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

A Additional Visualizations of SignAvatars Dataset

In this section, we present more samples and visualizations of our SignAvatars dataset for each of the subsets categorized by the annotation type: spoken language (sentence-level), HamNoSys, and word-level prompt annotation.

A.1 Qualitative Analysis of SignAvatars Dataset

We provide further details of our SignAvatars dataset and present more visualization of our data in Fig. 1, Figs. 2 and 3. Being the first large-scale multi-prompt 3D sign language (SL) motion dataset with accurate holistic mesh representations, our dataset enables 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. We also provide a demo video in the supplementary materials and our project page: https://signavatars.github.io/.

A.2 More generation samples from SignVAE

We now share snapshot examples produced from our SignVAE, demonstrating the application potential for 3D sign language production in our demo video on project page.

B Analysis of annotation pipeline

In this section, we provide further analysis of our annotation pipeline. Since there is not yet an existing benchmark for SL reconstruction while our method is not limited to SL video, we provide more in-the-wild examples with our annotation methods in Fig. 6 to demonstrate the reconstruction ability of our annotation pipeline. Moreover, Fig. 5 illustrates more qualitative comparison with state-of-the-art methods on EHF dataset [23], where we can observe that our method provides significantly better quality regarding **pixel alignment**, especially with more natural and plausible hand poses. Subsequently, the biomechanical constraints can serve as a prior for eliminating the implausible poses, which happens frequently in complex interacting-hands scenarios for other monocular capture methods, as shown in Fig. 7.

† Work done during an internship at Tencent AI Lab.
⋆ Corresponding author.

Fig. 1: More sentence-level spoken language examples of SignAvatars, *ASL* subset. We have different shapes of annotations presenting the accurate body and hand estimation.

B.1 Further comparison with State-of-the-art Methods

In this section, we compared our method against SGNify [10] . Tab. 1 and Fig. 4 clearly present the gains form our method. Note that our method is also running significantly faster than SGNify. On a single V100 GPU, our method takes $3 \sim 5$ minutes to run on a 60-frame video of HamNoSys where two hands are presented while SGNify takes hours.

C Further Evaluation of SL generation

C.1 More Experiments of Baselines

In our main paper, We adapt MDM (phrased as SignDiffuse) to use our prompt encoder as the semantic adaptor instead of the pre-trained CLIP. We provide a

Fig. 2: More HamNoSys-level examples of SignAvatars, *HamNoSys* subset. We have different shapes of annotations presenting the accurate body and hand estimation.

further comparison with different prompts using CLIP. Unfortunately, the MDM with CLIP feature did not work for complex sentence-level sign language generation and yielded random meaningless gestures that do not match the text. For the word-level generation, we follow the "*Holistic*" setting as in Tab. 6 achieving 0.246, 0.375, 0.527 for R-Precision, 4.668 for FID, 5.974 for MM-dist and 0.297, 0.411, 0.575 for MR-Precision, further demonstrating the ability of our SignVAE.

C.2 Evaluation of SignVAE generation on other benchmarks

In this section, we aim to conduct further experiments with our SignVAE on 3D SLP from spoken language on other benchmarks to further showcase its ability. To the best of our knowledge, no publicly available benchmark for **3D mesh & motion-based** SLP exists. Progressive Transformer [27] and its continuation series [26, 28, 29] on RWTH-PHOENIX-Weather 2014 T dataset [6] provides a **keypoint-based** 3D Text2Pose (Language2Motion) benchmark. Unfortunately, since, at the time of submission, this benchmark was not publicly available. Note that, conducting back-translation evaluations as in [27] must strictly follow the rule to use the same back-translation model checkpoint for a fair comparison. This is also the same for the human motion generation area, where all the evaluations should be conducted with the same evaluation checkpoints such as the

