SignAvatars: A Large-scale 3D Sign Language Holistic Motion Dataset and Benchmark

Zhengdi Yu[1,2†], Shaoli Huang[2⋆], Yongkang Cheng[2], and Tolga Birdal[1]

[1] Imperial College London, London, United Kingdom
[2] Tencent AI Lab, Shenzhen, China

Abstract. We present SignAvatars[3], the first large-scale, multi-prompt 3D sign language (SL) motion dataset designed to bridge the communication gap for Deaf and hard-of-hearing individuals. While there has been an exponentially growing number of research regarding digital communication, the majority of existing communication technologies primarily cater to spoken or written languages, instead of SL, the essential communication method for Deaf and hard-of-hearing communities. Existing SL datasets, dictionaries, and sign language production (SLP) methods are typically limited to 2D as annotating 3D models and avatars for SL is usually an entirely manual and labor-intensive process conducted by SL experts, often resulting in unnatural avatars. In response to these challenges, we compile and curate the SignAvatars dataset, which comprises 70,000 videos from 153 signers, totaling 8.34 million frames, covering both isolated signs and continuous, co-articulated signs, with multiple prompts including HamNoSys, spoken language, and words. To yield 3D holistic annotations, including meshes and biomechanically-valid poses of body, hands, and face, as well as 2D and 3D keypoints, we introduce an automated annotation pipeline operating on our large corpus of SL videos. SignAvatars facilitates various tasks such as 3D sign language recognition (SLR) and the novel 3D SL production (SLP) from diverse inputs like text scripts, individual words, and HamNoSys notation. Hence, to evaluate the potential of SignAvatars, we further propose a unified benchmark of 3D SL holistic motion production. We believe that this work is a significant step forward towards bringing the digital world to the Deaf and hard-of-hearing communities as well as people interacting with them.

Keywords: Sign Language · Digital Avatars · Hand Motion

1 Introduction

According to the World Health Organization, there are 466 million Deaf and hard-of-hearing people [8]. Among them, there are over 70 million who communicate via sign languages (SLs) resulting in more than 300 different SLs across

† Work done during an internship at Tencent AI Lab.
⋆ Corresponding author.
[3] `https://signavatars.github.io/`

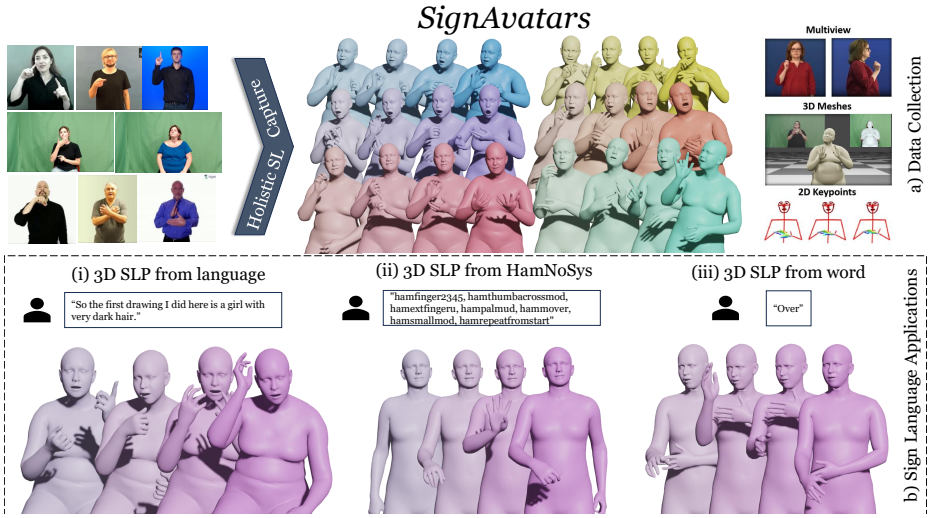

Fig. 1: Overview of SignAvatars, the first publicly available, large-scale multi-prompt 3D sign language holistic motion dataset. (**upper row**) We introduce a generic method to automatically annotate a large corpus of video data. (**lower row**) We propose a 3D SLP benchmark to produce plausible 3D holistic mesh motion and provide a neural architecture as well as baselines tailored for this novel task.

different communities [51]. While the field of (spoken) natural language processing (NLP) and language assisted computer vision (CV) are well explored, this is not the case for the alternate and important communicative tool of SL, and accurate generative models of holistic 3D avatars as well as dictionaries are highly desired for efficient learning [36].

We argue that the lack of large scale, targeted SL datasets is an important reason for this gap putting a barrier in front of downstream tasks such as digital simultaneous SL translators. On one hand, existing SL datasets and dictionaries [1,2,7,9,16,19] are typically limited to 2D videos or 2D keypoints annotations, which are insufficient for learners [28] as different signs could appear to be the same in 2D domain due to *depth ambiguity*. On the other hand, while parametric holistic models exist for human bodies [37] or bodies & faces [56], there is no unified, large-scale, multi-prompt 3D holistic motion dataset with accurate hand mesh annotations, which are crucial for SL. The reason for this is that the creation of 3D avatar annotation for SL is a labor-intensive, entirely manual process conducted by SL experts and the results are often unnatural [3].

To address this challenge, we begin by gathering various data sources from public datasets of continuous online videos with mixed-prompt annotations including HamNoSys, spoken language, and word and introduce the SignAvatars dataset. Overall, we compile $70K$ videos from 153 signers amounting to $8.34M$ frames. Unlike [14], our dataset is not limited to isolated signs, *i.e.* single sign per video, where HamNoSys-annotations are present, but includes continuous

and co-articulated signs. To augment our dataset with 3D full-body annotations, including 3D body, hand and face meshes as well as 2D & 3D keypoints, we design an automated and generic annotation pipeline, in which we perform a multi-objective optimization over 3D poses and shapes of face, hands and body. Our optimizer considers the temporal information of the motion and respects the biomechanical constraints in order to produce accurate hand poses, even in presence of complex, interacting hand gestures. Apart from meshes and SMPL-X [37] models, we also provide a *hand-only* subset with MANO [41] annotations.

SignAvatars enables multitude of tasks such as 3D sign language recognition (SLR) or the novel 3D sign language production (SLP) from text scripts, individual words, and HamNoSys notation. To address the latter challenge and accommodate diverse forms of semantic input, we further propose a novel SLP baseline, Sign-VQVAE, utilizing a semantic Variational Autoencoder (VQVAE) [53], capable of *parallel linguistic feature generation* (PLFG), effectively mapping various input data types to discrete code indices. The output of PLFG module is fused with a discrete motion encoder within an auto-regressive model to generate sequences of code indices derived from these semantic representations, strengthening the text-motion correlation. Consequently, our method can efficiently generate sign motion from an extensive array of textual inputs, enhancing its versatility and adaptability to various forms of semantic information. Sec. 5 will demonstrate that building such reliance and correlation between the low-level discrete representations leads to accurate, natural and sign-motion consistent SL production compared to direct regression from a high-level CLIP feature.

