3.1), resulting in the creation of binary mask $\mathcal{M} = \{m_1, m_2, \ldots, m_T\}$, where $m_t = 1$ denotes a valid keyframe.

***Control Signal Construction.*** Given the keyframe mask $\mathcal{M} = \{m_t\}_{t=1}^{T}$, we then construct the control signal to condition SignSparK. First, let $\boldsymbol{x}_1 \in \mathbb{R}^{T\times d}$ denote the ground-truth pose sequence for the channels under consideration, where $d = 60$ for body and $d = 90$ for hands. Then, following the CFM formulation (preliminaries in supplementary), we define a continuous probability path between the original sequence $\boldsymbol{x}_1$ and a noise sample $\boldsymbol{x}_0$ via linear interpolation:

$$\boldsymbol{x}_t = t\,\boldsymbol{x}_1 + (1-t)\,\boldsymbol{x}_0, \quad \boldsymbol{x}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad t \in [0,1]. \tag{1}$$

The keyframe mask $\mathcal{M}$ is subsequently projected across the feature dimension and used to preserve the ground-truth poses at keyframe locations while injecting the interpolated noised poses elsewhere, yielding the control signal:

$$\mathbf{C} = \mathcal{M} \odot \boldsymbol{x}_1 \;+\; (1-\mathcal{M}) \odot \boldsymbol{x}_t, \tag{2}$$

where $\odot$ denotes element-wise multiplication. This sparsity forces the model to infer intermediate motion, promoting the learning of natural signing dynamics.

***Flow Matching Model.*** To model the rich variability of human signing, we employ a conditional Flow Matching network $f_\theta$ that predicts a time-dependent vector field $\boldsymbol{v}_t \in \mathbb{R}^{T\times d}$, guided by control signal $\mathbf{C}_t \in \mathbb{R}^{T\times d}$ and timestep $t$.

Additionally, to enable T2P capabilities, the model is also conditioned on spoken translations $\mathcal{T}$, embedded via a text encoder $\boldsymbol{z}_{\mathcal{T}} = \text{Enc}(\mathcal{T})$, yielding the final conditional vector field parameterization $\boldsymbol{v}_t = f_\theta(t, \mathbf{C}_t, \boldsymbol{z}_{\mathcal{T}})$. Following CFM [49], the target vector field is defined as $\boldsymbol{u}_t(\mathbf{C}_t \mid \boldsymbol{x}_1) = \boldsymbol{x}_1 - \mathbf{C}_t$, capturing the residual direction needed to move the noised control signal $\mathbf{C}_t$ towards the ground-truth $\boldsymbol{x}_1$. The model is therefore trained to approximate this field by minimizing:

$$\mathcal{L}_{\text{CFM}}(\theta) = \mathbb{E}_{t,\,\mathbf{C}_t,\,\boldsymbol{x}_1} \left\| f_\theta(t, \mathbf{C}_t, \boldsymbol{z}_{\mathcal{T}}) - (\boldsymbol{x}_1 - \mathbf{C}_t) \right\|^2. \tag{3}$$

To enable flexible inference with or without guidance inputs $\mathbf{C}_t$ and $\boldsymbol{z}_{\mathcal{T}}$, we adopt classifier-free guidance (CFG) [32]; where during training, guidance inputs are randomly dropped with probability $\rho = 0.1$, and replaced by a null token $\varnothing$, teaching the model to handle both keyframe conditioned and unconditioned scenarios. At inference, the guidance's effect can then be modulated by interpolating the two predicted vector fields via $\boldsymbol{v}_t^{(\gamma)} = \boldsymbol{v}_t^{\text{uncond}} + \gamma\,(\boldsymbol{v}_t^{\text{cond}} - \boldsymbol{v}_t^{\text{uncond}})$, where $\boldsymbol{v}_t^{\text{cond}} = f_\theta(t, \mathbf{C}_t, \boldsymbol{z}_{\mathcal{T}})$, $\boldsymbol{v}_t^{\text{uncond}} = f_\theta(t, \varnothing, \varnothing)$, and $\gamma \geq 0$ controls the guidance scale. This formulation grants explicit control over the model's guidance adherence, and also directly enables end-to-end T2P translation at inference.

***Reconstruction Regularization.*** While the CFM objective encourages the model to capture motion via vector fields, we observed that explicitly guiding the prediction towards the target pose improves convergence and sampling fidelity. To this end, we introduce a *flow-based reconstruction loss*, which forces the vector field $\boldsymbol{v}_t = f_\theta(t, \mathbf{C}_t, \boldsymbol{z}_{\mathcal{T}})$ to recover the original pose $\boldsymbol{x}_1$ from the noised control signal $\mathbf{C}_t$. Specifically, we estimate the original pose via a single-step Euler integration along the vector field from the current timestep $t$ towards $t = 1$:

$$\boldsymbol{x}_1^{\text{est}} = \boldsymbol{x}_t + (1 - t)\,\boldsymbol{v}_t, \quad t \in [0, 1]. \tag{4}$$

We then supervise the network using mean-squared error to penalize significant deviations from the ground-truth poses:

$$\mathcal{L}_{\text{recon}} = \mathbb{E}_{\boldsymbol{x}_1, \boldsymbol{x}_t, t} \left\| \boldsymbol{x}_1^{\text{est}} - \boldsymbol{x}_1 \right\|^2. \tag{5}$$

Crucially, this reconstruction loss optimizes the vector field for *one-step sampling*, enabling the extreme inference efficiency of SignSparK. To enforce temporal coherence and smooth dynamics, we further introduce a *velocity-matching loss* that aligns the predicted inter-frame displacements with the ground truth:

$$\mathcal{L}_{\text{vel}} = \mathbb{E}_{\boldsymbol{x}_1, \boldsymbol{x}_t, t} \left\| (\boldsymbol{x}_{1,f+1}^{\text{est}} - \boldsymbol{x}_{1,f}^{\text{est}}) - (\boldsymbol{x}_{1,f+1} - \boldsymbol{x}_{1,f}) \right\|^2, \tag{6}$$

with $f$ indexing the consecutive frames in the motion sequence. The final objective is then the combination of the CFM and reconstruction losses:

$$\mathcal{L}_{\text{SignSparK}} = \lambda_{\text{CFM}} \mathcal{L}_{\text{CFM}} + \lambda_{\text{recon}} \mathcal{L}_{\text{recon}} + \lambda_{\text{vel}} \mathcal{L}_{\text{vel}}, \tag{7}$$

where $\lambda_{\text{recon}}$ and $\lambda_{\text{vel}}$ weight the auxiliary terms. By explicitly reconstructing the target pose at each timestep via one-step integrations, SignSparK consequently learns more faithful and fluid signing motion, improving generation quality (Sec. 4.3) while requiring far fewer sampling steps.

## 4 Experiments

In this section, we evaluate both of our main contributions: the sign language segmentor **FAST** and the SLP framework **SignSparK**. We first outline the datasets and evaluation metrics used, followed by state-of-the-art comparisons, ablations, and qualitative results. As human evaluation remains the gold standard for assessing sign correctness, we also include a comprehensive user study.