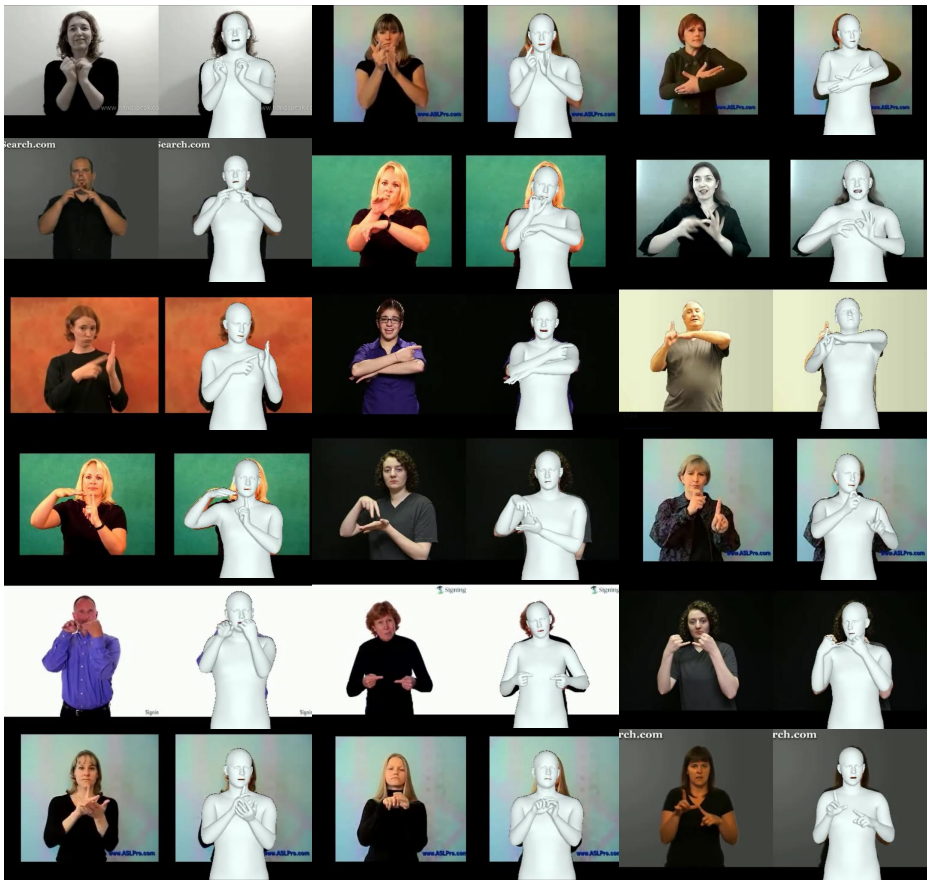

Fig. 3: More word-level examples of SignAvatars, *word* subset. We have different shapes of annotations presenting the accurate body and hand estimation.

popular HumanML3D benchmark does [12]. Unfortunately, the pretrained evaluation model checkpoint or its reproductions are available neither on the project website https://github.com/BenSaunders27/ProgressiveTransformersSLP (with an open issue) or on other sites, We have not managed to get in touch with the corresponding authors. For this reason, **we have re-evaluated the benchmark method in [27]** as follows:

Experimental Details. To conduct evaluations on Phoenix-2014T using the Progressive Transformer (PT) [27], we trained our network as well as PT on this dataset and recorded new results under our metrics. We conduct the re-evaluation experiments by:

– First, we generate mesh annotations for the Phoenix-2014T dataset and add them as our subsets GSL. We follow the original data distribution and official split to train our network.

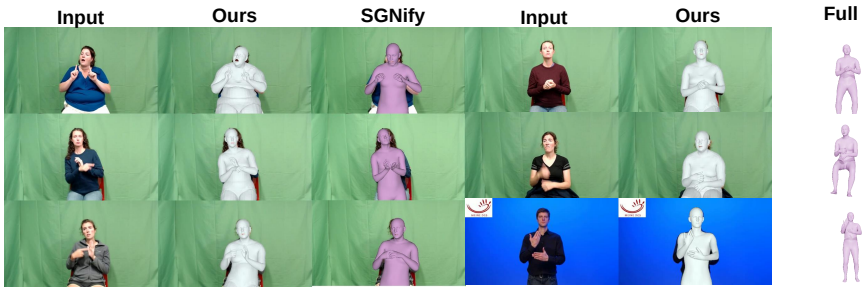

Fig. 4: Qualitative comparisons with SGNify (left) and full-body samples from our dataset (right).