Besides leveraging the existing benchmarks, to quantitatively & qualitatively evaluate the potential of SignAvatars, we introduce a new SLP benchmark and present the first results for 3D SL holistic mesh motion production from multiple prompts including HamNoSys, spoken language, and word. On this benchmark, we assess the performance of our Sign-VQVAE against the other baselines we introduce, where we show a relative improvement of 200%. Though, none of the assessed models can truly match the desired accuracy, confirming the timeliness and the importance of SignAvatars. As depicted in Fig. 1, our contributions are:

- We introduce SignAvatars, the first large-scale multi-prompt 3D holistic motion SL dataset, containing diverse forms of semantic input.
- We provide accurate annotations for SignAvatars, in the form of expressive 3D avatar meshes. We do so by utilizing a multi-objective optimization capable of dealing with the complex interacting hands scenarios, while respecting the biomechanical hand constraints. We initialize this fitting procedure by a novel multi-stage, hierarchical process.
- We provide a new 3D sign language production (SLP) benchmark for SignAvatars, considering multiple prompts and full-body meshes.
- We further develop a VQVAE-based strong 3D SLP network significantly outperforming the baselines, which are also introduced as part of our work.

We believe SignAvatars is a significant stepping stone towards bringing the 3D digital world and 3D SL applications to the Deaf and hard-of-hearing communities, by fostering future research in 3D SL understanding.

2 Related Work

3D holistic mesh reconstruction. Recovering holistic 3D human body avatars from RGB videos and parsing them into parametric forms like SMPL-X [37] or Adam [22] is a well explored area [31, 37, 56]. Arctic [12] introduces a full-body dataset annotated by SMPL-X, for 3D object manipulation. [17] provide a hand-object constellations datasets with MANO annotations. However, such expressive parametric models (like TalkShow [56] or OsX [31]) have rarely been applied to the SL domain. [26] use off-the-shelf methods to estimate a holistic 3D mesh on existing dataset [52] but cannot deal with the challenging occlusions and interactions, making them unsuitable for complex, real scenarios. SignBERT+ [18] proposed the first self-supervised pre-trainable framework with model-aware hand prior for sign language understanding (SLU). The latest concurrent work [14] can reconstruct 3D holistic mesh for SL videos using linguistic priors with group labels obtained from a sign-classifier trained on Corpus-based Dictionary of Polish Sign Language (CDPSL) [32], which is annotated with HamNoSys As such, it utilizes an existing sentence segmentation methods [40] to generalize to multiple-sign videos. Overall, the literature lacks a robust yet generic method handling continuous and co-articulated SL videos with complex hand interactions.

SL datasets. While there have been many well-organized continuous 2D SL motion datasets [1, 2, 7, 9, 16, 19], the only existing 3D SL motion dataset with 3D holistic mesh annotation is in [14]. As mentioned, this rather small dataset only includes a single sign per video only with HamNoSys-prompts. In contrast, SignAvatars provides a multi-prompt 3D SL holistic motion dataset with continuous and co-articulated signs and fine-grained hand mesh annotations.

SL applications. [4] can generate 2D motion sequences from HamNoSys. [45] and [46] are able to generate 3D keypoint sequences relying on glosses. The avatar approaches are often hand-crafted and produce robotic and unnatural movements. Apart from them, there are also early avatar approaches [5, 10, 11, 15, 61] with a pre-defined protocol and character. To the best of our knowledge, we present the first large-scale 3D holistic SL motion dataset, SignAvatars. Built upon the dataset, we also introduce the novel task and benchmark of 3D sign language production, through different prompts (language, word, HamNoSys).

3 SignAvatars Dataset

Overview. SignAvatars is a holistic motion dataset composed of $70K$ video clips having $8.34M$ frames in total, containing body, hand and face motions as summarized in Tab. 2. We compile SignAvatars by gathering various data sources from public datasets to online videos and form seven subsets, whose distribution is reported in Fig. 2. Since the individual subsets do not naturally contain expressive 3D whole-body motion labels and 2D keypoints, we introduce a unified automatic annotation framework providing rich 3D holistic parametric SMPL-X annotations along with MANO subsets for hands. Overall, we provide

Table 1: Modalities of **publicly available** sign language datasets. C, I represent isolated and co-articulated (continuous) separately. * means the annotation has not been released yet. To the best of our knowledge, our dataset is the first publicly available 3D SL holistic continuous motion dataset with whole-body and hand mesh annotations with the most parallel modalities.

Data	Video	Frame	Duration (hrs.)	Co-articulated	Pose Annotation (to date)	Signer
RWTH-Phoenix-2014T [7]	8.25K	0.94M	11	C	-	9
DGS Corpus [16]	-	-	50	C	2D keypoints	327
BSL Corpus [47]	-	-	125	C	-	249
MS-ASL [23]	25K	-	25	I	-	222
WL-ASL [29]	21K	1.39M	14	I	2D keypoints	119
How2Sign [9]	34K	5.7M	79	C	2D keypoints, depth*	11
CSL-Daily [19]	21K	-	23	C	2D keypoints, depth	10
SIGNUM [54]	33K	-	55	C	-	25
AUTSL [48]	38K	-	21	I	depth	43
SGNify [14]	0.05K	4K	-	I	body mesh vertices	-
SignAvatars (Ours)	70K	8.34M	117	Both	SMPL-X, MANO, 2D&3D keypoints	153

117 hours of $70K$ video clips with $8.34M$ frames of motion data with accurate expressive holistic 3D mesh as motion annotations.

3.1 Dataset Characteristics

Expressive motion representation. To fill in the gaps of previous 2D-only SL data, our expressive 3D holistic body annotation consists of face, hands, and body, which is achieved by adopting SMPL-X [37]. It uses standard vertex-based linear blend skinning with learned corrective blend shapes and has $N = 10475$ vertices and $K = 67$ joints. For time interval $[1 : T]$, $V = (v_1, ..., v_T), J = (j_1, ..., j_T), \theta = (\theta_1, ..., \theta_T)$, represent mesh vertices, 3d joints, and poses in 6D representation [60], respectively. Here the pose $\theta_t := [\theta_t^b, \theta_t^h]$ includes the body pose $\theta_t^b \in R^{23\times6}$ with global orientation and the hand pose $\theta_t^h \in R^{30\times6}$. Moreover, $\theta_t^f \in R^6$ and $\phi = \{\phi_1, \ldots, \phi_T\}$ represents the yaw pose and facial expressions respectively. For each of the sequences, we use an optimized consistent shape parameter $\tilde{\beta}$ as there is no signer change within each clip. Overall, a motion state M_t is represented as: $M_t = (\theta_t^b, \theta_t^h, \theta_t^f, \phi_t, \tilde{\beta})$. Moreover, as shown in Tab. 1, our dataset also provides a hand motion subset by replacing the parametric representation from SMPL-X to MANO [41]: $M_t^h = (\theta_t^h, \tilde{\beta}^h)$, where h is the *handed-ness*.

Sign language notation. Similar to spoken languages, sign languages have special structures with a set of linguistic rules [6] (*e.g.* grammar, lexicons). Unlike spoken languages, they have no standard written forms. Moreover, there are over 300 different sign languages across the world, with Deaf and hard-of-hearing people who do not know any SL. Hence, having only a single type of annotation is insufficient in practice. To enable more generic applications targeting different users, our SL annotations include various modalities that can be categorized into four common types: HamNoSys, spoken language, word, and gloss, which can be used for a variety of downstream applications such as SLP and SLR.