***Datasets.*** For sign segmentation, we follow standard practice [52–54] and adopt MeineDGS [43] as both our training and evaluation benchmark. This corpus offers precise frame-level boundary annotations for continuous signing, making it the only resource suitable for this task. Nonetheless, we show that FAST generalizes well even to unseen signers and datasets in the supplementary.

For SLP, we construct a multilingual training corpus following SOKE [104] by merging Phoenix14T [10], CSLDaily [102], and How2Sign [21]. However, we further extend their setup by adding British Sign Language data from the large-scale BOBSL corpus [1], enhancing linguistic diversity and coverage. To prevent data dominance, only 10% of BOBSL is used for training. For downstream evaluations, SignSparK is rigorously benchmarked across multiple SLP regimes. To ensure fair comparison with prior work, we align our test sets accordingly: (i) Sign Stitching (Gloss-to-Pose, G2P) is evaluated on Phoenix14T, MeineDGS, and BSLCorpus [18] following [89], while (ii) the Text-to-Pose (T2P) and Keyframe-to-Pose (KF2P) tasks are assessed on Phoenix14T, CSLDaily, and How2Sign, following [104]. Additional dataset statistics are provided in the supplementary.

***Evaluation Metrics.*** To evaluate segmentation performance, we follow the metrics adopted in prior work [52–54]. Specifically, we report frame-level F1 score (F1) for BIO label accuracy, Intersection over Union (IoU) to measure temporal overlap between predicted and ground-truth segments, and the Segment Ratio (SR) to quantify over- or under-segmentation based on segment counts.

Meanwhile, to evaluate performance across SLP regimes, we employ task-specific protocols. For Sign Stitching, we follow [89], utilizing Dynamic Time Warping on joint positions (DTW-JPE) to evaluate motion coherence and a Back-Translation (B-T) model to measure semantic intelligibility. For the T2P and KF2P tasks, we adopt the protocol of SOKE [104]. Here, motion and spatial fidelity are measured via DTW errors on both Procrustes-aligned (DTW-PA-JPE) and unaligned (DTW-JPE) joint positions. For KF2P, we also report B-T BLEU-4 scores [55] using the SL-Transformer model [11], consistent with most prior approaches [33, 71, 73, 74, 89]. However, since BLEU scores vary heavily based on B-T configurations, we additionally report *relative* BLEU-4 scores, computed as the percentage drop from the ground-truth, for a more interpretable comparison. As for FLAME facial expressions, we evaluate them separately in Sec. 4.5, as no prior SLP methods to our knowledge report comparable metrics.

### 4.1 Sign Language Segmentation

***State-of-the-Art Comparisons.*** In Tab. 1, we compare our approach (FAST), against recent state-of-the-art (SOTA) segmentation models and demonstrate

**Table 1:** We compare FAST against state-of-the-art sign segmentation models on the MeineDGS [43] dataset.

| Method | Input Features | F1↑ | IoU↑ | SR |
|---|---|---|---|---|
| Single-Modality | | | | |
| IO Tagger [54] | Optical Flow | – | 0.460 | 1.090 |
| BIO Tagger [53] | 3D Body Pose | 0.630 | 0.690 | 1.110 |
| Hands-On [52] | HaMeR | 0.830 | – | – |
| Multi-Modality | | | | |
| BIO Tagger [53] | 3D Pose + Hand | 0.590 | 0.630 | 1.130 |
| Hands-On [52] | 3D Pose + HaMeR | 0.857 | 0.760 | 0.980 |
| **Ours (FAST)** | **WiLoR** | **0.860** | **0.772** | **1.010** |

**Table 2:** We compare SignSparK to prior works on G2P and T2P stitching. We report absolute scores along with relative drops (B-T Score | Drop %) instead, as our B-T models yield higher BLEU-4 scores.

| Method | Phoenix-2014T | | MeineDGS | | BSLCorpus | |
|---|---|---|---|---|---|---|
| | DTW-J↓ | B-T↑ \| Drop↓ | DTW-J↓ | B-T↑ \| Drop↓ | DTW-J↓ | B-T↑ \| Drop↓ |
| *SS GT* [89] | *0.00* | *11.32* | *0.00* | *0.80* | *0.00* | *0.54* |
| SS T2P (Iso) [89] | 59.40 | 3.12 \| ↓72% | 58.10 | 0.22 \| ↓73% | 58.80 | 0.28 \| ↓41% |
| SS T2P (Cont) [89] | 57.20 | 6.67 \| ↓48% | 63.70 | 0.39 \| ↓51% | 57.50 | 0.41 \| ↓24% |
| SS G2P (Iso) [89] | 59.30 | 2.49 \| ↓78% | – | – | – | – |
| SS G2P (Cont) [89] | 58.70 | 5.95 \| ↓47% | – | – | – | – |
| *GT* | *0.00* | *14.77* | *0.00* | *0.99* | *0.00* | *1.23* |
| **SignSparK** | **9.01** | **11.70** \| ↓21% | **8.57** | **0.75** \| ↓24% | **5.24** | 1.13 \| ↓8% |

consistent improvements across all metrics, including F1 (+0.3%), IoU (+1.2%), and Segment Ratio, which is closer to the ideal value of 1 (1.01). Importantly, these gains are achieved using a unimodal setup, without relying on additional modalities. While the absolute improvements are modest, FAST is primarily designed for efficiency and scalability. In particular, FAST significantly outperforms Hands-On [52] in speed, achieving a 45× acceleration in hand detection by leveraging WiLoR and an additional 2× speedup by avoiding 3D body pose extraction. Moreover, FAST also operates with substantially lower dimensionality, requiring $\mathbb{R}^{192}$ for WiLoR's 6D rotations compared to Hands-On's combined hand and body features of $\mathbb{R}^{288+104}$. Together, these characteristics afford us the accuracy, efficiency, and speed needed to scale segmentation to the large datasets used in this work. We show additional timing details in the supplementary.

### 4.2 Sign Language Production

***Sign Stitching Evaluation.*** Sign stitching is an SLP task that generates continuous motion by concatenating isolated glosses. In Tab. 2, we find that SignSparK outperforms all variants of the SOTA Sign Stitcher (SS) [89], achieving lower DTW-JPE across all datasets. Furthermore, minimal degradation from upper-bound B-T scores also confirms that our generated sequences preserve high linguistic fidelity. SignSparK notably achieves this superior performance while conditioning on only three keyframes per gloss, whereas SS relies on full-sequence pose inputs. Finally, our framework also exhibits robust zero-shot generalization, surpassing SS on both the unseen datasets of MeineDGS and BSLCorpus.

***T2P Evaluation.*** We evaluate SignSparK against SOTA T2P models in both gloss-free (GF) and sign-retrieval (SR) regimes (Tab. 3). As GF-T2P requires the direct regression of continuous motion from spoken text, we evaluate SignSparK using zero-keyframe inference; a capability enabled by CFG. Notably, even in the complete absence of guiding spatial anchors, SignSparK still outperforms prior GF-T2P approaches across all metrics on all three datasets. This highlights that despite being optimized primarily via sparse conditioning, SignSparK still internalizes the complex linguistic mapping between text and signing kinematics.