Method	Upper Body	Left Hand	Right Hand	Both Hands	Avg. Runtime (min)
SMPLify-SL	56.07	22.23	18.83	20.53	-
SGNify	55.63	19.22	17.50	18.36	> 480
Ours	**38.5**	**15.56**	**13.28**	**14.42**	5

Table 1: Quantitative comparison with SGNify on their released ground truth mocap annotations. we compute the mean per-vertex error following SGNify to remove the lower body and face.

– Second, because in addition to the absence of the evaluation model, the generation model checkpoints are also lacking, we re-train PT using the official implementation on both 3D-lifted OpenPose keypoints J_{PT} and the 3D keypoints J_{ours} regressed from our mesh representation, corresponding to PT (J_{PT}) and PT (J_{ours}).
– Third, we train two 3D keypoints-based SL motion evaluation models on this subset with J_{PT} and J_{ours}, resulting in two model checkpoints C_{PT} and C_{ours}.

Comparisons. We conduct both quantitative and qualitative comparisons between the PT and our method, following the official split with both C_{PT} and C_{ours} in Tab. 2 under our evaluation metrics introduced in Sec. 5 and Appendix D.1. As shown in Tab. 2, our method significantly outperforms PT, especially regarding the R-precision and MR-precision, which indicates better prompt-motion consistency. Moreover, we can discover from the evaluation of Real Motion that the evaluation model C_{ours} utilizing the 3D keypoints J_{ours} regressed from our mesh representation can provide essentially better matching accuracy with less noise (MM-dist) than the noisy canonical 3d-lifted *OpenPose* keypoints J_{PT}, yielding better performance than using C_{PT}. A carefully designed evaluation model with proper training data will significantly improve the ability to reflect the authentic performance of the experiments and will be less likely to disturb our analysis as those in the results of C_{PT}.

Eval. Model	Method	R-Precision(↑)			FID (↓)	MM-dist (↓)	MR-Precision (↑)		
		top 1	top 3	top 5			top 1	top 3	top 5
$\mathbf{C}_{PT}$	**Real Motion**	$0.193^{\pm.006}$	$0.299^{\pm.002}$	$0.413^{\pm.005}$	$0.075^{\pm.066}$	$5.151^{\pm.033}$	-	-	-
	PT (J_{PT})	$0.035^{\pm.009}$	$0.082^{\pm.005}$	$0.195^{\pm.004}$	$4.855^{\pm.062}$	$7.977^{\pm.023}$	$0.088^{\pm.012}$	$0.145^{\pm.012}$	$0.212^{\pm.019}$
	PT (J_{ours})	$0.078^{\pm.004}$	$0.149^{\pm.002}$	$0.267^{\pm.003}$	$5.135^{\pm.024}$	$8.135^{\pm.019}$	$0.138^{\pm.009}$	$0.195^{\pm.023}$	$0.311^{\pm.011}$
	Ours	$0.165^{\pm.006}$	$0.275^{\pm.009}$	$0.356^{\pm.003}$	$4.194^{\pm.037}$	$4.899^{\pm.029}$	$0.219^{\pm.017}$	$0.325^{\pm.015}$	$0.443^{\pm.056}$
$\mathbf{C}_{ours}$	**Real Motion**	$0.425^{\pm.004}$	$0.635^{\pm.006}$	$0.733^{\pm.009}$	$0.015^{\pm.059}$	$2.413^{\pm.051}$	-	-	-
	PT (J_{PT})	$0.095^{\pm.004}$	$0.155^{\pm.005}$	$0.286^{\pm.002}$	$3.561^{\pm.035}$	$4.565^{\pm.027}$	$0.175^{\pm.002}$	$0.301^{\pm.010}$	$0.419^{\pm.034}$
	PT (J_{ours})	$0.134^{\pm.002}$	$0.285^{\pm.003}$	$0.395^{\pm.005}$	$3.157^{\pm.021}$	$3.977^{\pm.024}$	$0.216^{\pm.005}$	$0.363^{\pm.006}$	$0.489^{\pm.002}$
	Ours	$\mathbf{0.389}^{\pm.006}$	$\mathbf{0.575}^{\pm.009}$	$\mathbf{0.692}^{\pm.005}$	$\mathbf{1.335}^{\pm.003}$	$\mathbf{2.856}^{\pm.009}$	$\mathbf{0.497}^{\pm.006}$	$\mathbf{0.691}^{\pm.004}$	$\mathbf{0.753}^{\pm.015}$