Data sources. As shown in Tab. 2, SignAvatars leverages our unified automatic annotation framework to collect SL motion sequences in diverse modalities from various different sources. Specifically, for co-articulated SL datasets like How2Sign [9] and How2 [43] with American Sign Language (ASL) transcriptions, we collect *sentence-level* clips from the *Green Screen studio* subset with multi-view frames, resulting in $34K$ clips for the **ASL** subset. For German Sign Language **(GSL)** subset, we mostly gathered data from the publicly available PHOENIX14T dataset [7] following the official split to have $8.25K$ video clips. For **HamNoSys** subset, we collect $5.8K$ isolated-sign SL video clips from Polish SL corpus [33] for PJM, and German Sign Language (DGS), Greek Sign Language (GRSL) and French Sign Language (LSF) from DGS Corpus [38] and Dicta-Sign [34]. We finally gathered $21K$ clips from **word**-level sources such as WLASL [29] to curate the isolated-sign word subset. Overall, we gather our 7 subsets into four groups: (i) Word, (ii) ASL, (iii) HamNoSys (PJM, DGS, GRSL, LSF), (iv) GSL based on the prompt categories as shown in Fig. 2.

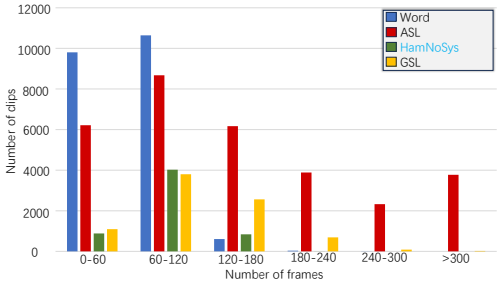

Fig. 2: Distribution of subsets. The number of frames for each clip in different subsets. PJM, LSF, DGS and GSL are gathered in one group.

Data	Video	Frame	Type	Signer
Word	$21K$	$1.39M$	W	119
PJM	$2.6K$	$0.21M$	H	2
DGS	$1.9K$	$0.12M$	H	8
GRSL	$0.8K$	$0.06M$	H	2
LSF	$0.4K$	$0.03M$	H	2
ASL	$34K$	$5.7M$	S	11
GSL	$8.3K$	$0.83M$	S, SG	9
Ours	$70K$	$8.34M$	S, H, W, SG	153

Table 2: Statistics of data sources. W, H, S, SG represent **w**ord, **H**amNoSys, sentence-level **s**poken language and sentence-level **g**loss.

3.2 Automatic Holistic Annotation

To efficiently auto-label the SL videos with motion data given only RGB online videos, we design an automatic 3D SL annotation pipeline that is not limited to isolated signs. To ensure motion stability and 3D shape accuracy, while maintaining efficiency during holistic 3D mesh recovery from SL videos, we propose an iterative fitting algorithm minimizing an objective heavily regularized both holistically and by *biomechanical hand constraints* [49]:

$$E(\theta, \tilde{\beta}, \phi) = \lambda_J L_J + \lambda_\theta L_\theta + \lambda_\alpha L_\alpha + \lambda_\beta L_\beta + \lambda_s L_{\text{smooth}} + \lambda_a L_{\text{angle}} + L_{\text{bio}} \quad (1)$$

where L_J represents the joint loss of 2D re-projection, which optimizes the difference between joints extracted from the SMPL-X model, projected into the image, with joints predicted with ViTPose [55] and MediaPipe [24]. To gain more robustness, we only optimize the joints that are visible in the image thus enforce a standing pose when the lower body is fully occluded. L_θ is the pose prior term following SMPLify-X [37]. Moreover, L_α is a prior penalizing extreme bending only for elbows and knees and L_β is the shape prior term. In addition, L_{smooth},

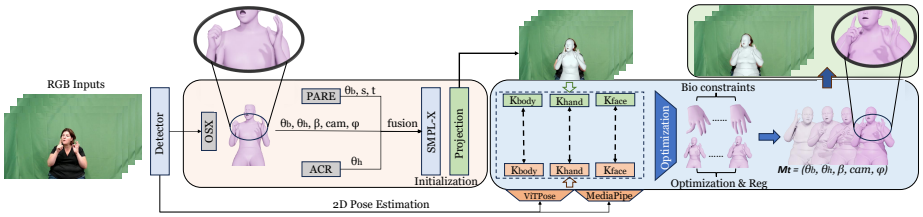

Fig. 3: Overview of our automatic annotation pipeline. Given an RGB image sequence as input, we perform a hierarchical initialization, followed by an optimization involving temporal smoothness and biomechanical constraints. Finally, our pipeline outputs the final 3D motion results as a sequence of SMPL-X parameters.

L_{angle} and L_{bio} are the smooth-regularization loss, angle loss and biomechanical constraints, separately. Finally, each λ denotes the influence weight of each loss term. In what follows, we describe in detail our regularizers.

Holistic regularization. To reduce the jitter on body and hand motion, caused by the noisy 2D detected keypoints, we employ a smoothness term defined as:

$$L_{\text{smooth}} = \|\hat{\theta}^b_{1:T}\|_2 + \|\hat{\theta}^h_{1:T}\|_2 + \|\theta^h_{2:T} - \theta^h_{1:T-1}\|_2 + \|\theta^b_{2:T} - \theta^b_{1:T-1}\|_2 \tag{2}$$

where $\hat{\theta}^b_{1:T} \in R^{N\times j_b\times 3}$ is the selected subset of pose parameters from $\theta^b_{1:T} \in R^{N\times J\times 3}$, and N is the frame number of the video. $\hat{\theta}^h \in R^{N\times j_h}$ is the selected subset of hand parameters from $\theta^b_{1:T} \in R^{N\times J\times 3}$. j_b and j_h are the numbers of selected body joints and hand parameters, Moreover, this could prevent implausible poses along the bone direction such as twists. Additionally, we penalize the hand poses lying outside the plausible range by adding an angle limit prior term:

$$L_{\text{angle}} = \mathcal{I}(\|\hat{\theta}^h_{1:T}\|_2; \theta^h_{\text{min}}, \theta^h_{\text{max}}) + \mathcal{I}(\|\hat{\theta}^b_{1:T}\|_2; \theta^b_{\text{min}}, \theta^b_{\text{max}}) \tag{3}$$

where $\mathcal{I}$ is the interval loss penalizing the outliers, $\theta^{h,b}_{\text{min}}, \theta^{h,b}_{\text{max}}$ is the pre-defined interval. Specifically, our fitting procedure is split into **five** stages, where we will optimize $\tilde{\beta}$ for the first **three** stages to derive the mean shape and freeze the shape in the following stages.

