# Neural Sign Actors: A diffusion model for 3D sign language production from text

Vasileios Baltatzis[1], Rolandos Alexandros Potamias[1], Evangelos Ververas[1],
Guanxiong Sun[2], Jiankang Deng[1], Stefanos Zafeiriou[1]

[1]Imperial College London, [2]Queen's University Belfast

{vasileios.baltatzis18, r.potamias, e.ververas16, j.deng16, s.zafeiriou}@imperial.ac.uk,
gsun02@qub.ac.uk

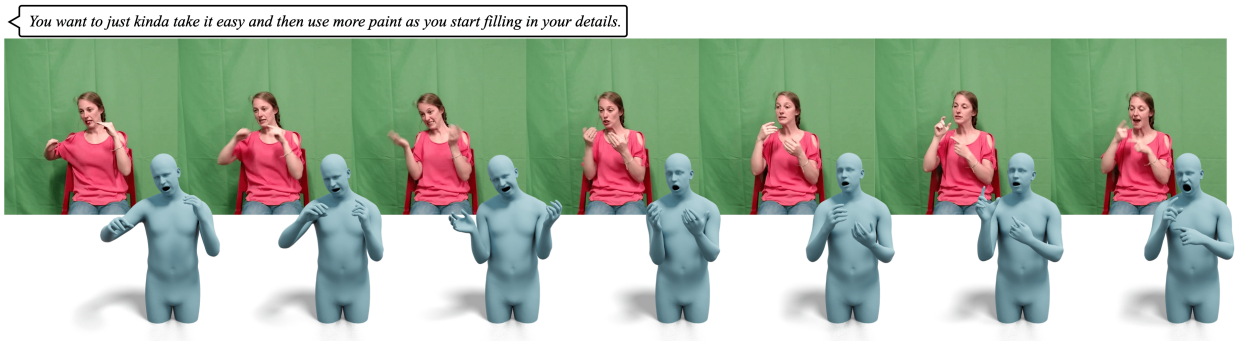


Figure 1. The proposed method takes raw text as input and generates a realistic and coherent motion of its corresponding sign language translation. From top to bottom: the input text, the ground truth sign language video (shown just for reference), and the generated motion.

## Abstract

*Sign Languages (SL) serve as the primary mode of communication for the Deaf and Hard of Hearing communities. Deep learning methods for SL recognition and translation have achieved promising results. However, Sign Language Production (SLP) poses a challenge as the generated motions must be realistic and have precise semantic meaning. Most SLP methods rely on 2D data, which hinders their realism. In this work, a diffusion-based SLP model is trained on a curated large-scale dataset of 4D signing avatars and their corresponding text transcripts. The proposed method can generate dynamic sequences of 3D avatars from an unconstrained domain of discourse using a diffusion process formed on a novel and anatomically informed graph neural network defined on the SMPL-X body skeleton. Through quantitative and qualitative experiments, we show that the proposed method considerably outperforms previous methods of SLP. This work makes an important step towards realistic neural sign avatars, bridging the communication gap between Deaf and hearing communities.*[1]

[1]Project page: https://baltatzisv.github.io/neural-sign-actors/

## 1. Introduction

Sign language (SL) is a form of language in which visual-manual modalities are used instead of spoken words to convey meaning. It is the predominant form of communication for more than 70 million Deaf and Hard of Hearing people around the world. Akin to verbal languages, SLs have extremely rich vocabulary and grammar, yet the complexities differ drastically [55]. To enable effective visual communication, they consist of both manual and non-manual components [35]. The manual modality encompasses hand articulation, orientation, position, and motion, while non-manual elements include arm movements and facial expressions [6]. Whilst it is possible to convey some meaning using just hand articulations, expressiveness is limited since non-manual elements often convey emotions [3, 55].

Recently, several methods have been proposed to bridge the domain gap between sign and spoken languages. Most methods focus on Sign Language Recognition (SLR) which includes the translation of a specific sign to its corresponding meaning, as well as Sign Language Translation (SLT) that extends SLR to the translation of a sign sequence to its spoken word equivalent. This is usually tackled using glosses [10, 14, 15, 28], which are simplified mid-level

representations that relate each sign with a corresponding meaning. However, even though glosses have provided a substantial enhancement to SLT methods, they have a predefined informative bottleneck, which limits the translation accuracy, and they usually fail to provide long-range dependencies and contextual information [10, 32, 64].

Despite the significant number of individuals with hearing difficulties, only ∼5% of television programs are interpreted into sign language, which shows the vital need for 3D signing avatars. Compared to SLR and SLT, only a small number of methods have attempted to tackle the task of Sign Language Production (SLP), either by using directly stand-alone glosses [53, 62] or by training a network to map text to glosses [50]. In the SLP setting, a network is given a text sentence and attempts to generate a motion that reflects the corresponding sign language translation. Usually, this is done using 2D and 3D joints to represent the human body [32, 49, 53]. However, joints provide an unrealistic representation of the animation, limiting their practical use in real-world avatar actor applications. Recently, Stoll *et al*. [54] proposed to extend SLP to 3D meshes using an optimization step that fits a SMPL-X [41] model to the predicted 2D joints. In contrast, the proposed method directly regresses the poses of SMPL-X [41] model to generate an animatable 3D signing avatar.

Aside from its challenging nature, SLP remains relatively unexplored due to the absence of large-scale available datasets with a sufficient vocabulary size. In particular, the majority of current SLP methods rely on German sign language datasets composed of only a few thousand words [9, 24] and use 2D landmarks detected from off-the-shelf pose estimation methods [11]. In contrast, we employ a hybrid regression-optimization method to accurately annotate, with SMPL-X pose and shape parameters [41], a large-scale video dataset [19] composed of over 16k word tokens. Using the acquired 3D pose annotations, we train a dynamic diffusion model to learn SLP from English texts. Our method directly translates text to signs without using any intermediate representation [49], which increases the generative capacity of the network. Given that sign language cannot be translated word-for-word [35], we utilize an off-the-shelf sentence encoder [43] which also enables out-of-distribution generalization. To sum up, the contributions of this study can be summarized as:

- We introduce the task of direct 3D signing *avatar* generation from text, without relying on 2D fitting optimizations or any intermediate gloss representations. In this paper, we aim to make a step towards neural sign avatars to aid the Deaf and Hard of Hearing community [40].
- We derive the first large-scale 3D dataset of American Sign Language by designing a state-of-the-art pipeline to annotate the How2Sign dataset [19].
- We propose a text-conditioned dynamic diffusion model founded on a novel, anatomically inspired graph neural network that facilitates SLP. The proposed model achieves remarkable results that outperform the current state-of-the-art models, by a large margin.