While this confirms strong GF-T2P performance, retrieval-based SLP (SR-T2P, or T2G2P) is the regime we primarily target, as it grounds the signing in

**Table 3: Comparisons with state-of-the-art Text-to-Pose (T2P) models.** SignSparK consistently reduces both body and hand Joint Position Error (JPE) across all datasets in both gloss-free (GF) and sign-retrieval (SR) regimes. For SOKE [104], we report its non-dictionary and final performances for the GF and SR settings, respectively. Results marked with † are reproduced by [104]. B-T metrics are excluded as the required evaluation model and ground-truth baselines are publicly unavailable, while How2Sign is omitted for SR-T2P evaluation due to its lack of gloss annotations.

| Method | How2Sign | | | | CSLDaily | | | | Phoenix-2014T | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | DTW-PA-JPE↓ | | DTW-JPE↓ | | DTW-PA-JPE↓ | | DTW-JPE↓ | | DTW-PA-JPE↓ | | DTW-JPE↓ | |
| | Body | Hand | Body | Hand | Body | Hand | Body | Hand | Body | Hand | Body | Hand |
| Gloss-Free Text-to-Pose (GF-T2P) | | | | | | | | | | | | |
| Prog. Trans.† [71] | 14.15 | 11.57 | 14.74 | 30.17 | 15.98 | 12.91 | 16.30 | 32.63 | 13.67 | 11.95 | 15.01 | 31.77 |
| Text2Mesh† [81] | 13.99 | 13.47 | 15.50 | 32.97 | 13.47 | 12.10 | 13.76 | 30.37 | 13.48 | 12.06 | 14.04 | 31.64 |
| T2S-GPT† [94] | 11.48 | 6.39 | 12.65 | 18.44 | 11.94 | 5.93 | 12.32 | 15.43 | 10.38 | 6.47 | 11.65 | 19.09 |
| S-MotionGPT† [38] | 11.23 | 4.39 | 12.41 | 13.74 | 10.81 | 3.78 | 11.58 | 11.31 | 9.45 | 3.41 | 10.42 | 9.08 |
| SOKE (no dict.) [104] | 7.91 | 3.10 | – | – | 7.58 | 2.17 | – | – | 6.16 | 1.85 | – | – |
| **SignSparK (no KF)** | **7.26** | **2.72** | **6.30** | **11.43** | **7.26** | **2.00** | **6.27** | **10.63** | **5.24** | **1.52** | **4.40** | **7.10** |
| Sign Retrieval Text-to-Pose (SR-T2P) | | | | | | | | | | | | |
| SOKE [104] | 6.82 | 2.35 | 7.75 | 10.08 | 6.24 | 1.71 | 7.38 | 9.68 | 4.77 | 1.38 | 6.04 | 7.72 |
| **SignSparK** | – | – | – | – | **4.87** | **1.37** | **4.89** | **6.56** | **3.68** | **1.18** | **3.51** | **4.82** |

**Table 4: Keyframe-to-Pose (KF2P) evaluation.** We compare SignSparK against a standard Spherical Linear Interpolation (SLERP) baseline, demonstrating that our data-driven approach outperforms traditional rotational interpolation on all metrics.

| Method | How2Sign | | | | | CSLDaily | | | | | Phoenix-2014T | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | DTW-PA-J↓ | | DTW-J↓ | | **B-T**↑ \| Drop↓ | DTW-PA-J↓ | | DTW-J↓ | | **B-T**↑ \| Drop↓ | DTW-PA-J↓ | | DTW-J↓ | | **B-T**↑ \| Drop↓ |
| | Bdy | Hnd | Bdy | Hnd | B-4 \| % | Bdy | Hnd | Bdy | Hnd | B-4 \| % | Bdy | Hnd | Bdy | Hnd | B-4 \| % |
| *GT* | *0.00* | *0.00* | *0.00* | *0.00* | *3.37* | *0.00* | *0.00* | *0.00* | *0.00* | *6.83* | *0.00* | *0.00* | *0.00* | *0.00* | *14.77* |
| SLERP Baseline | 4.18 | 0.98 | 4.36 | 3.76 | 2.78 \| ↓18% | 4.39 | 0.85 | 4.05 | 4.35 | 4.88 \| ↓29% | 4.10 | 0.89 | 3.71 | 4.12 | 9.18 \| ↓38% |
| **SignSparK** | **2.11** | **0.86** | **1.68** | **2.80** | **2.99 \| ↓11%** | **2.50** | **0.63** | **2.02** | **2.71** | **5.85 \| ↓14%** | **2.39** | **0.71** | **1.92** | **2.25** | **11.70 \| ↓21%** |

linguistically valid exemplars, guaranteeing articulation and understandability. We thus compare against SOKE's retrieval-augmented generation by adopting the T2G2P pipeline [74,105]: a text-to-gloss (T2G) model first predicts the gloss sequence, which then indexes an isolated sign dictionary at inference. However, while prior methods retrieve *full video clips* per gloss, we instead retrieve only *three keyframes* per gloss (drastically more storage efficient), pad the keyframes at a fixed 8-frame interval and then jointly condition SignSparK with spoken text. We provide additional implementation details in the supplementary.

Shown in Tab. 3, SignSparK significantly outperforms SOKE. On CSLDaily, we reduce body and hand DTW-PA-JPE to 4.87 (↓22%) and 1.37 (↓20%), respectively, with comparable reductions of 23% and 14% on Phoenix14T. Additionally, our reconstruction-based CFM also achieve unprecedented efficiency, accelerating inference by 100× (10-step SignSparK's 0.01s/vid vs. SOKE's 1.55s [104]). These sampling steps can be further reduced to flexibly trade fidelity for speed.

***KF2P Evaluation.*** Lastly, we benchmark SignSparK against a SLERP baseline on the newly proposed KF2P regime (Tab. 4), and find clear performance gains

**Table 5:** Ablation study on FAST's architectural choices.

| Model Variant | F1↑ | IoU↑ |
|---|---|---|
| Single-Stream (baseline) | 0.848 | 0.744 |
| Two-Stream | 0.850 | 0.747 |
| + Full Temporal Res. (*No Downsampling*) | 0.860 | 0.765 |
| + Temporal Conv. (*Kernel Size* = 3) | **0.861** | **0.768** |
| + Dropout | 0.857 | 0.761 |

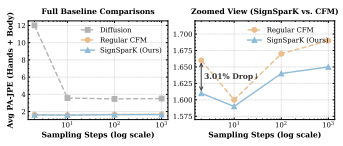


**Fig. 4:** Comparing SignSparK to diffusion and CFM models.

**Table 6: Ablation on keyframe selection policy and loss configurations.** We evaluate the keyframe selection strategies and the contribution of each loss term towards body and hand reconstruction.