Biomechanical hand constraints. Hand pose estimation from monocular RGB images is challenging due to fast movements, interaction, frequent occlusion and confusion. To further improve the hand motion quality and eliminate implausible hand pose, we apply biomechanically constrain the hand poses, using three losses: (i) L_{bl} for bone length, (ii) L_{palm} for palmar region optimization, and (iii) L_{ja} for joint angle priors. Specifically, the final biomechanical loss L_{bio} is defined as the weighted sum $L_{\text{bio}} = \lambda_{bl}L_{\text{bl}} + \lambda_{palm}L_{\text{palm}} + \lambda_{ja}L_{\text{ja}}$, with:

$$\begin{aligned} L_{\text{bl}} &= \sum_i \mathcal{I}(\|b^i_{1:T}\|_2; b^i_{\text{min}}, b^i_{\text{max}}), \quad L_{\text{ja}} = \sum_i D(\alpha^i_{1:T}, H^i) \\ L_{\text{palm}} &= \sum_i (\mathcal{I}(\|c^i_{1:T}\|_2; c^i_{\text{min}}, c^i_{\text{max}}) + \mathcal{I}(\|d^i_{1:T}\|_2; d^i_{\text{min}}, d^i_{\text{max}})), \end{aligned} \tag{4}$$

where $\mathcal{I}$ is the interval loss penalizing $\|b^i_{1:T}\|_2$ if lie outside of the valid bone length range $[b^i_{min}, b^i_{max}]$, b_i is the bone length of i^{th} finger bone and the optimization constraints the whole sequence $[1:T]$. We further constrain the curvature $\|c^i_{1:T}\|_2$ and angular distance $\|d^i_{1:T}\|_2$ for the four root bones supporting the palmar structures by penalizing the outliers of curvature range $c^i_{\text{max}}, c^i_{\text{min}}$ and angular distance range $d^i_{\text{max}}, d^i_{\text{min}}$. Inspired by [49], we also apply constraints to the sequence of joint angles $\alpha^i_{1:T} = (\alpha^f_{1:T}, \alpha^a_{1:T})$ by approximating the convex hull on (α^f, α^a) plane with point set H^i and minimizing their distance D, where (α^f, α^a) is the flexion and abduction angles. The biomechanical loss is then computed as the weighted sum of them: $L_{\text{bio}} = \lambda_{bl} L_{\text{bl}} + \lambda_{palm} L_{\text{palm}} + \lambda_{ja} L_{\text{ja}}$. We refer the reader to our appendix for more details.

Hierarchical initialization. Given an RGB image sequence, we initialize the holistic SMPL-X parameters from OSX [31]. Though, due to the frequent occlusion and hand interactions, OSX is not always sufficient for a good initialization. Therefore, we further fuse OSX with ACR [57], PARE [25] to improve stability under occlusion and truncation. For 2D holistic keypoints initialization, we first train a whole-body 2D pose estimation model on COCO-WholeBody [21] based on ViTPose [55] and subsequently incorporate it with MediaPipe [24] by fusing and feeding through a confidence-guided filter.

4 SignVAE: A Strong 3D SLP Baseline

Our SignAvatars dataset enables the first applications to generate high-quality and natural 3D sign language holistic motion along with 3D meshes from both isolated and continuous SL prompts. To achieve this goal, motivated by the fact that the text prompts are highly correlated and aligned with the motion sequence, our method consists of a two-stage process designed to enhance the understanding of varied inputs by focusing on both semantic and motion aspects. In the first stage, we develop two codebooks - a shared semantic codebook and a motion codebook - by employing two Vector Quantized Variational Auto-Encoders (VQ-VAE). This allows us to map various kinds of input data to their corresponding semantic code indices and link motion elements to motion code indices. In the second stage, we utilize an auto-regressive model to generate motion code indices based on the previously determined semantic code indices. This integrated approach ensures a coherent and logical understanding of the input data, effectively capturing both the semantic and motion-related information.

SL motion generation. To produce stable and natural holistic poses in space and time, instead of directly mapping prompts to motion, we leverage the generative model VQ-VAE as our SL motion generator. As illustrated in Fig. 4, our SL motion VQVAE consists of an autoencoder structure and a learnable codebook Z_m, which contains I codes $Z_m = \{z_i\}_{i=1}^{I}$ with $z_i \in R^{d_z}$, where d_z is the dimension of the codes. We first encode the given 3D SL motion sequence $M_{1:T} = (\theta^b_{1:T}, \theta^h_{1:T}, \theta^f_{1:T}, \phi_{1:T})$ into a latent feature $F^m_{1:\tau} = (f^m_1, ..., f^m_{1:\tau}) \in R^{d_z}$, where $\tau = \frac{T}{w}$ and $w = 4$ is used as the downsampling rate for the window size. Subsequently, we quantize the latent feature embedding by searching for

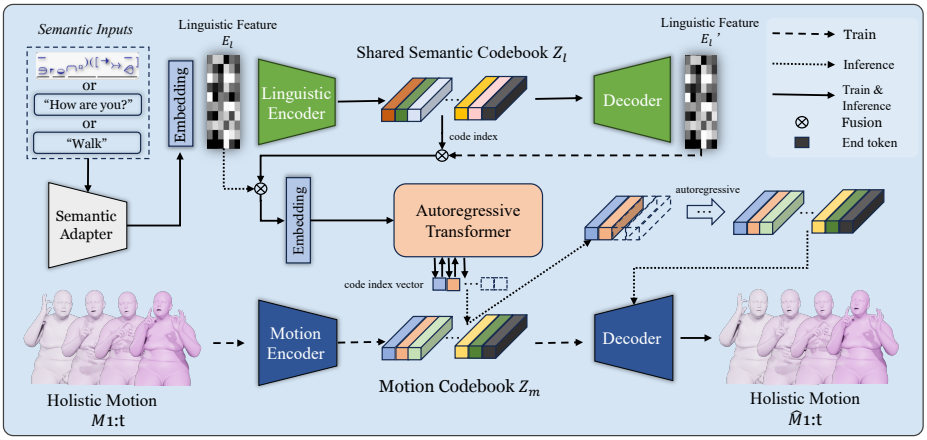

Fig. 4: Our 3D SLP network, SignVAE, consists of two-stages. We first create semantic and motion codebooks using two VQ-VAEs, mapping inputs to their respective code indices. Then, by an auto-regressive model, we generate motion code indices based on semantic code indices, ensuring a coherent understanding of the data.

the nearest neighbour code in the codebook Z_m. For the j^{th} feature, the quantization code $\hat{F}^m_{1:\tau}$ is found by: $\hat{f^m_j} = \underset{z_i \in Z_m}{\arg\min} \|f^m_j - z_i\|_2$. Finally, the quantized latent features are fed into decoders for reconstruction. For training of the SL motion generator, we apply the standard optimization scheme with L_{motion_vq}:

$$L_{m-vq} = L_{rec}(M_{1:T}, \hat{M}_{1:T}) + \|sg[F^m_{1:\tau}] - \hat{F}^m_{1:\tau}\|_2 + \omega\|F^m_{1:\tau} - \text{sg}[\hat{F}^m_{1:\tau}]\|_2 \quad (5)$$

where L_{rec} is the MSE loss, $\hat{M}_{1:T}$ is the reconstructed motion of $M_{1:T}$ and ω is a hyper-parameter. sg is the *detach* operation to stop the gradient. We provide more details regarding the network architecture and training in our appendix.

Prompt feature extraction for parallel linguistic feature generation. For efficient learning, typical motion generation tasks usually leverage an LLM to produce linguistic prior (condition) c given an input prompt. In our task of spoken language and word-level annotation, we leverage CLIP [39] as our prompt encoder to obtain the text embedding E^l. However, this does not extend to all the other SL annotations we desire. As a remedy, to enable applications with different prompts such as HamNoSys, instead of relying on the existing pre-trained CLIP, we define a new prompt encoder for embedding. After quantizing the prompt (*e.g.* HamNoSys glyph) into tokens with length s, we use an embedding layer to produce the linguistic feature $\hat{E}^l_{1:s} = (\hat{e}^l_1, ..., \hat{e}^l_s)$ with same dimension d_l as the text embeddings of CLIP [39]. For simplicity, we use "text" to represent all different input prompts. Subsequently, motivated by the fact that the text prompts are highly correlated and aligned with the motion sequence, we propose a linguistic VQVAE as our *parallel linguistic feature generator* (PLFG) module coupled with the SL motion generator. In particular, we leverage a similar quan-

Table 3: Quantitative comparisons on EHF dataset. *, † ‡ denote the optimization-based, regression-based method, and hybrid methods, respectively.