Table 2: Quantitative comparison on Phoenix-2014 dataset, where **Real Motion** and **Ours** are evaluated by extracting the 3D keypoints from our mesh representation. The J_{PT} and J_{ours} in the bracket represent being trained on the corresponding keypoints. Furthermore, we also qualitative comparison results in Fig. 8. Please see more visualizations in our supplementary video, and project page.

Discussion. With SignAvatars, our goal is to provide an up-to-date, publicly available 3D holistic mesh **motion-based** SLP benchmark and we invite the community to participate. As an alternative for the re-evaluation, we can also develop a brand new 3D sign language translation (SLT) method to **re**-evaluate PT and compare it with our method on BLEU and ROUGE. As a part of our future work on SL understanding, we also encourage the SL community to develop back-translation and mesh-based SLT methods trained with our benchmark. We believe that the 3D holistic mesh representation presents significant improvements for the accurate SL-motion correlation understanding, compared to the pure 2D methods as shown in Tab. 4 and Tab. 5 of the main paper, which was also proved to be true in a latest 3D SLT work [17].

D Implementation details for experiments and evaluation

Optimization strategy of automatic annotation pipeline. During optimization, we utilize an iterative five-stage fitting procedure to minimize the objective function and use Adam optimizer with 1e-2 as the learning rate. Moreover, a good initialization can significantly boost the fitting speed of our annotation pipeline. At the same time, a well-pixel-aligned body pose will also help the reconstruction of hand meshes. Motivated by this, we apply 2000 fitting steps for a clip and split the fitting steps into five stages with 400 steps in each stage to formulate our iterative fitting pipeline. In the meantime, the Limited-memory BFGS [22] with a strong Wolfe line is applied to our optimization. In the first three stages, all the loss and parameters are optimized together. The weights $w_{body} = w_{hand} = 1$ are applied for L_J to obtain a good body pose estimation. In the last two stages, we will first extract a mean pose from the record of the previous optimization to gain a stable body shape and freeze it as a fixed shape, as the signer will not change in a video by default. Subsequently, to obtain accurate and detailed hand meshes, we will enlarge the w_{hand} to 2 to reach the final holistic mesh reconstruction with a natural and accurate hand pose.

D.1 Evaluation Protocols

In this subsection, we will elaborate on the computational details of our used evaluation protocol. To start with, our evaluation relies on a text-motion embedding model following prior arts [17, 30, 33]. For simplicity, we use the same symbols and notations as in our Sec. 3 and Sec. 4 of the main paper. Through the GRU embedding layer, we embed our motion representation $M_{1:T}$ and linguistic feature $E^l_{1:s}$ into $f_m \in R^d$ and $f_l \in R^d$ with the same dimensions to apply contrastive loss and minimize the feature distances, where $d = 512$ is used in our experiments. After motion and prompt feature extraction, we compute each of the evaluation metrics, which are summarized below:

- **Frechet Inception Distance (FID)** (↓), the distributional distance between the generated motion and the corresponding real motion based on the extracted motion feature.
- **Diversity**, the average Euclidean distance in between the motion features of $N_D = 300$ randomly sampled motion pairs.
- **R-precision** (↑), the average accuracy at top-k positions of sorted Euclidean distances between the motion embedding and each GT prompt embedding.
- **Multimodality**, average Euclidean distance between the motion feature of $N_m = 10$ pairs of motion generated with the same single input prompt.
- **Multimodal Distance (MM-Dist)** (↓), average Euclidean distance between each generated motion feature and its input prompt feature.
- **MR-precision** (↓), the average accuracy at top-k positions of sorted Euclidean distance between a generated motion feature and 16 motion samples from dataset (1 positive + 15 negative).

We now provide further details in each of those. For simplicity, we denote the dataset length as N below.