| Ablation Experiment | Body | | Hand | |
|---|---|---|---|---|
| | PA-MPJPE$_\downarrow$ | MPJPE$_\downarrow$ | PA-MPJPE$_\downarrow$ | MPJPE$_\downarrow$ |
| KEYFRAME SELECTION POLICY | | | | |
| Random | $2.12_{\pm 1.0}$ | $1.76_{\pm 0.8}$ | $1.47_{\pm 1.0}$ | $4.37_{\pm 2.4}$ |
| **Segments Only** | $\mathbf{1.93}_{\pm 0.8}$ | $\mathbf{1.61}_{\pm 0.7}$ | $\mathbf{1.27}_{\pm 0.5}$ | $\mathbf{3.94}_{\pm 1.4}$ |
| Segments w/ Random Masking | $2.02_{\pm 0.9}$ | $1.68_{\pm 0.7}$ | $1.31_{\pm 0.5}$ | $4.11_{\pm 1.4}$ |
| LOSS CONFIGURATION | | | | |
| $\mathcal{L}_{\text{CFM}}$ | $1.93_{\pm 0.8}$ | $1.61_{\pm 0.7}$ | $1.27_{\pm 0.5}$ | $3.97_{\pm 1.4}$ |
| $\mathcal{L}_{\text{recon}}$ | $2.02_{\pm 0.9}$ | $1.68_{\pm 0.7}$ | $1.27_{\pm 0.5}$ | $4.09_{\pm 1.4}$ |
| $\mathcal{L}_{\text{vel}}$ | $8.43_{\pm 2.2}$ | $7.12_{\pm 1.9}$ | $1.51_{\pm 0.5}$ | $9.60_{\pm 2.7}$ |
| $\mathcal{L}_{\text{CFM}} + \mathcal{L}_{\text{recon}}$ | $1.92_{\pm 0.8}$ | $1.60_{\pm 0.7}$ | $1.27_{\pm 0.5}$ | $3.91_{\pm 1.3}$ |
| $\mathcal{L}_{\text{CFM}} + \mathcal{L}_{\text{vel}}$ | $1.95_{\pm 0.8}$ | $1.63_{\pm 0.7}$ | $1.27_{\pm 0.5}$ | $4.01_{\pm 1.4}$ |
| $\boldsymbol{\mathcal{L}_{\text{CFM}} + \mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{vel}}}$ | $\mathbf{1.92}_{\pm 0.8}$ | $\mathbf{1.60}_{\pm 0.7}$ | $\mathbf{1.26}_{\pm 0.5}$ | $\mathbf{3.91}_{\pm 1.4}$ |

across all datasets. This indicates that SignSparK successfully learnt the complex dynamics of signing rather than merely interpolating between sparse frames.

### 4.3 Ablation Studies

***Sign Segmentor.*** In Tab. 5, we ablate the design components of FAST. Using a single-stream design, consistent with prior work, yields an F1 of 0.85 and IoU of 0.74. Adopting a dual-stream architecture, where each stream processes one hand independently, then leads to minor gains in both metrics. Removing the 2× temporal downsampling used in [52] further improves performance (F1 +1%, IoU +1.8%). Adding temporal convolution layers further boosts performances.

***Generative Architecture Comparison.*** Fig. 4 compares SignSparK against standard diffusion and CFM models across 1, 10, 100, and 1000 sampling steps. Here, the noise-based diffusion model severely underperforms in low-step regimes, requiring hundreds of steps to achieve competitive fidelity due to indirect noise optimization. Conversely, while SignSparK and standard CFM perform comparably at 10 steps, our framework demonstrates superior robustness elsewhere; notably yielding a 3% gain in single-step generation. This stable fidelity retention stems directly from our reconstruction-based formulation, which explicitly penalizes pose prediction errors and enforces one-step sampling during training.

***Keyframe Selection Strategy.*** We then ablate the effectiveness of our segment-based keyframe selection policy by comparing it against random sampling and random segment masking during training (Tab. 6). Random sampling yields the weakest results, suggesting that unstructured frame selection disrupts the model's ability to generate coherent signing motion. In contrast, our segment-based keyframes consistently yields strong performances across all metrics. Further adding random masking as augmentation then degrades results, indicating that the model learns better when trained directly on the structured anchors. Further ablations, such as joint masking, are provided in the supplementary.

***Loss Configuration.*** We analyze the effects of different loss combinations in SignSparK (Tab. 6). The baseline $\mathcal{L}_{\text{CFM}}$ already achieves strong performance,

**Table 7: Dataset and language token.** We ablate the contribution of multilingual training data and verify whether prepending language token identifiers (e.g., <ASL>) to text improves model performance.

| Lang. | Datasets | | | | Body | | Hand | |
|---|---|---|---|---|---|---|---|---|
| Tok. | Ph-T | CSL | H2S | BSL | PA-JPE$_\downarrow$ | JPE$_\downarrow$ | PA-JPE$_\downarrow$ | JPE$_\downarrow$ |
| ✗ | ✓ | ✗ | ✗ | ✗ | $2.87_{\pm 1.2}$ | $2.49_{\pm 1.0}$ | $1.51_{\pm 0.5}$ | $5.20_{\pm 1.9}$ |
| ✗ | ✓ | ✓ | ✗ | ✗ | $2.66_{\pm 1.1}$ | $2.29_{\pm 0.9}$ | $1.36_{\pm 0.5}$ | $4.58_{\pm 1.6}$ |
| ✗ | ✓ | ✓ | ✓ | ✗ | $1.92_{\pm 0.8}$ | $1.61_{\pm 0.7}$ | $1.27_{\pm 0.5}$ | $3.90_{\pm 1.4}$ |
| ✗ | ✓ | ✓ | ✓ | ✓ | $1.93_{\pm 0.8}$ | $1.61_{\pm 0.7}$ | $1.27_{\pm 0.5}$ | $3.94_{\pm 1.4}$ |
| ✓ | ✓ | ✓ | ✓ | ✓ | $\mathbf{1.90}_{\pm 0.8}$ | $\mathbf{1.59}_{\pm 0.7}$ | $\mathbf{1.27}_{\pm 0.5}$ | $\mathbf{3.90}_{\pm 1.4}$ |

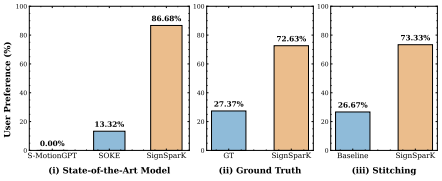


**Fig. 5: SignSparK user study.** We conduct a study with *6 Deaf and 10 hearing BSL signers*, comparing SignSparK against SOTA and baseline models.

outperforming models trained with only $\mathcal{L}_{\text{recon}}$ or $\mathcal{L}_{\text{vel}}$. Adding $\mathcal{L}_{\text{recon}}$ further reduces joint errors, indicating complementary benefits between the CFM and reconstruction objectives. Combining all three losses then yields the best overall performance, with a 2:1:1 weighting of $(\mathcal{L}_{\text{CFM}}, \mathcal{L}_{\text{recon}}, \mathcal{L}_{\text{vel}})$ working best (see supplementary). We retain $\mathcal{L}_{\text{vel}}$ despite its small numerical contribution, as its role is qualitative: by penalizing inter-frame deviations from the pseudo-GT, it suppresses *micro-jitters* that JPE-style metrics fail to capture. Overall, $\mathcal{L}_{\text{CFM}}$ lays a strong foundation, but the regularizations are what enforce the reconstruction accuracy and perceptual smoothness in SignSparK's generated motion.