## 2. Related Work

### 2.1. Sign Language Production

Despite nearly two decades of research [16, 39], the development of highly effective sign language production methods remains challenging. Stoll *et al*. [53] proposed the first neural SLP method forming a seq2seq architecture to map text to glosses. To decode poses to 2D joint locations they proposed an empirical lookup table paradigm. To avoid the two-stage generation, Zelinka *et al*. [62] utilized OpenPose [11] to extract joint locations from Czech weather forecasting videos and train a network to directly regress the 2D joint poses. Recently, several methods [30, 49] proposed transformer-based architectures to tackle German sign language production [10]. However, their generations suffer from under-articulation and limited expressiveness in hand and body motion. Follow-up works attempted to improve the generation quality using adversarial training [48], mixture density networks [50] and dictionary representations [51]. However, most of the aforementioned methods, apart from being contingent to intermediate glosses representation, rely on the regression of 2D/3D joint positions, a process that inherently encounters difficulties in realistically conveying meanings. In an attempt to tackle such limitations, Stoll *et al*. [54] proposed the application of a post-regression SMPL-X [41] fitting to lift 2D joints to 3D meshes. On the contrary, we make a step towards realistic signing avatars and propose a diffusion pipeline that directly regresses SMPL-X poses from an unconstrained domain of discourse, without relying on any intermediate representations such as glosses.

### 2.2. Sign Language Datasets

A major contributing factor to the slow-paced advancements in sign language research is the absence of large-scale datasets [8]. Earlier datasets were designed with a focus on sign language recognition using *isolated signs* [4, 31, 33, 59, 60], containing a limited vocabulary. To address the challenges of sign language recognition and translation within the context of complete sentences, several continuous sign language datasets have been introduced. More specifically, RWTH-BOSTON-50 [61], Dreuw *et al*. [18], SIGNUM [1], and BSL [52] along with DictaSign Corpus, which was developed in several languages [7, 20, 21], were among the first datasets with sentence level annotations. While additional datasets featuring an expanded set of signs have been introduced [12, 31, 33], it is crucial to emphasize the importance of continuous sign language

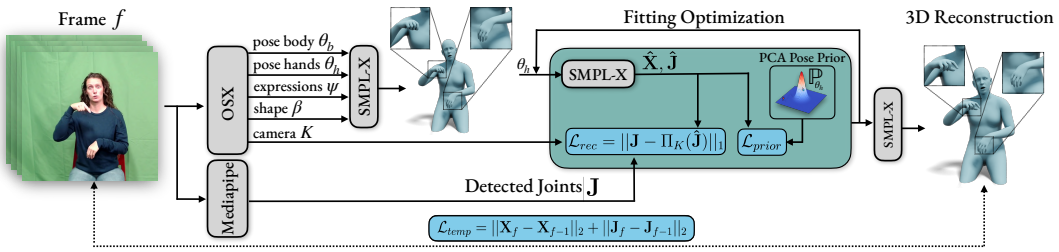


Figure 2. **Overview of the fitting pipeline.** A set of input frames $F$ are first processed by OSX [36] to obtain an initial set of pose parameters $\mathbf{p}_{1:F}^{init}$. Then, using the Mediapipe algorithm [37], we fine-tune the predicted hand poses to match the detected joints $\mathbf{J}$ while constraining the hand poses $\boldsymbol{\theta_h}$ to lie in the space of plausible poses. Finally, using a temporal coherence loss, we acquire smooth and high-fidelity annotations of 3D signing avatars.

for the purposes of translation and production. S-pot [58] was among the first large-scale continuous sign language datasets with a vocabulary of over 1K signs of Finnish sign language collected in a constrained environment. To enforce the robustness of sign language translation methods, RWTH-Phoenix [9] contained a collection of TV clips with German sign language, remaining amongst one of the most popular datasets used for sign language translation [10] and production [49]. Similarly, BSL-1K, as presented in [2], curated a dataset featuring British sign language (BSL) used in casual conversations, encompassing a total vocabulary of 1K signs. Recently, Duarte *et al*. [19] collected How2Sign, a large-scale dataset of American Sign Language (ASL), that is aligned with speech signals from the How2 dataset [46]. How2Sign is equipped with a vocabulary of over 16K signs, captured in a total of seventy-nine hours of continuous sign language. A major limitation of the above datasets, is the lack of available 3D annotations, that not only aid the translation tasks, but are also essential for training realistic 3D signing avatars. In this work, we have extended the How2Sign dataset by incorporating high-quality SMPL-X [41] annotations, thereby establishing it as the first publicly available 3D sign language dataset.

## 3. Dataset

To train a high-fidelity SLP method, capable of generating realistic sign actors, we curate a large-scale dataset of 3D dynamic ASL sign sequences paired with their corresponding text transcripts. To do so, we devise a robust 4D reconstruction pipeline, crafted specifically for hand gestures, to estimate dynamic hand and body poses of signing avatars in the SMPL-X format [41]. How2Sign dataset [19] provides the optimum candidate since it is composed of 35K high-resolution clips of *co-articulated* ASL with a substantial vocabulary size featuring over 16K word tokens.

To acquire high-fidelity 4D reconstructions of How2Sign clips, we build our pipeline upon the powerful OSX [36]. Specifically, we initialize our fitting optimization using the SMPL-X pose and shape parameters acquired from OSX for each one of the $F$ frames of the clip as:

$$\mathbf{p}_{1:F}^{init} = [\boldsymbol{\theta_b}||\boldsymbol{\theta_h}||\boldsymbol{\psi}||\boldsymbol{\beta}], \tag{1}$$

where $\boldsymbol{\theta_b}, \boldsymbol{\theta_h}, \boldsymbol{\psi}, \boldsymbol{\beta}$ denote the body pose, hand pose, expression, and shape parameters respectively, and $||$ the concatenation symbol.

Recognizing that hand poses constitute the pivotal component in conveying SL, we adopt an optimization procedure to enhance the precision of hand poses and rectify any potential misalignments by leveraging body and hand joints detected from the Mediapipe framework [37]. More specifically, we optimize the initial pose parameters $\mathbf{p}_{1:F}^{init}$ to minimize the re-projection loss $\mathcal{L}_{rec}$ between the regressed joints $\hat{\mathbf{J}}_{1:F}$ and the joints predicted from Mediapipe $\mathbf{J}_{1:F}$:

$$\mathcal{L}_{rec} = ||\mathbf{J}_{1:F} - \Pi_K(\hat{\mathbf{J}}_{1:F})||_1, \tag{2}$$

where $\Pi_K$ is the intrinsic camera projection matrix. Following extensive experimentation, we observed that optimization is only necessary for the arm and hand joints, as the OSX-regressed body joints are sufficiently accurate. Fitting hand poses using 2D keypoints is an exceptionally challenging task, primarily due to the numerous articulations and the inherent ambiguities within the solutions. While several methods have been proposed to constrain SMPL to feasible body poses [17, 41, 57], pose prior models for the hand models [42, 44] remain unexplored. To constrain the optimization to plausible human and hand poses, we propose a simple but intuitive approach using Principal Component Analysis (PCA) to model the subspace of anatomically feasible poses. Specifically, we trained a PCA pose prior model on two large datasets of human body [38] and hand [22] poses, to model the distribution of feasible arm and hand poses. To formulate the prior loss we measure the reconstruction error of a mesh $\mathbf{X}$ projected and reconstructed from the PCA space $\mathbf{U}$ as:

$$\mathcal{L}_{prior} = ||\mathbf{X} - [(\mathbf{X} - \boldsymbol{\mu})\mathbf{U}^T]\mathbf{U} + \boldsymbol{\mu}||_2, \tag{3}$$

where $\mathbf{U} \in \mathbb{R}^{N \cdot 3 \times d}$ is the eigenvector basis of $d$ components and $\boldsymbol{\mu}$ is the mean mesh. Intuitively, realistic poses

will result in smaller reconstruction errors compared to infeasible articulations.