Frechet Inception Distance (FID) is used to evaluate the distribution distance between the generated motion and the corresponding real motion:

$$FID = \|\mu_{gt} - \mu_{pred}\|_2 - Tr(C_{gt} + C_{pred} - 2(C_{gt}C_{pred})^{1/2}) \tag{1}$$

where μ_{gt}, μ_{pred} are the mean values for the features of real motion and generated motion, separately. C, Tr are the covariance matrix and trace of a matrix.

Diversity is used for evaluating the variance of the generated SL motion. Specifically, we randomly sample $N_D = 300$ motion feature pairs $\{f_m, f_m^{'}\}$ and compute the average Euclidean distance between them by:

$$Diversity = \frac{1}{N_D} \sum_{i}^{N_D} \|f_m^i - f_m^{i'}\| \tag{2}$$

Multimodality is leveraged to measure the diversity of the SL motion generated from the same prompts. Specifically, we compute the average Euclidean distance between the extracted motion feature of $N_m = 10$ pairs $\{f_m^j, f_m^{j'}\}$ of motion generated with the same single input prompt. Through the full dataset, it can

be written as:

$$Multimodality = \frac{1}{NN_m} \sum_{i}^{N} \sum_{j}^{N_M} \| f_m^{ij} - f_m^{ij'} \| \tag{3}$$

Multimodal Distance (MM-Dist) is applied to evaluate the text-motion correspondency. Specifically, it computes the average Euclidean distance between each generated motion feature and its input prompt feature:

$$MM\text{-}Dist = \frac{1}{N} \sum_{i}^{N} \| f_m^i - f_l^i \| \tag{4}$$

E Discussion

E.1 Related Work

In this section, we present more details about the related work as well as the open problems.

Background. Existing SL datasets, and dictionaries are typically limited to 2D, which is ambiguous and insufficient for learners as introduced in [17], different signs could appear to be the same in 2D domain due to depth ambiguity. In that, 3D avatars and dictionaries are highly desired for efficient learning [21], teaching, and many downstream tasks. However, 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]. As a result, there is not a unified large-scale multi-prompt 3D sign language holistic motion dataset with precise hand mesh annotations. The lack of such 3D avatar data is a huge barrier to bringing these meaningful applications to Deaf community, such as 3D sign language production (SLP), 3D sign language recognition (SLR), and many downstream tasks such as digital simultaneous translators between spoken language and sign language in VR/AR.

Open problems. Overall, the open problems chain is: **1)** Current 3D avatar annotation methods for sign language are mostly done manually by SL experts and are labor-intensive. **2)** Lack of generic automatic 3D expressive avatar annotation methods with detailed hand pose. **3)** Due to the lack of a generic annotation method, there is also a lack of a unified large-scale multi-prompt 3D co-articulated continuous sign language holistic motion dataset with precise hand mesh annotations. **4)** Due to the above constraints, it is difficult to extend sign language applications to highly desired 3D properties such as 3D SLR, 3D SLP, which can be used for many downstream applications like virtual simultaneous SL translators, 3D dictionaries, etc.

According to the problem chain, we will introduce the SoTA from three aspects: 3D holistic mesh annotation pipeline, 3D sign language motion dataset, and 3D SL applications.

3D holistic mesh annotation: There are a lot of prior works for reconstructing holistic human body from RGB images with parametric models like SMPL-X [23], Adam [15]. Among them, TalkSHOW [32] proposes a fitting pipeline based on SMPLify-X [23] with a photometric loss for facial details. OSX [19] proposes a time-consuming finetune-based weakly supervision pipeline to generate pseudo-3D holistic annotations. However, such expressive parametric models have rarely been applied to the SL domain. [16] use off-the-shelf methods to estimate holistic 3D mesh on the GSLL sign-language dataset [31]. In addition to that, only a concurrent work [10] can reconstruct 3D holistic mesh annotation using linguistic priors with group labels obtained from a sign-classifier trained on Corpus-based Dictionary of Polish Sign Language (CDPSL) [20], which is annotated with HamNoSys As such, it utilizes an existing sentence segmentation methods [25] to generalize to multiple-sign videos. These methods cannot deal with the challenging self-occlusion, hand–hand and hand–body interactions which makes them insufficient for complex interacting hand scenarios such as sign language. There is not yet a generic annotation pipeline that is sufficient to deal with complex interacting hand cases in **continuous and co-articulated** SL videos.