***Contribution of Datasets.*** In Tab. 7, we find that training SignSparK on more multilingual datasets generally improves downstream performance. However, this trend was initially disrupted by the inclusion of BOBSL, which increased overall error rates. Interestingly, introducing a language identifier token then resolves this issue, suggesting that the model was confusing American and British Sign Languages. Since both shared English as a spoken language, the model's ability to produce language-specific signing motion was likely hindered.

***User Study.*** Beyond quantitative metrics, we also conducted a forced-choice user study with 16 signers, evaluating perceived naturalness and visual alignment (Fig. 5). Across three distinct scenarios, SignSparK was overwhelmingly preferred. Against SOTA baselines (SOKE and S-MotionGPT), our model was favored in 86.68% of trials. Interestingly, SignSparK was even preferred over pseudo-ground-truth extractions in 72.63% of cases, as our learned signing prior significantly reduced the temporal jitter inherent to frame-wise 3D estimators. Finally, for sign stitching, our keyframe-conditioned generation also outperformed standard interpolation (73.33% preference) due to its more fluid coarticulation. Comprehensive study details are provided in the supplementary material.

### 4.4 Qualitative Results

In Fig. 6, we present qualitative examples on all four training datasets. On the challenging BOBSL dataset, SignSparK was able to faithfully reproduce body posture, hand positions, and common handshapes such as open palms and pointing gestures when given keyframe anchors. While occasional inaccuracies, such

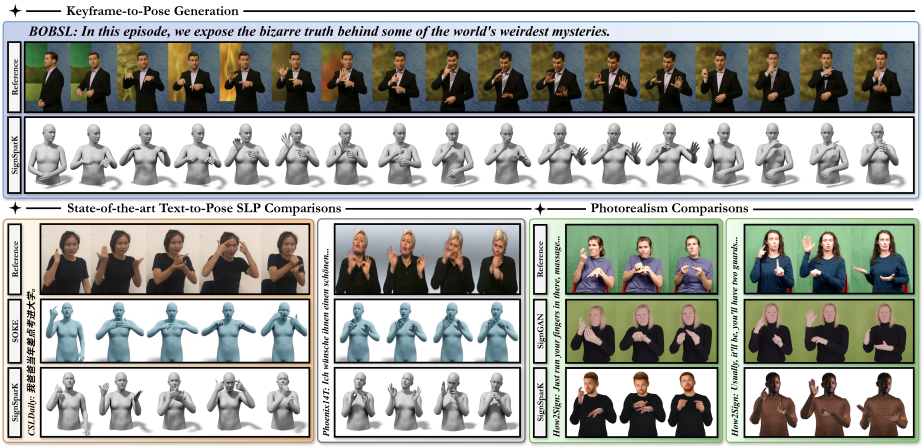


**Fig. 6: Qualitative Results.** SignSparK exhibits strong visual results across diverse tasks and sign languages. Top: KF2P predictions on BOBSL. Bottom-left: SOTA T2P Comparisons against SOKE [104] on CSL-Daily and Phoenix14T. Bottom-right: Photorealism comparison between SignSparK with 3DGS and SignGAN [74] on How2Sign.

as incomplete hand closures, were observed, these were often caused by imprecisions in the ground-truth SMPL-X and MANO reconstructions provided by current single-view extraction approaches. On T2P SLP, SignSparK demonstrates clear advantages over SOKE [104], as we find that SignSparK achieves consistently accurate signing with coherent articulations across consecutive signs. To render photorealistic avatars, we pair SignSparK with HuGeDiff [35] due to its lightweight design, but note that SignSparK is inherently compatible with any SMPL-based 3DGS pipeline (e.g., GUAVA [98]). As seen in Fig. 6, the 3DGS renderings preserve precise hand geometry and yield more intelligible signing than GAN-based pipelines such as SignGAN [74], which struggle to model fine-grained bimanual interactions. Further video comparisons are also provided in the supplementary to better demonstrate SignSparK's generative performance.

### 4.5 Facial Expression Evaluation

Non-manual features, and facial expressions in particular, carry essential grammatical and prosodic information in sign language [61], yet to our knowledge, prior SMPL-based SLP methods have rarely, if ever, modelled them explicitly. Given how central facial articulations are to comprehension, we considered its inclusion in SignSparK essential. Identifying a feedforward extraction pipeline of sufficient fidelity for the subtle expressions characteristic of sign language proved a persistent challenge, but we ultimately adopted TEASER [50] for its balance of scalability and reconstruction quality, and used it to extract FLAME parameters across our datasets. As shown in Fig. 7, training SignSparK on these extractions yields tractable expressions that match both the reference video and the pseudo-GT, despite being synthesised from only sparse keyframes. As facial expression


**Fig. 7: Facial Expression Comparison.** We provide close-up comparisons between SignSparK's synthesised facial expressions against the pseudo-ground truth from TEASER [50] and confirm that our formulation transfers seamlessly to FLAME parameters.

**Table 8: Facial PA-VPE** values on How2Sign and CSL-Daily. Reported as a standalone reference as no comparable FLAME-based SLP baselines exist.

| Dataset | PA-VPE$_{\downarrow}$ |
|---|---|
| How2Sign | $0.07_{\pm 0.02}$ |
| CSL-Daily | $0.06_{\pm 0.02}$ |

metrics have, to our knowledge, not been reported on standard SLP benchmarks, we provide SignSparK's PA-VPE$_{\text{face}}$ as a standalone reference, measuring 0.07 on How2Sign and 0.06 on CSL-Daily (Tab. 8). We hope these initial numbers can serve as a starting point for future SMPL-based SLP methods.

## 5 Conclusions

This paper presents a unified framework for large-scale SLP built upon two core contributions. First, we introduce FAST, an ultra-efficient segmentor that establishes a new SOTA and enables the massive-scale extraction of linguistic keyframes. Second, we propose SignSparK, a CFM model driven by a novel training paradigm: synthesizing high-fidelity 3D signing motion directly from sparse keyframes. Coupled with a reconstruction-based objective, SignSparK achieves a $>100\times$ efficiency gain over prior methods. This efficiency directly enables our expansion across ASL, BSL, CSL, and DGS, establishing the largest multilingual SLP framework to date. Finally, by integrating 3DGS, SignSparK also overcomes the limitations of bare meshes to render photorealistic, identity-diverse avatars. Supported by extensive evaluations demonstrating SOTA kinematic fidelity and robust generalization, this work equips the community with a highly scalable segmentation tool and a multilingual generative prior for diverse SLP tasks.

## Acknowledgements

This work was supported by EPSRC grant APP24554 (SignGPT-EP/Z535370/1), EPSRC grant APP78083 (UMCS UKRI3927), as well as through funding from Google.org via the AI for Global Goals scheme. The authors acknowledge the use of Isambard-AI National AI Research Resource (AIRR) funded by UK DSIT via UKRI and STFC [ST/AIRR/I-A-I/1023]. Jianhe Low additionally acknowledges a bursary from the Rabin Ezra Scholarship Trust. This work reflects only the authors' views and the funders are not responsible for any use that may be made of the information it contains.