Finally, a common issue with blurry videos is that the OSX reconstruction and the Mediapipe detections may include jittering and spatio-temporal noise. To tackle this, we enforce temporal coherence using a loss function on both vertex and joint space that enforces smooth transitions between adjacent frames $f$, $f-1$:

$$\mathcal{L}_{temp} = ||\mathbf{X}_f - \mathbf{X}_{f-1}||_2 + ||\mathbf{J}_f - \mathbf{J}_{f-1}||_2. \quad (4)$$

The overall loss function can be defined as:

$$\mathcal{L} = \mathcal{L}_{rec} + \lambda_{prior}\mathcal{L}_{prior} + \lambda_{temp}\mathcal{L}_{temp}, \quad (5)$$

where $\lambda_{prior}, \lambda_{temp}$ are hyperparameters. An overview of the proposed fitting pipeline is depicted in Fig. 2.

# 4. Method

We propose Neural Sign Actors, a diffusion-based generative model that generates motion sequences conditioned on text transcripts. Similar to traditional diffusion architectures, our method is composed of the deterministic forward diffusion process that gradually adds noise to the input distributions and the reverse denoising model $\epsilon_\theta(\cdot)$ that predicts the noise introduced by the forward process at each time-step. To reduce the computational requirements and facilitate the generation quality, we train a diffusion model on the low-dimensional pose space defined by SMPL-X [41] model instead of the vertex space. Given that sign language is solely related to hand motion and facial expressions, we focus on modeling the pose and the expression parameters on the canonical shape. An overview of the proposed approach can be found in Fig. 3.

## 4.1. Forward Diffusion Process

During the forward diffusion process, noise sampled from a Gaussian distribution $\mathcal{N}(\mu, \sigma\mathbf{I})$ is gradually added to the sequence of SMPL-X parameters $\mathbf{p}_{1:F}$, which consist of the concatenated poses $\boldsymbol{\theta}_{1:F}$, and expressions $\boldsymbol{\psi}_{1:F}$. This process iterates a total of $T$ times as a Markov chain, ultimately transforming the poses into a Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$. In line with the approach delineated in [29], we establish the forward diffusion process as follows:

$$q(\mathbf{p}^t_{1:F}|\mathbf{p}^{t-1}_{1:F}) = \mathcal{N}(\mathbf{p}^t_{1:F}|\sqrt{\alpha_t}\mathbf{p}^{t-1}_{1:F}, (1-\alpha_t)\mathbf{I}) \quad (6)$$

where $\alpha_t$ is the variance schedule parameter that controls the noise scheduling of the process.

## 4.2. Reverse Diffusion Process

Following the forward process, the goal of the denoising module $\epsilon_\Theta$ is to learn the reverse process, *i.e.* learn a mapping from the noised distribution to the real pose space $p_\theta(\mathbf{p}^{t-1}_{1:F}|\mathbf{p}^t_{1:F})$. Following the reparameterization trick of [29], we train a denoising model $\epsilon_\theta$ that predicts the time conditioned noise $\epsilon_t$ as:

$$\mathcal{L}_t = ||\epsilon_t - \epsilon_\Theta(\mathbf{p}^t_{1:F}, t, \boldsymbol{w}_{1:F})||_2, \quad (7)$$

where $\epsilon_t$ is the noise added at time-step $t$ of the forward diffusion process and $w_{1:F}$ denotes the target text transcript. To further enforce the generation of accurate hand articulations, we modify $\mathcal{L}_t$ to double the weighting factor of the hand poses. The proposed denoising module can be divided into three main components: the anatomically informed pose and expression encoders, the text encoder, and the auto-regressive decoder.

**Anatomically Informed Encoder.** Previous methods for human motion generation attempted to model poses and joint rotations independently, using permutation equivariant layers such as MLPs [13, 27]. We observed that such equivariance limits the generative ability of the network and results in mild motion intensities. To tackle this limitation, we propose to break the permutation equivariance using a novel, anatomically inspired, graph neural network (GNN) combined with a pose embedding layer. In particular, for a joint $i$, we build a message passing layer that updates the joint $i$ features $\mathbf{f}_i$ based on the relative features of the SMPL-X kinematic tree $\mathcal{K}$. Additionally, to break the permutation equivariance of the proposed message passing layer, we introduce a pose embedding that encodes joint index $i$ into a unique token feature $\mathcal{P}_i$. With this formulation, the network learns to disentangle the joint distributions since each joint is uniquely defined by its token feature $\mathcal{P}$. The update function of the proposed message passing layer can be defined as:

$$\mathbf{f}'_i = \gamma\left(\sum_{j\in\mathcal{K}_i} g_{ij}(\mathbf{f}_j - \mathbf{f}_i) + \mathcal{P}_i\right), \quad (8)$$

where $\mathbf{f}_j$ denotes the features of joint $j$ which is anatomically connected to joint $i$, in the kinematic tree $\mathcal{K}$, $g_{ij}$ is an anisotropic function between the joints $i, j$, $\mathcal{P}_i$ refers to the positional encoding of joint $i$, and $\gamma$ is a non-linearity. We establish the anisotropy of kernel $g_{ij}$ by assigning a different set of learnable weights to each set of neighbors.

Similarly, to break the permutation equivariance of the expression encoder layers, we append each of the expression parameters with a learnable expression token $\mathcal{E}$. Given that expression blendshapes cannot be represented in graph form, we utilize an MLP to encode their latent features as:

$$\mathbf{g}'_i = \gamma\left(\text{MLP}\left(\mathbf{g}_i + \mathcal{E}_i\right)\right), \quad (9)$$

where $\mathbf{g}_i$ denotes the latent features of expression parameter $i$ and $\mathcal{E}_i$ refers to its corresponding expression embedding.