Method	MPVPE			PA-MPVPE			PA-MPJPE	
	Holistic	Hands	Face	Holistic	Hands	Face	Body	Hands
SMPLify-X [37]*	-	-	-	65.3	75.4	12.3	62.6	12.9
FrankMocap [42]†	107.6	42.8	-	57.5	12.6	-	62.3	12.9
PIXIE [13]†	89.2	42.8	32.7	55.0	11.1	4.6	61.5	11.6
Hand4Whole [35]†	76.8	39.8	26.1	50.3	10.8	5.8	60.4	10.8
PyMAF-X [13]†	64.9	29.7	19.7	50.2	10.2	5.5	52.8	10.3
OSX [31]†	70.8	-	-	48.7	-	-	55.6	-
Motion-X [30]‡	44.7	-	-	31.8	-	-	33.5	-
Motion-X w/GT 3Dkpt [30]‡	30.7	-	-	19.7	-	-	23.9	-
Ours (w/o bio)*	21.6	12.5	7.8	14.2	5.4	4.3	16.5	6.2
Ours*	**20.1**	**9.7**	**7.8**	**12.9**	**4.7**	**4.3**	**15.6**	**5.8**

tization process using the codebook Z_l and training scheme as in the SL motion generator to yield linguistic features:

$$L_{l-vq} = L_{rec}(E^l_{1:s}, \hat{E}^l_{1:s}) + \|sg[F^l_{1:s}] - \hat{F^l_{1:s}}\|_2 + \omega\|F^l_{1:s} - sg[\hat{F^l_{1:s}}]\|_2 \quad (6)$$

where $F^l_{1:s}$ is the latent feature after encoding the initial linguistic feature. $\hat{F^l_{1:s}}$ is the quantized linguistic feature after applying $\hat{f^l_j} = \underset{z_i \in Z_l}{\arg\max} \|f^l_j - z_i\|_2$ to $F^l_{1:s}$.

Sign-motion cross modelling and production. After training the VQVAE-based SL motion generator, we can map any motion sequence $M_{1:T}$ to a sequence of indices $X = [x_1, ..., x_{T/w}, x_{\text{EOS}}]$ through the motion encoder and quantization, where x_{EOS} is a learnable end token representing the *stop* signal. After training both the SL motion generator and the linguistic feature generator, our network will be jointly optimized in a parallel manner. Specifically, we fuse the linguistic feature embedding E_l and the codebook index vectors of Z_l to formulate the final condition for our autoregressive code index generator. The objective for training the code index generator can be seen as an autoregressive next-index prediction task, learned with a cross − entropy loss between the likelihood of the full predicted code index sequence and the real ones as $L_{\text{SLP}} = \mathbb{E}_{X \sim p(X)}[-\log p(X|c)]$ given an input linguistic prompt c.

Lastly, with the quantized motion representation, we generate the codebook vectors in a temporal autoregressive manner and predict the distribution of the next codebook indices. After mapping the codebook indices X to the quantized motion representation $\hat{F}^m_{1:\tau}$, we can decode and produce the final 3D holistic motion with mesh representations $M_{1:T}$.

5 Experimental Evaluation

We now showcase the effectiveness of our contributions individually, namely, the 3D reconstruction and annotation pipeline as well as the 3D sign language production. Note that, with SignAvatars we present the first benchmark results for 3D holistic SL motion production yielding mesh representations.

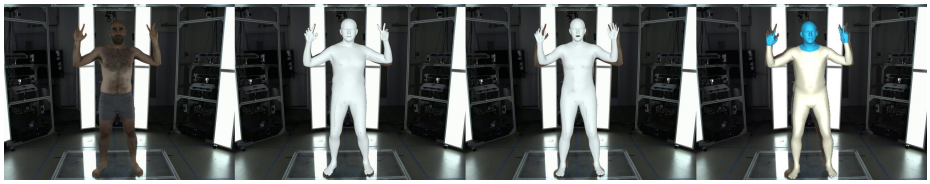

Fig. 6: Comparing **3D holistic human mesh reconstruction** methods on EHF dataset [37]. Our annotation method produces significantly better holistic reconstructions with plausible poses and the best pixel alignment. (Zoom in for a better view)

5.1 Evaluating the Annotation Pipeline

We start by assessing our optimization-based automatic annotation approach. Due to the availability of ground truth, we evaluate our method against the state-of-the-art hand and holistic human mesh recovery methods including SMPLify-x [37], OSX [31], PyMAF-X [58], PIXIE [13], on the standard EHF dataset [37].

Evaluation metrics. Our quantitative evaluations follow the prior works and compute per-vertex error (MP-VPE), mean per-vertex error after Procrusters alignment (PA-MPVPE), and mean per-joint error after PA (PA-MPJPE).

Results. It can be seen from Tab. 3 that our method significantly surpasses the leading monocular holistic reconstruction methods by a large margin. Notably, our PA-MPJPE we achieve a 40% improvement over SoTA [30]. Specifically, our hand reconstruction error drops down to 4.7 on PA-MPVPE when the biomechanical constraints are integrated. The qualitative results presented in Fig. 5 and in Fig. 6 show significantly more natural body movement, accurate hand poses and better pixel-mesh aligned body shapes (β) in favour of our method. This superior reconstruction quality consistently translate to the quality of our dataset.

Fig. 5: Comparison of 3D holistic body reconstruction. The results from PIXIE [13], PyMAF-X [58], and Ours on our dataset, SignAvatars.

5.2 3D SL Motion Generation on the SignAvatars Benchmark

We now evaluate the generative capabilities of our SignVAE model on the SignAvatars dataset and provide ablations studies on the PFLG module.

Evaluation metrics. To fully assess the quality of our motion generation, we evaluate the holistic motion as well as the arm motion. Our network is only

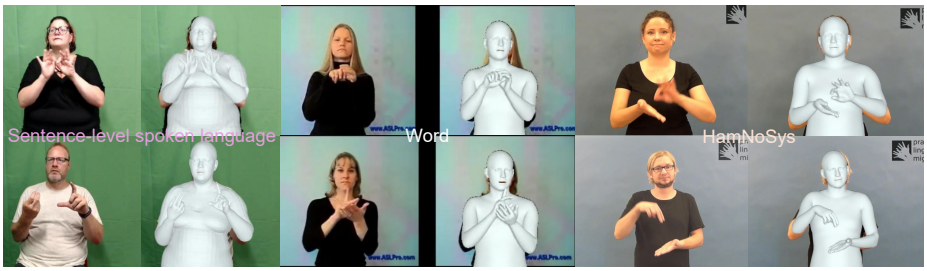

Fig. 7: Output of our reconstruction-based annotation pipeline for different types of input. Specifically, we present examples from the subsets of SignAvatars (**left**) sentence-level spoken language from ASL subset, (**mid**) HamNoSys-level examples from *HamNoSys*-subset, and (**right**) word-level examples of the *word*-subset.

trained on the **upper body** [4]. Based on an evaluation model trained following prior arts in motion generation [50, 59], we use the scores and metrics of FID, Diversity, Multimodality (MM), MM-Dist, MR-precision, whose details are provided in our supplementary material. Unfortunately, there is no de-facto standard for evaluating 3D SLP in the literature. While [28] can back-translate 3D SL motion, it is tailored only for word-level back-translation. While BLEU and ROUGE are commonly used in the back-translation evaluation [45, 46], they are not generic for other types of annotations such as HamNoSys or glosses. Since the generated motion might differ in length from the real motion, absolute metrics like MPJPE would also be unsuited. Inspired by [4, 20], we propose a new **MR-Precision** for motion retrieval as well as DTW-MJE (Dynamic Time Warping - Mean Joint Error) [27] with standard SMPL-X keypoint set without lower body, for evaluating the performance of our method as well as the baselines.