## References


1. Albanie, S., Varol, G., Momeni, L., Bull, H., Afouras, T., Chowdhury, H., Fox, N., Woll, B., Cooper, R., McParland, A., Zisserman, A.: BOBSL: BBC-Oxford british sign language dataset. arXiv preprint arXiv:2111.03635 (2021)
2. Arkushin, R.S., Moryossef, A., Fried, O.: Ham2pose: Animating sign language notation into pose sequences. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 21046–21056 (2023)
3. Baltatzis, V., Potamias, R.A., Ververas, E., Sun, G., Deng, J., Zafeiriou, S.: 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 (CVPR). pp. 1985–1995 (2024)
4. Bangham, J.A., Cox, S., Elliott, R., Glauert, J.R., Marshall, I., Rankov, S., Wells, M.: Virtual signing: Capture, animation, storage and transmission-an overview of the visicast project. In: IEE Seminar on Speech and Language Processing for Disabled and Elderly People (Ref. No. 2000/025). pp. 6–1. IET (2000)
5. Bensabath, L., Petrovich, M., Varol, G.: Text-driven 3d hand motion generation from sign language data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 23095–23105 (2026)
6. Buehler, P., Zisserman, A., Everingham, M.: Learning sign language by watching tv (using weakly aligned subtitles). In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2961–2968 (2009)
7. Bull, H., Afouras, T., Varol, G., Albanie, S., Momeni, L., Zisserman, A.: Aligning subtitles in sign language videos. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 11532–11541 (2021)
8. Bull, H., Gouiffès, M., Braffort, A.: Automatic segmentation of sign language into subtitle-units. In: European Conference on Computer Vision Workshops (ECCVW). pp. 186–198. Springer (2020)
9. Cai, H., Bai, C., Tai, Y.W., Tang, C.K.: Deep video generation, prediction and completion of human action sequences. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 366–382 (2018)
10. Camgoz, N.C., Hadfield, S., Koller, O., Ney, H., Bowden, R.: Neural sign language translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7784–7793 (2018)
11. Camgoz, N.C., Koller, O., Hadfield, S., Bowden, R.: 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 (CVPR). pp. 10023–10033 (2020)
12. Carlsson, F., Eisen, P., Rekathati, F., Sahlgren, M.: Cross-lingual and multilingual clip. In: International Conference on Language Resources and Evaluation (LREC). pp. 6848–6854 (2022)
13. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4724–4733 (2017)
14. Chen, S., Wang, Q., Wang, Q.: Semantic-driven diffusion for sign language production with gloss-pose latent spaces alignment. Computer Vision and Image Understanding (CVIU) **246**, 104050 (2024)
15. Chen, X., Jiang, B., Liu, W., Huang, Z., Fu, B., Chen, T., Yu, G.: Executing your commands via motion diffusion in latent space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 18000–18010 (2023)

16. Cohan, S., Tevet, G., Reda, D., Peng, X.B., van de Panne, M.: Flexible motion in-betweening with diffusion models. In: Special Interest Group on Computer Graphics and Interactive Techniques (SIGGRAPH). pp. 1–9. ACM (2024)
17. Cooper, H., Bowden, R.: Learning signs from subtitles: A weakly supervised approach to sign language recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2568–2574 (2009)
18. Cormier, K., Fenlon, J., Gulamani, S., Smith, S.: Bsl corpus annotation conventions (2017), https://bslcorpusproject.org/wp-content/uploads/BSLCorpus_AnnotationConventions_v3.0_-March2017.pdf, accessed: 2026-06-24
19. Cox, S., Lincoln, M., Tryggvason, J., Nakisa, M., Wells, M., Tutt, M., Abbott, S.: Tessa, a system to aid communication with deaf people. In: Proceedings of the International ACM Conference on Assistive Technologies (ASSETS). pp. 205–212 (2002)
20. Dong, L., Chaudhary, L., Xu, F., Wang, X., Lary, M., Nwogu, I.: Signavatar: Sign language 3d motion reconstruction and generation. In: IEEE International Conference on Automatic Face and Gesture Recognition (FG). pp. 1–10. IEEE (2024)
21. Duarte, A., Palaskar, S., Ventura, L., Ghadiyaram, D., DeHaan, K., Metze, F., Torres, J., Giro-i Nieto, X.: How2sign: A large-scale multimodal dataset for continuous american sign language. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2735–2744 (2021)
22. Efthimiou, E., Fotinea, S.E., Hanke, T., Glauert, J., Bowden, R., Braffort, A., Collet, C., Maragos, P., Lefebvre-Albaret, F.: The dicta-sign wiki: Enabling web communication for the deaf. In: International Conference on Computers for Handicapped Persons (ICCHP). pp. 205–212. Springer (2012)
23. ElGhoul, O., Jemni, M.: Websign: A system to make and interpret signs using 3d avatars. In: Proceedings of the International Workshop on Sign Language Translation and Avatar Technology (SLTAT). vol. 23 (2011)
24. Fang, S., Sui, C., Zhou, Y., Zhang, X., Zhong, H., Tian, Y., Chen, C.: Signdiff: Diffusion model for american sign language production. In: IEEE International Conference on Automatic Face and Gesture Recognition (FG). pp. 1–11. IEEE (2025)
25. Farag, I., Brock, H.: Learning motion disfluencies for automatic sign language segmentation. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 7360–7364 (2019)
26. Fish, E., Bowden, R.: Geo-sign: Hyperbolic contrastive regularisation for geometrically aware sign language translation. In: Advances in Neural Information Processing systems (NeurIPS) (2025)
27. Guo, C., Mu, Y., Javed, M.G., Wang, S., Cheng, L.: Momask: Generative masked modeling of 3d human motions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1900–1910 (2024)
28. Guo, C., Zou, S., Zuo, X., Wang, S., Ji, W., Li, X., Cheng, L.: Generating diverse and natural 3d human motions from text. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5152–5161 (2022)
29. Guo, C., Zuo, X., Wang, S., Cheng, L.: Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 580–597. Springer (2022)