**Text Encoding.** Sign language is not merely a direct word-for-word translation of spoken language, rather it possesses

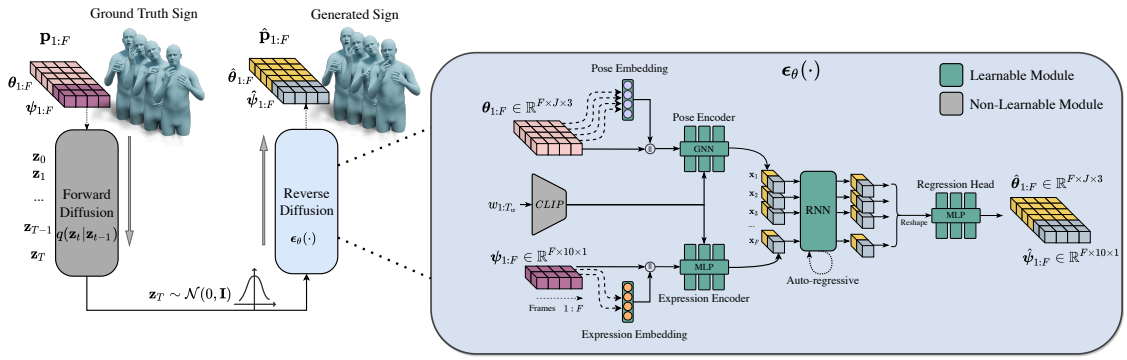


Figure 3. **Overview of the proposed method.** We employ a diffusion model to learn a mapping between text scripts and 3D sign language. The proposed framework consists of an auto-regressive denoising module $\epsilon_\Theta$ that is founded on the novel anatomically informed pose encoder to model the sign motions.

Table 1. Mean per vertex error (mm) of the proposed and the baseline methods on the SGNify mocap dataset [25].

| Method | Body | Left Hand | Right Hand |
|---|---|---|---|
| FrankMoCap[45] | 78.07 | 20.47 | 19.62 |
| PIXIE[23] | 60.11 | 25.02 | 22.42 |
| PyMAF-X [63] | 68.61 | 21.46 | 19.19 |
| SMPLify-X [41] | 56.07 | 22.23 | 18.83 |
| SGNify [25] | 55.63 | 19.22 | 17.50 |
| OSX [36] | 47.32 | 18.34 | 18.12 |
| **Proposed** | **46.42** | **16.17** | **15.23** |

its own unique grammar, semantic structure, and distinct language logic [34]. Contingent upon this, we avoid using a sequence of word embeddings to condition the motion generation and propose to utilize CLIP [43] as a powerful sentence encoder that is able to generalize to arbitrary text prompts. We condition pose and expression encoders on the text embedding using a gating approach described in [26].

**Auto-regressive Decoder.** Considering that motion can be conceptualized as a sequence of poses where each pose is contingent upon its predecessor, we constructed our motion generative network utilizing an auto-regressive model. As we experimentally illustrate in Sec. 5.4, we utilize a Long-Short-Term-Memory (LSTM) model as our pose decoder since it has less memory requirement than transformer architectures and provides better auto-regressive capabilities. Finally, we map the output of the autoregressive model back to the pose space using an MLP layer.

# 5. Experiments

## 5.1. Dataset Evaluation

Given that accurate 3D annotations are a requisite for the training of a potent SLP model, we quantitatively and qualitatively evaluated the performance of the pipeline introduced in Sec. 3 on the task of sign language reconstruction from videos. To assess the fitting quality, we apply the proposed pipeline to the SGNify mocap dataset [25] that contains ground truth annotations. In Tab. 1, we report the

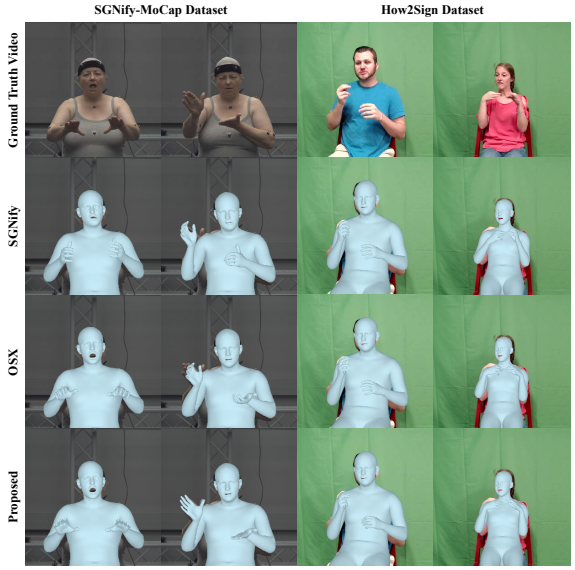


Figure 4. Qualitative comparison between the proposed and the baseline fitting frameworks on SGNify [25] and How2Sign [19].

reconstruction error of the proposed pipeline and compare it with OSX [36] and the SGNify method, which is the current state-of-the-art method for 3D fitting from SL videos. It must be noted that unlike the SGNify method, the proposed fitting pipeline achieves a smaller reconstruction error despite not including any SL-driven losses. The powerful prior model that guides the fitting optimization leads our method to valid poses and articulation Fig. 4.

## 5.2. Sign Language Production

**Baselines.** To evaluate the performance of our method we selected the current state-of-the-art methods for text-driven sign language generation, *i.e.* Saunders *et al.* [49], Saunders

Table 2. Quantitative evaluation of the proposed and the baseline methods on the How2Sign test dataset.

| | Body | | | | Left Hand | | | | Right Hand | | | | Back-Translation | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Method** | **MPVPE ↓** | **MPJPE ↓** | **FID ↓** | **DTW ↓** | **MPVPE** | **MPJPE** | **FID** | **DTW** | **MPVPE** | **MPJPE** | **FID** | **DTW** | **BLEU-4 ↑** | **BLEU-3 ↑** | **BLEU-2 ↑** | **BLEU-1 ↑** | **ROUGE ↑** |
| Saunders *et al.* [49] | 67.21 | 70.06 | 4.71 | 14.15 | 73.49 | 74.13 | 0.68 | 11.21 | 75.57 | 77.47 | 0.75 | 11.93 | 2.75 | 5.87 | 8.21 | 13.82 | 29.87 |
| Saunders *et al.* [48] | 63.19 | 65.25 | 3.98 | 13.78 | 71.43 | 72.39 | 0.59 | 11.02 | 68.54 | 70.14 | 0.51 | 11.32 | 6.21 | 8.98 | 12.01 | 18.22 | 32.33 |
| Hwang *et al.* [30] | 62.74 | 63.25 | 4.45 | 13.94 | 78.95 | 70.34 | 0.63 | 11.33 | 68.65 | 69.59 | 0.60 | 12.26 | 5.75 | 8.21 | 11.62 | 17.55 | 31.98 |
| Stoll *et al.* [54] | 55.02 | 60.32 | 4.96 | 13.99 | 68.48 | 69.45 | 0.56 | 11.59 | 60.18 | 62.73 | 0.64 | 12.29 | 7.51 | 10.72 | 13.92 | 19.56 | 33.17 |
| **Proposed** | **31.47** | **35.87** | **1.56** | **7.83** | **36.24** | **38.82** | **0.24** | **6.74** | **39.68** | **40.56** | **0.36** | **7.91** | **13.12** | **18.25** | **25.44** | **41.31** | **47.55** |