Subsets & training settings. Specifically, we report results on three representative subsets: (i) the *ASL* subset for spoken language (corresponding to *language* in Tab. 6), (ii) the *word* subset with 300 vocabularies, (iii) combined subset HamNoSys. For training, we follow the official splits for (i) and (ii). For (iii), we leverage a four-fold strategy where we train on three of them and test on the other, repeated four times to have the final results.

Benchmarking & results. To the best of our knowledge, there is no publicly available benchmark for 3D mesh & motion-based SLP[5]. To evaluate SignAvatars as the first 3D motion-based SLP benchmark, we present detailed quantitative results in Tab. 6. It can be seen that the 3D SLP with word-level prompts can achieve the best performance reaching the quality of real motions. Learning from spoken languages is a naturally harder task and we invite the community to develop stronger methods to produce 3D SLP from spoken languages. To further evaluate the sign accuracy and effect of body movement, we report separate results for individual arms (*e.g.* "*Gesture*"), with slight improvements in FID

[4] The lower body is not factored in our evaluations as it is unrelated to the SL motion.
[5] [44] does not provide a public evaluation model as discussed in our Appendix.

Table 6: Quantitative evaluation results for the 3D holistic SL motion generation. *Real motion* denotes the motions sampled from the original holistic motion annotation in the dataset. *Holistic* represents the generation results regarding holistic motion. *Gesture* stands for the evaluation conducted on two arms. *Div.* refers to Diversity.

Data Type		R-Precision (↑)			FID (↓)	Div. (→)	MM (→)	MM-dist(↓)	MR-Precision (↑)		
		top 1	top 3	top 5					top 1	top 3	top 5
Real motion	Language	$0.375^{\pm.005}$	$0.545^{\pm.007}$	$0.679^{\pm.008}$	$0.061^{\pm.153}$	$12.11^{\pm.075}$	-	$3.786^{\pm.057}$	-	-	-
	HamNoSys	$0.455^{\pm.002}$	$0.689^{\pm.006}$	$0.795^{\pm.004}$	$0.007^{\pm.062}$	$8.754^{\pm.028}$	-	$2.113^{\pm.023}$	-	-	-
	Word-300	$0.499^{\pm.003}$	$0.811^{\pm.002}$	$0.865^{\pm.003}$	$0.006^{\pm.054}$	$8.656^{\pm.035}$	-	$1.855^{\pm.019}$	-	-	-
Holistic	Language	$0.265^{\pm.007}$	$0.413^{\pm.008}$	$0.531^{\pm.0059}$	$4.359^{\pm.389}$	$12.35^{\pm.101}$	$3.451^{\pm.107}$	$4.851^{\pm.067}$	$0.356^{\pm.007}$	$0.525^{\pm.007}$	$0.645^{\pm.009}$
	HamNoSys	$0.429^{\pm.004}$	$0.657^{\pm.005}$	$0.756^{\pm.002}$	$0.884^{\pm.035}$	$9.451^{\pm.087}$	$0.941^{\pm.056}$	$2.651^{\pm.027}$	$0.552^{\pm.002}$	$0.745^{\pm.010}$	$0.813^{\pm.034}$
	Word-300	$0.475^{\pm.002}$	$0.731^{\pm.003}$	$0.815^{\pm.005}$	$0.756^{\pm.021}$	$8.956^{\pm.091}$	$0.815^{\pm.059}$	$2.101^{\pm.024}$	$0.615^{\pm.005}$	$0.797^{\pm.006}$	$0.875^{\pm.002}$
Gesture	Language	$0.245^{\pm.008}$	$0.405^{\pm.009}$	$0.519^{\pm.010}$	$3.951^{\pm.315}$	$10.12^{\pm.121}$	$3.112^{\pm.135}$	$5.015^{\pm.089}$	$0.375^{\pm.011}$	$0.535^{\pm003}$	$0.668^{\pm.004}$
	HamNoSys	$0.435^{\pm.005}$	$0.649^{\pm.004}$	$0.745^{\pm.006}$	$0.851^{\pm.033}$	$8.944^{\pm.097}$	$0.913^{\pm.036}$	$2.876^{\pm.015}$	$0.581^{\pm.004}$	$0.736^{\pm.006}$	$0.825^{\pm.008}$
	Word-300	$0.465^{\pm.001}$	$0.711^{\pm.003}$	$0.818^{\pm.003}$	$0.715^{\pm.016}$	$8.235^{\pm.055}$	$0.801^{\pm.021}$	$2.339^{\pm.027}$	$0.593^{\pm.006}$	$0.814^{\pm.005}$	$0.901^{\pm.006}$

and MR-Precision. However, it will also degenerate the text-motion consistency (R-Precision and MM-dist) due to the absence of body-relative hand position.

Due to the lack of works that can generate 3D holistic SL motion with mesh representation from any of the linguistic sources (*e.g.* spoken language, HamNoSys, gloss, ...), we modify the latest HamNoSys-based SLP work, Ham2Pose [4] (*Ham2Pose-3d* in Tab. 4), as well as MDM [50] (corresponding to *SignDiffuse* in Tab. 5), to take our linguistic feature as input and to output SMPL-X representations and evaluate on our dataset. We then train our SignVAE and *Ham2Pose-3d* along with the original *Ham2Pose* on their official split and use DTW-MJE for evaluation. Specifically, we also regress the keypoints from our holistic representation $M_{1:T}$ to align with the Ham2Pose 2D skeleton. As discovered in this benchmark, leveraging our SignAvatars dataset can easily enable more 3D approaches and significantly improve the existing SLP applications by simple adaptation compared to the original Ham2Pose. The results in Tab. 4 are reported on the HamNoSys *holistic* set for comparison. While our method drastically improves over the baseline, the result is far from ideal, motivating the need for better models for this new task.

Table 4: Comparison with state-of-the-art SLP methods from HamNoSys holistic subset. * represents using only 2D cues.

Method	DTW-MJE Rank (↑)		
	top 1	top 3	top 5
Ham2Pose*	$0.092^{\pm.031}$	$0.197^{\pm.029}$	$0.354^{\pm.032}$
Ham2Pose-3d	$0.253^{\pm.036}$	$0.369^{\pm.039}$	$0.511^{\pm.035}$
SignVAE (Ours)	$\mathbf{0.516}^{\pm.039}$	$\mathbf{0.694}^{\pm.041}$	$\mathbf{0.786}^{\pm.035}$

Table 5: Quantitative ablation study of SignVAE on HamNoSys *holistic* subset for comparison with prior arts.