Sign language datasets. While there have been many well-organized continuous SL motion datasets [1, 2, 6, 7, 13, 14] with 2D videos or 2D keypoints annotations, the only existing 3D SL motion dataset with 3D holistic mesh annotation is in [10], which is purely **isolated sign based** and not sufficient for tackling real-world applications in natural language scenarios. There is not yet a unified large-scale **multi-prompt 3D** SL holistic motion dataset with **continuous and co-articulated** signs and precise hand mesh annotations.

SL applications. Regarding the SL applications, especially sign language production (SLP), [4] can generate 2D motion sequences from HamNoSys. [27] and [29] are able to generate 3D keypoint sequences with glosses. The avatar approaches are often hand-crafted and produce robotic and unnatural movements. Apart from them, there are also early avatar approaches [5, 8, 9, 11, 34] with a pre-defined protocol and character.

E.2 Licensing

Our dataset will first be released under the CC BY-NC-SA (Attribution-NonCommercial-Share-Alike) license for research purposes. Specifically, we will release the SMPL-X/MANO annotation and provide the instruction to extract the data instead of distributing the raw videos. We also elaborate on the license of the data source we used in our dataset collection:

How2Sign [7]. Creative Commons Attribution-NonCommercial 4.0 International License.

DGS Corpus [24]. is under CC BY-NC license.

Dicta-Sign. is under CC-BY-NC-ND 4.0 license.

WLASL [18]. Computational Use of Data Agreement (C-UDA-1.0).