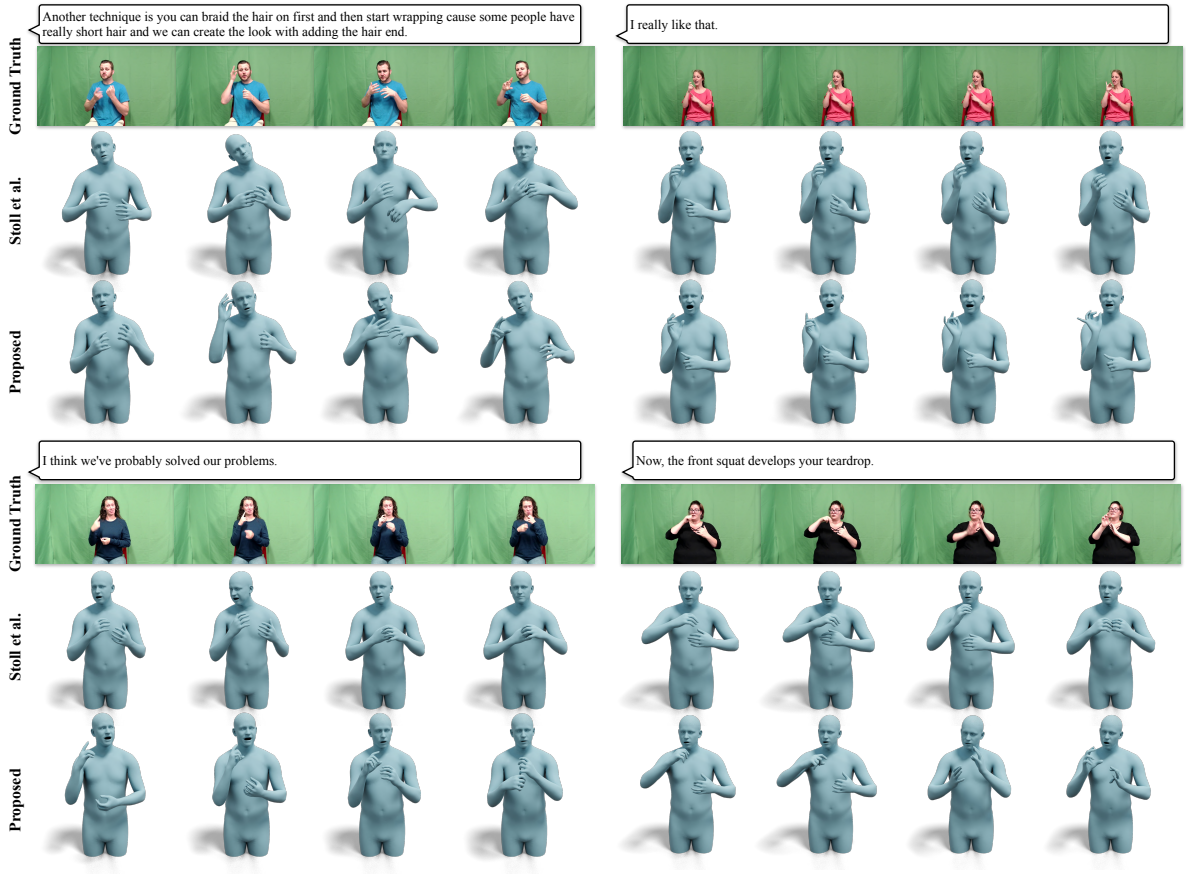


Figure 5. Qualitative comparison of generated signs conditioned on the text transcript between the proposed and Stoll *et al.* [54] methods. The ground truth video is given for reference.

*et al.* [48], Hwang *et al.* [30] and Stoll *et al.* [54]. Given that all methods have been trained on the German Sign Language RWTH-Phoenix dataset [9], we retrained the models using the same training set-up as the proposed method.

**Implementation Details.** To train the diffusion model we followed the implementation details of [29]. We implemented pose and expression embedding layers using a simple linear projection. Pose and expression encoders are composed of 4 stacked GNN and MLP layers, respectively, with an increasing number of channels. We employed the CLIP-ViT-L-14 model as our text encoder. The RNN decoder consists of 4 LSTM layers. We trained our model for 2K epochs using the Adam optimizer with a linearly decreasing learning rate from $10^{-3}$ to $10^{-6}$.

**Evaluation Metrics.** We quantitatively evaluate the generation quality of the proposed and the baseline methods under a set of metrics. The first two were Mean Per Vertex Position Error (MPVPE) and Mean Per Joint Position Error (MPJPE). Given that the motions generated from the proposed and baseline methods may not be correctly aligned with the ground truth annotations, we used Dynamic Time Warping (DTW) [5] to measure the similarity between the generated and the original sign sequences. To evaluate the quality of the generated poses we measured the Fréchet inception distance (FID) score between the generated and the ground truth poses. Finally, following [49], we trained a

Transformer-based back-translation network to map pose sequences back to text. For additional details, we refer the reader to the supplementary material.

**Evaluation of Generated Signs.** In the first three columns of Tab. 2 we quantitatively compare the proposed and the baseline methods on the test set of the curated dataset. The proposed method manages to outperform the baselines under all metrics, even by a large margin. Specifically, the generated signs not only demonstrate low reconstruction error across the entire upper body but also exhibit significant improvements on the hands region. Additionally, the proposed method is able to generate articulations that match the ground truth signs, which is translated to low FID scores. This can be also validated in Fig. 5, where the proposed method is able to generate signs with high-frequency articulations that match the ground truth videos. In contrast, current state-of-the-art SLP methods fail to model high-frequency articulations and can only generate small deviations around the canonical pose. This is quantified in Fig. 6, where we report the average per-frame pose deviations. The proposed method not only produces a larger variety of poses, compared to the small deviations of Stoll *et al*. [54], but also follows the ground truth pose distribution. To enhance readability, we focus on Stoll *et al*. for qualitative comparisons, as the best performing prior work.

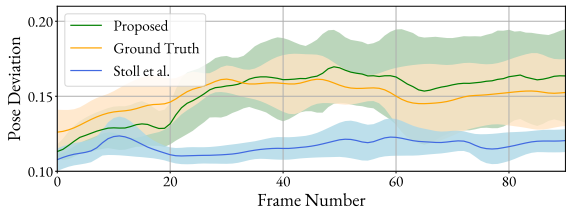


Figure 6. Mean and Standard Deviation of the absolute pose value across the sequence.