Method	R-Precision (↑)			MM-dist(↓)
	top 1	top 3	top 5	
Ham2Pose-3d	$0.291^{\pm.006}$	$0.386^{\pm.005}$	$0.535^{\pm.005}$	$3.875^{\pm.086}$
SignDiffuse	$0.285^{\pm.003}$	$0.415^{\pm.005}$	$0.654^{\pm.003}$	$3.866^{\pm.054}$
SignVAE(Base)	$0.385^{\pm.008}$	$0.613^{\pm.009}$	$0.745^{\pm.007}$	$3.056^{\pm.108}$
SignVAE(Ours)	$\mathbf{0.429}^{\pm.009}$	$\mathbf{0.657}^{\pm.008}$	$\mathbf{0.756}^{\pm.008}$	$\mathbf{2.651}^{\pm.119}$

Fig. 8 shows qualitative results from continuous 3D holistic body motion generation. As observed, our method can generate plausible and accurate holistic 3D motion from a variety of prompts while containing some diversity enriching the production results. We provide further examples on our supplementary material.

Ablation on PFLG. In order to study our unique text-sign cross-modeling module, we introduce a baseline, *SignVAE (Base)*, replacing the PLFG with a canonical CLIP-like encoder as input to the encoder as our Semantic Adap-

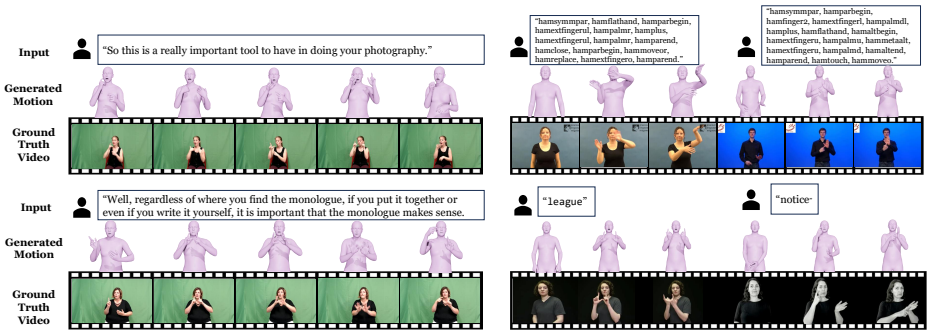

Fig. 8: Qualitative results of 3D holistic SLP from different prompts (left row: spoken language, top right: HamNoSys, bottom right: word). Within each sample, the first two rows are the input prompts and the generated results.

tor and E_l is directly fed into the embedding layer. As shown in Tab. 5, our joint scheme utilizing the PLFG can significantly improve the prompt-motion consistency, resulting in an increase in R-precision and MM-dist. Moreover, our VQVAE backbone quantizing the motion representation into a motion codebook, enables interaction with the linguistic feature codebook, leading to significant improvements in prompt-motion correspondences and outperforms other baselines built with our linguistic feature generator (SignDiffuse, Ham2Pose-3d) and generates more text-motion consistent results.

6 Conclusion

We introduced **SignAvatars**, the first large-scale 3D holistic SL motion dataset with expressive 3D human and hand mesh annotations, provided by our automatic annotation pipeline. SignAvatars enables a variety of application potentials for Deaf and hard-of-hearing communities. Built upon our dataset, we proposed the first 3D sign language production approach to generate natural holistic mesh motion sequences from SL prompts. We also introduced the first benchmark results for this new task, continuous and co-articulated 3D holistic SL motion production from diverse SL prompts. Our evaluations on this benchmark clearly showed the advantage of our SignVAE, over the baselines, we develop.

Limitations and future work. Having the first benchmark at hand opens up a sea of potential in-depth investigations for 3D SL motion generation. Especially, the lack of a sophisticated and generic 3D back-translation method may prevent our evaluations from fully showcasing the superiority of the proposed method. We leave this for a future study. Combining 3D SLT and SLP to formulate a multimodal generic SL framework will also be one of the future works. Developing a large 3D sign language motion model with more properties and applications in AR/VR will significantly benefit the Deaf and hard-of-hearing people around the world, as well as countless hearing individuals interacting with them. As such, We invite the research community to develop even stronger baselines.

Acknowledgements. T. Birdal acknowledges support from the Engineering and Physical Sciences Research Council [grant EP/X011364/1].

References

1. Albanie, S., Varol, G., Momeni, L., Afouras, T., Chung, J.S., Fox, N., Zisserman, A.: BSL-1K: Scaling up co-articulated sign language recognition using mouthing cues. In: European Conference on Computer Vision (2020)
2. 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. `https://www.robots.ox.ac.uk/~vgg/data/bobsl` (2021)
3. Aliwy, A.H., Ahmed, A.A.: Development of arabic sign language dictionary using 3d avatar technologies. Indonesian Journal of Electrical Engineering and Computer Science **21**(1), 609–616 (2021)
4. 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. pp. 21046–21056 (2023)
5. 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)
6. Blaisel, X.: David f. armstrong, william c. stokoe, sherman e. wilcox, gesture and the nature of language, cambridge, cambridge university press, 1995, x+ 260 p., bibliogr., index. Anthropologie et Sociétés **21**(1), 135–137 (1997)
7. Camgoz, N.C., Hadfield, S., Koller, O., Ney, H., Bowden, R.: Neural sign language translation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7784–7793 (2018)
8. Davis, A.C., Hoffman, H.J.: Hearing loss: rising prevalence and impact. Bulletin of the World Health Organization **97**(10), 646 (2019)
9. 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. pp. 2735–2744 (2021)
10. Ebling, S., Glauert, J.: Building a swiss german sign language avatar with jasigning and evaluating it among the deaf community. Universal Access in the Information Society **15**, 577–587 (2016)
11. Efthimiou, E., Fotinea, S.E., Hanke, T., Glauert, J., Bowden, R., Braffort, A., Collet, C., Maragos, P., Goudenove, F.: Dicta-sign: sign language recognition, generation and modelling with application in deaf communication. In: sign-lang@ LREC 2010. pp. 80–83. European Language Resources Association (ELRA) (2010)
12. Fan, Z., Taheri, O., Tzionas, D., Kocabas, M., Kaufmann, M., Black, M.J., Hilliges, O.: ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In: IEEE Conference on Computer Vision and Pattern Recognition (2023)
13. Feng, Y., Choutas, V., Bolkart, T., Tzionas, D., Black, M.J.: Collaborative regression of expressive bodies using moderation. In: International Conference on 3D Vision (3DV) (2021)
14. Forte, M.P., Kulits, P., Huang, C.H.P., Choutas, V., Tzionas, D., Kuchenbecker, K.J., Black, M.J.: Reconstructing signing avatars from video using linguistic priors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12791–12801 (2023)