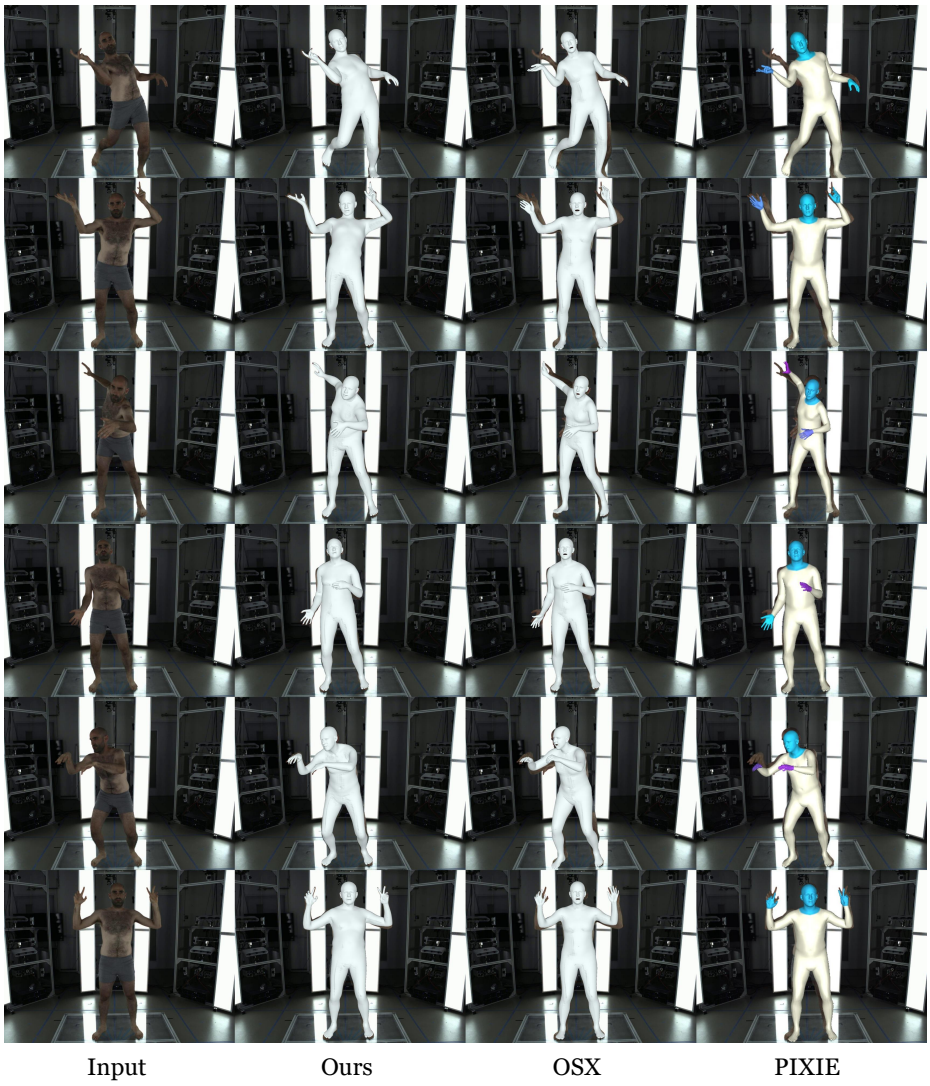

Fig. 5: Comparisons of existing 3D holistic human mesh reconstruction methods on EHF dataset. Our annotation method produces significantly better holistic reconstructions with plausible poses, as well as the best pixel alignment. (Zoom in for a better view)

Fig. 6: Our 3D holistic human mesh reconstruction methods on in-the-wild cases. (Zoom in for a better view)

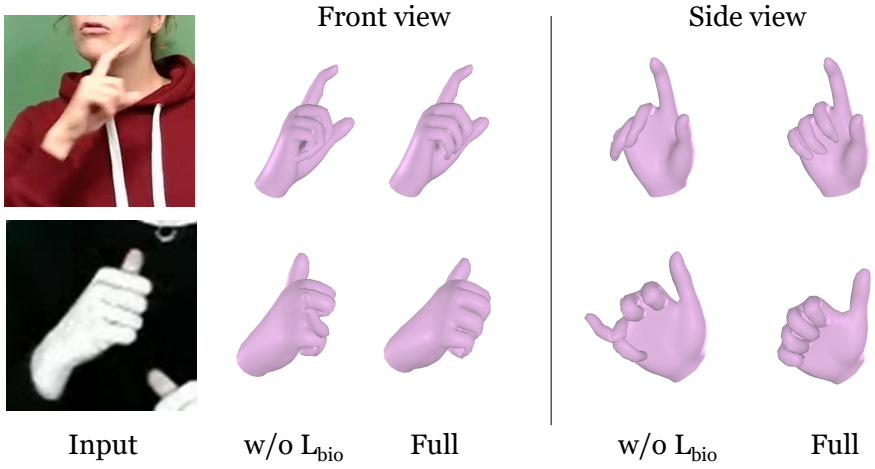

Fig. 7: Visualization examples and analysis of our regularization term. The biomechanical constraints can alleviate the implausible poses caused by monocular depth ambiguity, which happens occasionally in complex interacting-hands scenarios for other monocular capture methods.