**3D Pose Back-Translation.** We additionally evaluated the overall quality of the generated pose sequences using back-translation, which measures how much information of the input sentences has been maintained in the model's output. In particular, we trained a Transformer-based architecture on our curated How2Sign dataset to learn a mapping from pose sequences back to the original text transcripts. We then translated the generated 3D pose sequences of How2Sign back to spoken language using our back-translation network. To comply with the evaluations in [49, 54] we report BLEU n-grams from 1 to 4 and ROUGE scores. We repeated the above process for sequences generated by our model, as well as the baseline models [30, 48, 49, 54] and summarize the results in the last column of Tab. 2. Our model produces the highest back translation scores across all metrics, with BLEU-4 being at the level of BLEU-2 of the second best performing method (Stoll *et al*. [54]).

## 5.3. User-Study

Although evaluation metrics can provide insights into a network's performance, the most critical benchmark lies in the perceptual evaluation from Hard of Hearing individuals. Notably, we further assess the realism of the generated signs by designing a user study where 15 ASL fluent subjects, with ages ranging from 29 to 62, evaluated the generated signs. We divided the perceptual study into two parts, to assess: (a) how aligned the generated signs are with respect to the text transcripts and (b) the fidelity and readability of the proposed generations. For the first part of the user study, we presented 15 different generated signs from both the proposed and baseline methods, alongside the ground truth video and its corresponding fitting. Participants were asked to assign a value between 1-10 rating the alignment of each method with the corresponding text transcript.

To avoid potential biases between the methods, all videos were shown in a random order. In Fig. 7, we report the results of the first part of the user-study. As expected, the ground truth videos achieve the best average score of 8.7 while the fittings achieve slightly less with 8.1, which quantifies the high quality of the generated annotations. The proposed method achieves an average score of 5.8 whereas the method of Stoll *et al*. [54] fails to achieve reasonable results.

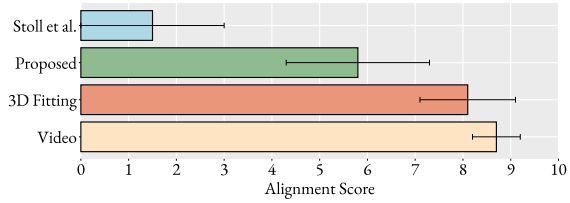


Figure 7. Human Evaluation of the alignment between the generated signs and the text transcript.

The second part of the perceptual study aimed to assess the fidelity of the generated signs. Each participant was shown 15 rendered videos with signs generated by the proposed method and was asked to rank five candidate text translations, from most likely to least likely translation. Apart from the ground truth, the candidate translations included both similar to the ground truth sentences with slightly modified meanings, *i.e*. by masking words, cropping sentences, or changing the word order, and also unrelated sentences from different topics and domains. Measuring the cumulative accuracy, the generated signs attained a top-1 accuracy of 40% and a top-2 accuracy of 80%, affirming the realism and fidelity of the generated signs.

## 5.4. Ablation

The proposed method consists of four main components: the anatomically inspired pose encoder, the pose and expression embeddings, the autoregressive decoder, and the

Table 3. Evaluation of individual components in the proposed method. Every row refers to a different ablated module. We include the performance of our method for reference.

| Method | MPVPE | MPJPE | FID | DTW |
|---|---|---|---|---|
| w/o GNN | 37.51 | 38.23 | 2.85 | 9.19 |
| w/o Pose and Expression Embedding | 66.56 | 68.34 | 6.65 | 11.98 |
| w/o LSTM | 36.17 | 39.46 | 2.12 | 10.82 |
| w. Transformer | 32.73 | **35.41** | 1.58 | 8.17 |
| w/o CLIP | 69.12 | 71.42 | 5.36 | 12.84 |
| w. BERT | 45.32 | 47.11 | 2.23 | 9.21 |
| Tevet *et al*. [56] | 36.11 | 38.23 | 2.45 | 8.92 |
| Chen *et al*. [13] | 35.23 | 37.12 | 2.15 | 8.26 |
| **Proposed** | **31.47** | 35.87 | **1.56** | **7.83** |

text encoding. In this section, we evaluate the contribution of each component to the final generations of the model.

**Effect of Pose Encoder and Embedding Layers.** Initially, we ablate the novel pose encoder and we substitute it with an MLP layer (*w/o GNN*), similar to the expression encoder. Under this setting, the generated motions are smoother and mainly deviate around the mean pose. Aligned with our hypothesis, updating poses in an anatomically inspired manner enables the network to learn higher frequencies, such as rare articulations, without an increase in the network's capacity. Additionally, as mentioned in Sec. 4.2, in the traditional setting of motion diffusion models, during the denoising step, poses and expressions are sampled from a Gaussian distribution and are then decoded back to their original space using a permutation equivariant network, such as an MLP. Such permutation equivariance, treats poses under a uniform setting that limits the generative power of the network. As shown in Tab. 3, without pose and expression embedding layer (*w/o Pose and Expression Embedding*), the model fails to produce any reasonable sign, resulting in a performance drop under all metrics.

**Effect of LSTM Encoder.** A core part of the proposed method is the autoregressive LSTM decoder. We initially evaluate its contribution compared to a simple frame positional encoding to transform the network to a frame-conditioned generative model, without having any temporal module (*w/o LSTM*). As expected, this results in poor DTW performance and generations that present increased jittering and lack of temporal coherence. Furthermore, we substitute the LSTM layer with a Transformer encoder layer (*w. Transformer*). Interestingly, the LSTM layer achieves similar performance to the Transformer layer while having 75% fewer parameters (1M vs. 4.5M).

**Effect of Text Encoding.** The text encoding module has a pivotal effect on motion generation. Firstly, we substitute the CLIP encoder with a word-level embedding layer that is trained with the rest of the method in an end-to-end fashion. We set the embedded size to 256 although we did not observe significant differences in performance. Following this, we utilized the pretrained DistilBERT [47], whose parameters remain frozen throughout training. As depicted in Tab. 3, training word embeddings from scratch strongly affects the generalization of the network, leading to large MPVPE and MPJPE metrics. In contrast, DistilBERT achieves better performance than learnable word embeddings, although it does not outperform CLIP embeddings. This is aligned with our assumption that sentence embeddings could provide better insights regarding the meaning of a sentence compared to word-level embeddings. Especially in the task of SLP, where there is not an explicit one-to-one mapping between words and poses, sentence level embeddings provide a more powerful text encoding solution.

**Comparison with Human Motion Diffusion Models.** Finally, we compare our model with state-of-the-art methods on human motion modeling [13, 56], which can be considered as deviations from the proposed framework. Unlike our anatomically inspired approach, both models rely on linear layers to handle pose motions, constraining their capacity to encode intricate hand movements with high-frequency details. In particular, although both methods can achieve smooth body motions, they fail to produce accurate hand articulations that match the ground truth distribution, which can be validated from the reconstruction errors (MPVPE, MPVJE), along with the FID measure.