15. Gibet, S., Lefebvre-Albaret, F., Hamon, L., Brun, R., Turki, A.: Interactive editing in french sign language dedicated to virtual signers: Requirements and challenges. Universal Access in the Information Society **15**, 525–539 (2016)
16. Hanke, T., Schulder, M., Konrad, R., Jahn, E.: Extending the public dgs corpus in size and depth. In: sign-lang@ LREC 2020. pp. 75–82. European Language Resources Association (ELRA) (2020)
17. Hasson, Y., Varol, G., Tzionas, D., Kalevatykh, I., Black, M.J., Laptev, I., Schmid, C.: Learning joint reconstruction of hands and manipulated objects. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11807–11816 (2019)
18. Hu, H., Zhao, W., Zhou, W., Li, H.: Signbert+: Hand-model-aware self-supervised pre-training for sign language understanding. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)
19. Huang, J., Zhou, W., Zhang, Q., Li, H., Li, W.: Video-based sign language recognition without temporal segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 32 (2018)
20. Huang, W., Pan, W., Zhao, Z., Tian, Q.: Towards fast and high-quality sign language production. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 3172–3181 (2021)
21. Jin, S., Xu, L., Xu, J., Wang, C., Liu, W., Qian, C., Ouyang, W., Luo, P.: Whole-body human pose estimation in the wild. In: Proceedings of the European Conference on Computer Vision (ECCV) (2020)
22. Joo, H., Simon, T., Sheikh, Y.: Total capture: A 3d deformation model for tracking faces, hands, and bodies. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8320–8329 (2018)
23. Joze, H.R.V., Koller, O.: Ms-asl: A large-scale data set and benchmark for understanding american sign language. arXiv preprint arXiv:1812.01053 (2018)
24. Kartynnik, Y., Ablavatski, A., Grishchenko, I., Grundmann, M.: Real-time facial surface geometry from monocular video on mobile gpus. arXiv preprint arXiv:1907.06724 (2019)
25. Kocabas, M., Huang, C.H.P., Hilliges, O., Black, M.J.: Pare: Part attention regressor for 3d human body estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11127–11137 (2021)
26. Kratimenos, A., Pavlakos, G., Maragos, P.: Independent sign language recognition with 3d body, hands, and face reconstruction. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 4270–4274. IEEE (2021)
27. Kruskal, J.B.: An overview of sequence comparison: Time warps, string edits, and macromolecules. SIAM review **25**(2), 201–237 (1983)
28. Lee, T., Oh, Y., Lee, K.M.: Human part-wise 3d motion context learning for sign language recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 20740–20750 (2023)
29. Li, D., Rodriguez, C., Yu, X., Li, H.: Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 1459–1469 (2020)
30. Lin, J., Zeng, A., Lu, S., Cai, Y., Zhang, R., Wang, H., Zhang, L.: Motionx: A large-scale 3d expressive whole-body human motion dataset. In: Advances in Neural Information Processing Systems (2023)

31. Lin, J., Zeng, A., Wang, H., Zhang, L., Li, Y.: One-stage 3d whole-body mesh recovery with component aware transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023)
32. Linde-Usiekniewicz, J., Czajkowska-Kisil, M., Łacheta, J., Rutkowski, P.: A corpus-based dictionary of polish sign language (pjm)
33. Linde-Usiekniewicz, J., Czajkowska-Kisil, M., Łacheta, J., Rutkowski, P.: A corpus-based dictionary of polish sign language (pjm). In: Proceedings of the XVI EURALEX International Congress: The user in focus. pp. 365–376 (2014)
34. Matthes, S., Hanke, T., Regen, A., Storz, J., Worseck, S., Efthimiou, E., Dimou, A.L., Braffort, A., Glauert, J., Safar, E.: Dicta-sign-building a multilingual sign language corpus. In: 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon. Satellite Workshop to the eighth International Conference on Language Resources and Evaluation (2012)
35. Moon, G., Choi, H., Lee, K.M.: Accurate 3d hand pose estimation for whole-body 3d human mesh estimation. In: Computer Vision and Pattern Recognition Workshop (CVPRW) (2022)
36. Naert, L., Larboulette, C., Gibet, S.: A survey on the animation of signing avatars: From sign representation to utterance synthesis. Computers & Graphics **92**, 76–98 (2020)
37. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3d hands, face, and body from a single image. In: Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2019)
38. Prillwitz, S., Hanke, T., König, S., Konrad, R., Langer, G., Schwarz, A.: Dgs corpus project–development of a corpus based electronic dictionary german sign language/german. In: sign-lang@ LREC 2008. pp. 159–164. European Language Resources Association (ELRA) (2008)
39. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
40. 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. pp. 3403–3412 (2021)
41. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: Modeling and capturing hands and bodies together. arXiv preprint arXiv:2201.02610 (2022)
42. Rong, Y., Shiratori, T., Joo, H.: Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration. In: IEEE International Conference on Computer Vision Workshops (2021)
43. Sanabria, R., Caglayan, O., Palaskar, S., Elliott, D., Barrault, L., Specia, L., Metze, F.: How2: a large-scale dataset for multimodal language understanding. arXiv preprint arXiv:1811.00347 (2018)
44. Saunders, B., Camgoz, N.C., Bowden, R.: Adversarial training for multi-channel sign language production. arXiv preprint arXiv:2008.12405 (2020)
45. Saunders, B., Camgoz, N.C., Bowden, R.: Progressive transformers for end-to-end sign language production. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16. pp. 687–705. Springer (2020)
46. 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. pp. 1919–1929 (2021)

47. Schembri, A., Fenlon, J., Rentelis, R., Reynolds, S., Cormier, K.: Building the british sign language corpus. LaNguagE DocumENtatIoN & coNSErvatIoN (2013)
48. Sincan, O.M., Keles, H.Y.: Autsl: A large scale multi-modal turkish sign language dataset and baseline methods. IEEE Access **8**, 181340–181355 (2020)
49. Spurr, A., Iqbal, U., Molchanov, P., Hilliges, O., Kautz, J.: Weakly supervised 3d hand pose estimation via biomechanical constraints. In: European conference on computer vision. pp. 211–228. Springer (2020)
50. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-Or, D., Bermano, A.H.: Human motion diffusion model. arXiv preprint arXiv:2209.14916 (2022)
51. The World Federation, o.t.D.: Who we are, `http://wfdeaf.org/who-we-are/`
52. Theodorakis, S., Pitsikalis, V., Maragos, P.: Dynamic–static unsupervised sequentiality, statistical subunits and lexicon for sign language recognition. Image and Vision Computing **32**(8), 533–549 (2014)
53. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. Advances in neural information processing systems **30** (2017)
54. Von Agris, U., Knorr, M., Kraiss, K.F.: The significance of facial features for automatic sign language recognition. In: 2008 8th IEEE international conference on automatic face & gesture recognition. pp. 1–6. IEEE (2008)
55. Xu, Y., Zhang, J., Zhang, Q., Tao, D.: Vitpose: Simple vision transformer baselines for human pose estimation. Advances in Neural Information Processing Systems **35**, 38571–38584 (2022)
56. Yi, H., Liang, H., Liu, Y., Cao, Q., Wen, Y., Bolkart, T., Tao, D., Black, M.J.: Generating holistic 3d human motion from speech. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023)
57. Yu, Z., Huang, S., Chen, F., Breckon, T.P., Wang, J.: Acr: Attention collaboration-based regressor for arbitrary two-hand reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
58. Zhang, H., Tian, Y., Zhang, Y., Li, M., An, L., Sun, Z., Liu, Y.: Pymaf-x: Towards well-aligned full-body model regression from monocular images. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)
59. Zhang, J., Zhang, Y., Cun, X., Huang, S., Zhang, Y., Zhao, H., Lu, H., Shen, X.: 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)
60. Zhu, H., Zuo, X., Wang, S., Cao, X., Yang, R.: Detailed human shape estimation from a single image by hierarchical mesh deformation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4491–4500 (2019)
61. 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) (2004)