30. Guo, Y., Yang, C., Rao, A., Agrawala, M., Lin, D., Dai, B.: Sparsectrl: Adding sparse controls to text-to-video diffusion models. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 330–348. Springer (2024)
31. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in Neural Information Processing systems (NeurIPS) **33**, 6840–6851 (2020)
32. Ho, J., Salimans, T.: Classifier-free diffusion guidance. In: Advances in Neural Information Processing systems (NeurIPS) Workshop (2021)
33. Huang, W., Pan, W., Zhao, Z., Tian, Q.: Towards fast and high-quality sign language production. In: Proceedings of the ACM International Conference on Multimedia (MM). pp. 3172–3181. ACM (2021)
34. Ivashechkin, M., Mendez, O., Bowden, R.: Improving 3d pose estimation for sign language. In: IEEE International Conference on Acoustics, Speech and Signal Processing Workshops (ICASSPW). pp. 1–5. IEEE (2023)
35. Ivashechkin, M., Mendez, O., Bowden, R.: HuGeDiff: 3D Human Generation via Diffusion with Gaussian Splatting. In: British Machine Vision Conference (BMVC). BMVA (2025)
36. Ivashechkin, M., Mendez, O., Bowden, R.: Signsplat: Rendering sign language via gaussian splatting. arXiv preprint arXiv:2505.02108 (2025)
37. Jain, S., Watson, D., Tabellion, E., Poole, B., Kontkanen, J., et al.: Video interpolation with diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7341–7351 (2024)
38. Jiang, B., Chen, X., Liu, W., Yu, J., Yu, G., Chen, T.: Motiongpt: Human motion as a foreign language. Advances in Neural Information Processing systems (NeurIPS) **36**, 20067–20079 (2023)
39. Karunratanakul, K., Preechakul, K., Suwajanakorn, S., Tang, S.: Guided motion diffusion for controllable human motion synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2151–2162 (2023)
40. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. In: Transactions on Graphics (TOG). vol. 42, pp. 139–1. ACM (2023)
41. Khan, N., Wu, B., Tan, S., Ishi, C.T., Nakadai, K.: Signflow: End-to-end sign language generation for one-to-many modeling using conditional flow matching. In: Proceedings of the 27th International Conference on Multimodal Interaction (ICMI). pp. 173–180 (2025)
42. Kipp, M., Nguyen, Q., Heloir, A., Matthes, S.: Assessing the deaf user perspective on sign language avatars. In: Proceedings of the International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS). pp. 107–114 (2011)
43. Konrad, R., Hanke, T., Langer, G., Blanck, D., Bleicken, J., Hofmann, I., Jeziorski, O., König, L., König, S., Nishio, R., Regen, A., Salden, U., Wagner, S., Worseck, S., Böse, O., Jahn, E., Schulder, M.: Meine dgs – annotiert. öffentliches korpus der deutschen gebärdensprache, 3. release / my dgs – annotated. public corpus of german sign language, 3rd release (2020). `https://doi.org/10.25592/dgs.corpus-3.0`, `https://doi.org/10.25592/dgs.corpus-3.0`, accessed: 2026-06-24
44. Lee, H.Y., Yang, X., Liu, M.Y., Wang, T.C., Lu, Y.D., Yang, M.H., Kautz, J.: Dancing to music. Advances in Neural Information Processing systems (NeurIPS) **32** (2019)
45. Li, T., Bolkart, T., Black, M.J., Li, H., Romero, J.: Learning a model of facial shape and expression from 4d scans. In: Transactions on Graphics (TOG). vol. 36, pp. 194–1. ACM (2017)

46. Li, Z., Zhou, W., Zhao, W., Wu, K., Hu, H., Li, H.: Uni-sign: Toward unified sign language understanding at scale. International Conference on Learning Representations (ICLR) (2025)
47. Lillo-Martin, D.C., Gajewski, J.: One grammar or two? sign languages and the nature of human language. Wiley Interdisciplinary Reviews. Cognitive Science **5**(4), 387–401 (2014)
48. Lin, P.: Handdiffuse: generative controllers for two-hand interactions via diffusion models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 5280–5288 (2025)
49. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. In: International Conference on Learning Representations (ICLR) (2023)
50. Liu, Y., Zhu, L., Lin, L., Zhu, Y., Zhang, A., Li, Y.: Teaser: Token enhanced spatial modeling for expressions reconstruction. In: International Conference on Learning Representations (ICLR) (2025)
51. Low, J., Sincan, O.M., Bowden, R.: Sage: Segment-aware gloss-free encoding for token-efficient sign language translation. In: International Conference on Computer Vision Workshops (ICCVW). IEEE (2025)
52. Low, J., Walsh, H., Sincan, O.M., Bowden, R.: Hands-on: Segmenting individual signs from continuous sequences. In: IEEE International Conference on Automatic Face and Gesture Recognition (FG) (2025)
53. Moryossef, A., Jiang, Z., Müller, M., Ebling, S., Goldberg, Y.: Linguistically motivated sign language segmentation. In: Conference on Empirical Methods in Natural Language Processing (EMNLP) (2023)
54. Moryossef, A., Tsochantaridis, I., Aharoni, R., Ebling, S., Narayanan, S.: Real-time sign language detection using human pose estimation. In: European Conference on Computer Vision Workshops (ECCVW). pp. 237–248. Springer (2020)
55. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL). pp. 311–318 (2002)
56. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A., Tzionas, D., Black, M.J.: 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 (CVPR). pp. 10975–10985 (2019)
57. Pavlakos, G., Shan, D., Radosavovic, I., Kanazawa, A., Fouhey, D., Malik, J.: Reconstructing hands in 3d with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9826–9836 (2024)
58. Perniss, P., Pfau, R., Steinbach, M.: Can't you see the difference? sources of variation in sign language structure. In: Visible variation: Cross-linguistic studies in sign language narratives, pp. 1–34. Mouton de Gruyter (2007)
59. Petrovich, M., Black, M.J., Varol, G.: Action-conditioned 3d human motion synthesis with transformer vae. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 10985–10995 (2021)
60. Petrovich, M., Black, M.J., Varol, G.: Temos: Generating diverse human motions from textual descriptions. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 480–497. Springer (2022)
61. Pfau, R., Quer, J.: Nonmanuals: their grammatical and prosodic roles. Sign Languages pp. 381–402 (2010)
62. Plappert, M., Mandery, C., Asfour, T.: The kit motion-language dataset. Big data **4**(4), 236–252 (2016)

63. Potamias, R.A., Zhang, J., Deng, J., Zafeiriou, S.: Wilor: End-to-end 3d hand localization and reconstruction in-the-wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12242–12254 (2025)
64. Punnakkal, A.R., Chandrasekaran, A., Athanasiou, N., Quiros-Ramirez, A., Black, M.J.: Babel: Bodies, action and behavior with english labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 722–731 (2021)
65. Renz, K., Stache, N.C., Albanie, S., Varol, G.: Sign language segmentation with temporal convolutional networks. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 2135–2139 (2021)
66. Renz, K., Stache, N.C., Fox, N., Varol, G., Albanie, S.: Sign segmentation with changepoint-modulated pseudo-labelling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 3403–3412 (2021)
67. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: modeling and capturing hands and bodies together. In: Transactions on Graphics (TOG). vol. 36, pp. 1–17. ACM (2017)
68. Santemiz, P., Aran, O., Saraclar, M., Akarun, L.: Automatic sign segmentation from continuous signing via multiple sequence alignment. In: International Conference on Computer Vision Workshops (ICCVW). pp. 2001–2008 (2009)
69. Sárándi, I., Pons-Moll, G.: Neural localizer fields for continuous 3d human pose and shape estimation. Advances in Neural Information Processing systems (NeurIPS) **37**, 140032–140065 (2024)
70. Saunders, B., Camgoz, N.C., Bowden, R.: Adversarial training for multi-channel sign language production. In: British Machine Vision Conference (BMVC). British Machine Vision Association (2020)
71. Saunders, B., Camgoz, N.C., Bowden, R.: Progressive transformers for end-to-end sign language production. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 687–705. Springer (2020)
72. Saunders, B., Camgoz, N.C., Bowden, R.: Continuous 3d multi-channel sign language production via progressive transformers and mixture density networks. International Journal of Computer Vision (IJCV) **129**(7), 2113–2135 (2021)
73. Saunders, B., Camgoz, N.C., Bowden, R.: Mixed signals: Sign language production via a mixture of motion primitives. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 1919–1929 (2021)
74. Saunders, B., Camgoz, N.C., Bowden, R.: 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 (CVPR). pp. 5141–5151 (2022)
75. Shafir, Y., Tevet, G., Kapon, R., Bermano, A.H.: Human motion diffusion as a generative prior. In: International Conference on Learning Representations (ICLR) (2024)
76. Shlizerman, E., Dery, L., Schoen, H., Kemelmacher-Shlizerman, I.: Audio to body dynamics. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7574–7583 (2018)
77. Sincan, O.M., Low, J.H., Asasi, S., Bowden, R.: Gloss-free sign language translation: An unbiased evaluation of progress in the field. Computer Vision and Image Understanding (CVIU) p. 104498 (2025)
78. Stokoe, W.C.: Language in hand: Why sign came before speech. Gallaudet University Press (2001)