## 6. Conclusion

Neural 3D sign language production is an important challenge that aims to aid the Deaf and Hard of Hearing community and can effectively increase their inclusion in any social environment. In this work, we made a step towards high fidelity 3D SLP, by deriving a large-scale 3D dataset to train a text conditioned diffusion-based model. The release of additional relevant databases will enable the training of even more robust architectures. We initially introduce a precise 3D sign language reconstruction pipeline that outperforms previous SL reconstruction methods. Then, we train a motion generative model using an autoregressive diffusion model. The core of our method is founded on a novel, anatomically inspired, graph neural network that learns the pose distribution and enables highly detailed articulations. Importantly, leveraging the powerful CLIP text embeddings, the proposed model can generalize to out-of-distribution samples. Extensive experiments on sign language generation tasks, including a perceptual study with ASL fluent subjects, demonstrate the superiority of the proposed method compared to the previous approaches.

**Acknowledgements.** S. Zafeiriou was supported by EPSRC Project DEFORM (EP/S010203/1) and GNOMON (EP/X011364). R.A. Potamias was supported by EPSRC Project GNOMON (EP/X011364).

## References


[1] Ulrich von Agris and Karl-Friedrich Kraiss. Signum database: Video corpus for signer-independent continuous sign language recognition. In *sign-lang@ LREC 2010*, pages 243–246. European Language Resources Association (ELRA), 2010. 2

[2] Samuel Albanie, Gül Varol, Liliane Momeni, Triantafyllos Afouras, Joon Son Chung, Neil Fox, and Andrew Zisserman. Bsl-1k: Scaling up co-articulated sign language recognition using mouthing cues. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 35–53. Springer, 2020. 3

[3] Epameinondas Antonakos, Anastasios Roussos, and Stefanos Zafeiriou. A survey on mouth modeling and analysis for sign language recognition. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–7. IEEE, 2015. 1

[4] Vassilis Athitsos, Carol Neidle, Stan Sclaroff, Joan Nash, Alexandra Stefan, Quan Yuan, and Ashwin Thangali. The american sign language lexicon video dataset. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8. IEEE, 2008. 2

[5] Donald J Berndt and James Clifford. Using dynamic time warping to find patterns in time series. In *Proceedings of the 3rd international conference on knowledge discovery and data mining*, pages 359–370, 1994. 6

[6] P Boyes Braem and RL Sutton-Spence. *The Hands Are The Head of The Mouth. The Mouth as Articulator in Sign Languages*. Hamburg: Signum Press, 2001. 1

[7] Annelies Braffort, Laurence Bolot, Emilie Chételat-Pelé, Annick Choisier, Maxime Delorme, Michael Filhol, Jérémie Segouat, Cyril Verrecchia, Flora Badin, and Nadège Devos. Sign language corpora for analysis, processing and evaluation. In *LREC*, 2010. 2

[8] Danielle Bragg, Oscar Koller, Mary Bellard, Larwan Berke, Patrick Boudreault, Annelies Braffort, Naomi Caselli, Matt Huenerfauth, Hernisa Kacorri, Tessa Verhoef, et al. Sign language recognition, generation, and translation: An interdisciplinary perspective. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility*, pages 16–31, 2019. 2

[9] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Neural sign language translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7784–7793, 2018. 2, 3, 6

[10] Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. Sign language transformers: Joint end-to-end sign language recognition and translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10023–10033, 2020. 1, 2, 3

[11] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 2

[12] Xiujuan Chai, Hanjie Wang, and Xilin Chen. The devisign large vocabulary of chinese sign language database and baseline evaluations. In *Technical report VIPL-TR-14-SLR-001. Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS)*. Institute of Computing Technology, 2014. 2

[13] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18000–18010, 2023. 4, 8

[14] Yutong Chen, Fangyun Wei, Xiao Sun, Zhirong Wu, and Stephen Lin. A simple multi-modality transfer learning baseline for sign language translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5120–5130, 2022. 1

[15] Yutong Chen, Ronglai Zuo, Fangyun Wei, Yu Wu, Shujie Liu, and Brian Mak. Two-stream network for sign language recognition and translation. *Advances in Neural Information Processing Systems*, 35:17043–17056, 2022. 1

[16] Stephen Cox, Michael Lincoln, Judy Tryggvason, Melanie Nakisa, Mark Wells, Marcus Tutt, and Sanja Abbott. Tessa, a system to aid communication with deaf people. In *Proceedings of the fifth international ACM conference on Assistive technologies*, pages 205–212, 2002. 2

[17] Andrey Davydov, Anastasia Remizova, Victor Constantin, Sina Honari, Mathieu Salzmann, and Pascal Fua. Adversarial parametric pose prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10997–11005, 2022. 3

[18] Philippe Dreuw, David Rybach, Thomas Deselaers, Morteza Zahedi, and Hermann Ney. Speech recognition techniques for a sign language recognition system. *hand*, 60:80, 2007. 2

[19] Amanda Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier Giro-i Nieto. How2sign: a large-scale multimodal dataset for continuous american sign language. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2735–2744, 2021. 2, 3, 5

[20] Eleni Efthimiou, Stavroula-Evita Fotinea, Thomas Hanke, John Glauert, Richard Bowden, Annelies Braffort, Christophe Collet, Petros Maragos, and François Goudenove. Dicta-sign: sign language recognition, generation and modelling with application in deaf communication. In *sign-lang@ LREC 2010*, pages 80–83. European Language Resources Association (ELRA), 2010. 2

[21] Eleni Efthimiou, Stavroula-Evita Fotinea, Thomas Hanke, John Glauert, Richard Bowden, Annelies Braffort, Christophe Collet, Petros Maragos, and François Lefebvre-Albaret. The dicta-sign wiki: Enabling web communication for the deaf. In *Computers Helping People with Special Needs: 13th International Conference, ICCHP 2012, Linz, Austria, July 11-13, 2012, Proceedings, Part II 13*, pages 205–212. Springer, 2012. 2

[22] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J. Black, and Otmar Hilliges. ARCTIC: A dataset for dexterous bimanual hand-

object manipulation. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[23] Yao Feng, Vasileios Choutas, Timo Bolkart, Dimitrios Tzionas, and Michael Black. Collaborative regression of expressive bodies using moderation. In *International Conference on 3D Vision (3DV)*, pages 792–804, 2021. 5

[24] Jens Forster, Christoph Schmidt, Oscar Koller, Martin Bellgardt, and Hermann Ney. Extensions of the sign language recognition and translation corpus rwth-phoenix-weather. In *LREC*, pages 1911–1916, 2014. 2

[25] Maria-Paola Forte, Peter Kulits, Chun-Hao Paul Huang, Vasileios Choutas, Dimitrios Tzionas, Katherine J. Kuchenbecker, and Michael J. Black. Reconstructing signing avatars from video using linguistic priors. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 12791–12801, 2023. 5

[26] Will Grathwohl, Ricky TQ Chen, Jesse Bettencourt, Ilya Sutskever, and David Duvenaud. Ffjord: Free-form continuous dynamics for scalable reversible generative models. *arXiv preprint arXiv:1810.01367*, 2018. 5

[27] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5152–5161, 2022. 4

[28] Aiming Hao, Yuecong Min, and Xilin Chen. Self-mutual distillation learning for continuous sign language recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11303–11312, 2021. 1

[29] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 4, 6

[30] Eui Jun Hwang, Jung-Ho Kim, and Jong C. Park. Non-autoregressive sign language production with gaussian space. In *The 32nd British Machine Vision Conference (BMVC 21)*. British Machine Vision Conference (BMVC), 2021. 2, 6, 7

[31] Hamid Reza Vaezi Joze and Oscar Koller. Ms-asl: A large-scale data set and benchmark for understanding american sign language. *arXiv preprint arXiv:1812.01053*, 2018. 2

[32] Parul Kapoor, Rudrabha Mukhopadhyay, Sindhu B Hegde, Vinay Namboodiri, and CV Jawahar. Towards automatic speech to sign language generation. *arXiv preprint arXiv:2106.12790*, 2021. 2

[33] Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. 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*, pages 1459–1469, 2020. 2

[34] Dongxu Li, Chenchen Xu, Xin Yu, Kaihao Zhang, Benjamin Swift, Hanna Suominen, and Hongdong Li. Tspnet: Hierarchical feature learning via temporal semantic pyramid for sign language translation. *Advances in Neural Information Processing Systems*, 33:12034–12045, 2020. 5

[35] Zeyu Liang, Huailing Li, and Jianping Chai. Sign language translation: A survey of approaches and techniques. *Electronics*, 12(12):2678, 2023. 1, 2

[36] Jing Lin, Ailing Zeng, Haoqian Wang, Lei Zhang, and Yu Li. One-stage 3d whole-body mesh recovery with component aware transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21159–21168, 2023. 3, 5

[37] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019. 3

[38] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. Amass: Archive of motion capture as surface shapes. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019. 3

[39] John McDonald, Rosalee Wolfe, Jerry Schnepp, Julie Hochgesang, Diana Gorman Jamrozik, Marie Stumbo, Larwan Berke, Melissa Bialek, and Farah Thomas. An automated technique for real-time production of lifelike animations of american sign language. *Universal Access in the Information Society*, 15:551–566, 2016. 2

[40] Lucie Naert, Caroline Larboulette, and Sylvie Gibet. A survey on the animation of signing avatars: From sign representation to utterance synthesis. *Computers & Graphics*, 92: 76–98, 2020. 2

[41] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985, 2019. 2, 3, 4, 5

[42] Rolandos Alexandros Potamias, Stylianos Ploumpis, Stylianos Moschoglou, Vasileios Triantafyllou, and Stefanos Zafeiriou. Handy: Towards a high fidelity 3d hand shape and appearance model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4670–4680, 2023. 3

[43] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 5

[44] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *arXiv preprint arXiv:2201.02610*, 2022. 3

[45] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration. In *IEEE International Conference on Computer Vision Workshops*, 2021. 5

[46] Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. How2: a large-scale dataset for multimodal language understanding. *arXiv preprint arXiv:1811.00347*, 2018. 3

[47] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019. 8

[48] Ben Saunders, Necati Cihan Camgöz, and Richard Bowden. Adversarial training for multi-channel sign language production. In *The 31st British Machine Vision Virtual Conference*. British Machine Vision Association. 2, 6, 7

[49] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 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*, pages 687–705. Springer, 2020. 2, 3, 5, 6, 7

[50] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. Mixed signals: Sign language production via a mixture of motion primitives. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1919–1929, 2021. 2

[51] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. Signing at scale: Learning to co-articulate signs for large-scale photo-realistic sign language production. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5141–5151, 2022. 2

[52] Adam Schembri, Jordan Fenlon, Ramas Rentelis, Sally Reynolds, and Kearsy Cormier. Building the british sign language corpus. 2013. 2

[53] Stephanie Stoll, Necati Cihan Camgöz, Simon Hadfield, and Richard Bowden. Sign language production using neural machine translation and generative adversarial networks. In *Proceedings of the 29th British Machine Vision Conference (BMVC 2018)*. British Machine Vision Association, 2018. 2

[54] Stephanie Stoll, Armin Mustafa, and Jean-Yves Guillemaut. There and back again: 3d sign language generation from text using back-translation. In *2022 International Conference on 3D Vision (3DV)*, pages 187–196. IEEE, 2022. 2, 6, 7

[55] Rachel Sutton-Spence and Bencie Woll. *The linguistics of British Sign Language: an introduction*. Cambridge University Press, 1999. 1

[56] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2022. 8

[57] Garvita Tiwari, Dimitrije Antic, Jan Eric Lenssen, Nikolaos Sarafianos, Tony Tung, and Gerard Pons-Moll. Pose-ndf: Modeling human pose manifolds with neural distance fields. In *European Conference on Computer Vision (ECCV)*, 2022. 3

[58] Ville Viitaniemi, Tommi Jantunen, Leena Savolainen, Matti Karppa, and Jorma Laaksonen. S-pot - a benchmark in spotting signs within continuous signing. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1892–1897, Reykjavik, Iceland, 2014. European Language Resources Association (ELRA). 3

[59] Ulrich Von Agris, Moritz Knorr, and Karl-Friedrich Kraiss. The significance of facial features for automatic sign language recognition. In *2008 8th IEEE international conference on automatic face & gesture recognition*, pages 1–6. IEEE, 2008. 2

[60] Ronnie Wilbur and Avinash C Kak. Purdue rvl-slll american sign language database. 2006. 2

[61] Morteza Zahedi, Daniel Keysers, Thomas Deselaers, and Hermann Ney. Combination of tangent distance and an image distortion model for appearance-based sign language recognition. In *Pattern Recognition: 27th DAGM Symposium, Vienna, Austria, August 31-September 2, 2005. Proceedings 27*, pages 401–408. Springer, 2005. 2

[62] Jan Zelinka, Jakub Kanis, and Petr Salajka. Nn-based czech sign language synthesis. In *Speech and Computer: 21st International Conference, SPECOM 2019, Istanbul, Turkey, August 20–25, 2019, Proceedings 21*, pages 559–568. Springer, 2019. 2

[63] Hongwen Zhang, Yating Tian, Yuxiang Zhang, Mengcheng Li, Liang An, Zhenan Sun, and Yebin Liu. Pymaf-x: Towards well-aligned full-body model regression from monocular images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 5

[64] Benjia Zhou, Zhigang Chen, Albert Clapés, Jun Wan, Yanyan Liang, Sergio Escalera, Zhen Lei, and Du Zhang. Gloss-free sign language translation: Improving from visual-language pretraining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20871–20881, 2023. 2