79. Stoll, S., Camgöz, N.C., Hadfield, S., Bowden, R.: Sign language production using neural machine translation and generative adversarial networks. In: British Machine Vision Conference (BMVC). British Machine Vision Association (2018)
80. Stoll, S., Camgoz, N.C., Hadfield, S., Bowden, R.: Text2sign: towards sign language production using neural machine translation and generative adversarial networks. International Journal of Computer Vision (IJCV) **128**(4), 891–908 (2020)
81. Stoll, S., Mustafa, A., Guillemaut, J.Y.: There and back again: 3d sign language generation from text using back-translation. In: Proceedings of the International Conference on 3D Vision (3DV). pp. 187–196. IEEE (2022)
82. Symeonidis-Herzig, A., Low, J., Sincan, O.M., Bowden, R.: M3t: Discrete multi-modal motion tokens for sign language production. arXiv preprint arXiv:2603.23617 (2026)
83. Tang, S., Hong, R., Guo, D., Wang, M.: Gloss semantic-enhanced network with online back-translation for sign language production. In: Proceedings of the ACM International Conference on Multimedia (MM). pp. 5630–5638 (2022)
84. Tang, S., Xue, F., Wu, J., Wang, S., Hong, R.: Gloss-driven conditional diffusion models for sign language production. ACM Transactions on Multimedia Computing, Communications and Applications (TOMM) **21**(4), 1–17 (2025)
85. Tanzer, G., Zhang, B.: Youtube-sl-25: A large-scale, open-domain multilingual sign language parallel corpus. International Conference on Learning Representations (ICLR) (2025)
86. Tevet, G., Gordon, B., Hertz, A., Bermano, A.H., Cohen-Or, D.: Motionclip: Exposing human motion generation to clip space. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 358–374. Springer (2022)
87. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-or, D., Bermano, A.H.: Human motion diffusion model. In: International Conference on Learning Representations (ICLR) (2023)
88. Varol, G., Momeni, L., Albanie, S., Afouras, T., Zisserman, A.: Scaling up sign spotting through sign language dictionaries. International Journal of Computer Vision (IJCV) **130**(6), 1416–1439 (2022)
89. Walsh, H.T., Saunders, B., Bowden, R.: Sign stitching: A novel approach to sign language production. In: British Machine Vision Conference (BMVC). British Machine Vision Association (2024)
90. Wang, C., Deng, Z., Jiang, Z., Yin, Y., Shen, F., Cheng, Z., Ge, S., Gan, S., Gu, Q.: Advanced sign language video generation with compressed and quantized multi-condition tokenization. Advances in Neural Information Processing systems (NeurIPS) **38**, 79519–79545 (2026)
91. Wang, X., Zhou, B., Curless, B., Kemelmacher-Shlizerman, I., Holynski, A., Seitz, S.: Generative inbetweening: Adapting image-to-video models for keyframe interpolation. In: International Conference on Learning Representations (ICLR) (2025)
92. Wang, Z., Yu, P., Zhao, Y., Zhang, R., Zhou, Y., Yuan, J., Chen, C.: Learning diverse stochastic human-action generators by learning smooth latent transitions. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 12281–12288 (2020)
93. Xing, J., Xia, M., Zhang, Y., Chen, H., Yu, W., Liu, H., Liu, G., Wang, X., Shan, Y., Wong, T.T.: Dynamicrafter: Animating open-domain images with video diffusion priors. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 399–417. Springer (2024)
94. Yin, A., Li, H., Shen, K., Tang, S., Zhuang, Y.: T2s-gpt: Dynamic vector quantization for autoregressive sign language production from text. In: Proceedings of

the Annual Meeting of the Association for Computational Linguistics (ACL). pp. 3345–3356 (2024)
95. Yu, P., Zhao, Y., Li, C., Yuan, J., Chen, C.: Structure-aware human-action generation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 18–34. Springer (2020)
96. Yu, Z., Huang, S., Cheng, Y., Birdal, T.: Signavatars: A large-scale 3d sign language holistic motion dataset and benchmark. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 1–19. Springer (2024)
97. Zelinka, J., Kanis, J.: Neural sign language synthesis: Words are our glosses. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 3395–3403 (2020)
98. Zhang, D., Liu, Y., Lin, L., Zhu, Y., Li, Y., Qin, M., Li, Y., Wang, H.: Guava: Generalizable upper body 3d gaussian avatar. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 14205–14217 (2025)
99. Zhang, J., Zhang, Y., Cun, X., Zhang, Y., Zhao, H., Lu, H., Shen, X., Shan, Y.: Generating human motion from textual descriptions with discrete representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14730–14740 (2023)
100. Zhang, M., Cai, Z., Pan, L., Hong, F., Guo, X., Yang, L., Liu, Z.: Motiondiffuse: Text-driven human motion generation with diffusion model. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **46**(6), 4115–4128 (2024)
101. Zhou, B., Chen, Z., Clapés, A., Wan, J., Liang, Y., Escalera, S., Lei, Z., Zhang, D.: Gloss-free sign language translation: Improving from visual-language pretraining. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 20871–20881 (2023)
102. Zhou, H., Zhou, W., Qi, W., Pu, J., Li, H.: Improving sign language translation with monolingual data by sign back-translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1316–1325 (2021)
103. Zhou, Y., Barnes, C., Lu, J., Yang, J., Li, H.: On the continuity of rotation representations in neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5745–5753 (2019)
104. Zuo, R., Potamias, R.A., Ververas, E., Deng, J., Zafeiriou, S.: Signs as tokens: A retrieval-enhanced multilingual sign language generator. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 23806–23816 (2025)
105. Zuo, R., Wei, F., Chen, Z., Mak, B., Yang, J., Tong, X.: A simple baseline for spoken language to sign language translation with 3d avatars. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 36–54. Springer (2024)
106. Zwitserlood, I., Verlinden, M., Ros, J., Van Der Schoot, S., Netherlands, T.: Synthetic signing for the deaf: Esign. In: Proceedings of the Conference and Workshop on Assistive Technologies for Vision and Hearing Impairment (CVHI). vol. 1 (2004)