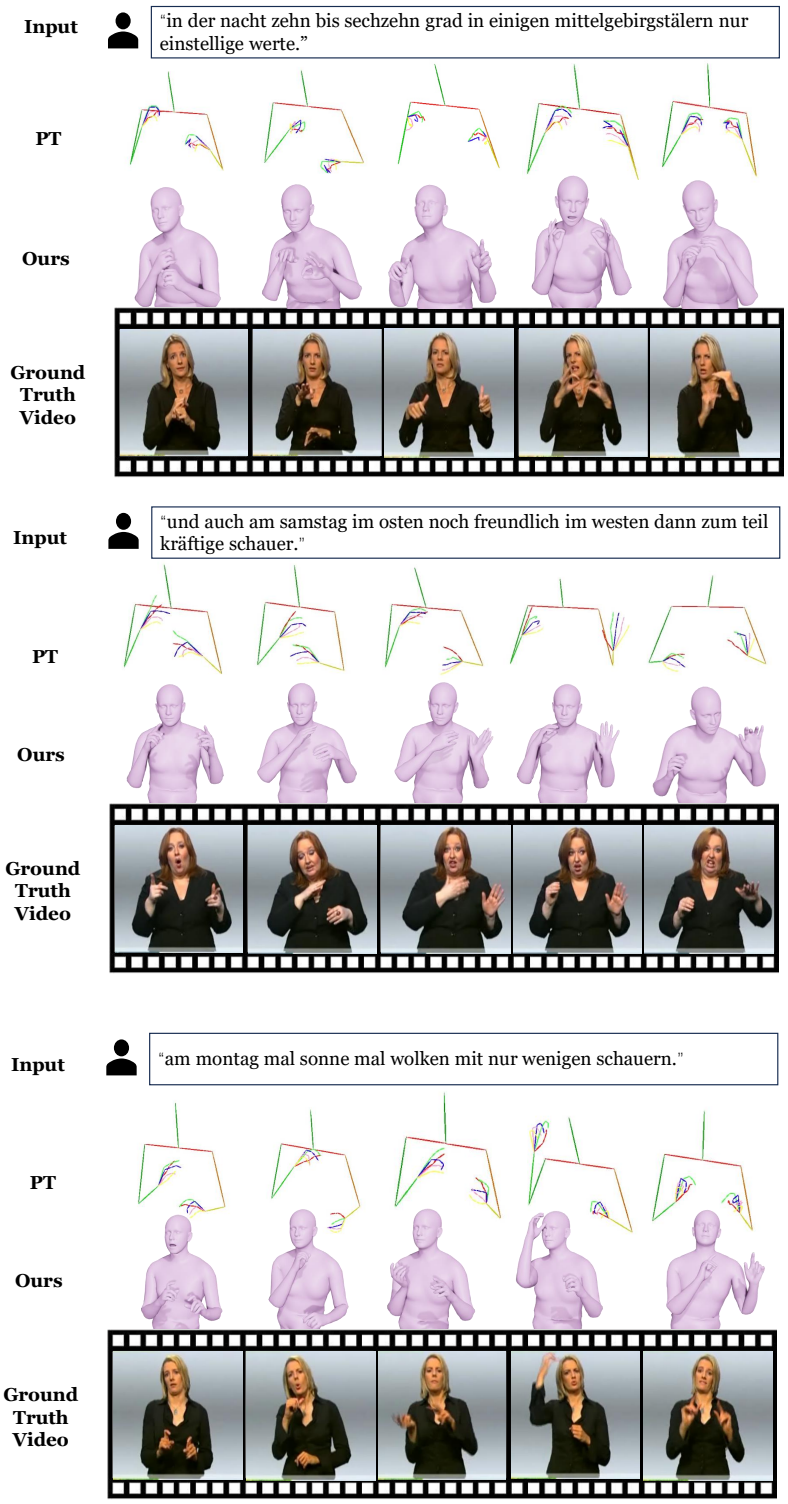

Fig. 8: Qualitative comparison with PT [27] on Phoenix-2014 T dataset.

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) 9
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) 9
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) 8
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) 9
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) 9
6. 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) 3, 9
7. 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) 9
8. 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) 9
9. 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) 9
10. 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) 2, 9
11. 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) 9
12. Guo, C., Zou, S., Zuo, X., Wang, S., Ji, W., Li, X., Cheng, L.: Generating diverse and natural 3d human motions from text. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5152–5161 (2022) 4
13. 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) 9
14. 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) 9

15. 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) 9
16. 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) 9
17. 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) 6, 7, 8
18. 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) 9
19. 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) 9
20. Linde-Usiekniewicz, J., Czajkowska-Kisil, M., Łacheta, J., Rutkowski, P.: A corpus-based dictionary of polish sign language (pjm) 9
21. 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) 8
22. Nocedal, J., Wright, S.J.: Numerical optimization. Springer (1999) 6
23. 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) 1, 9
24. 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) 9
25. 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) 9
26. Saunders, B., Camgoz, N.C., Bowden, R.: Adversarial training for multi-channel sign language production. arXiv preprint arXiv:2008.12405 (2020) 3
27. 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) 3, 4, 9, 12
28. Saunders, B., Camgoz, N.C., Bowden, R.: Continuous 3d multi-channel sign language production via progressive transformers and mixture density networks. International journal of computer vision **129**(7), 2113–2135 (2021) 3
29. 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) 3, 9
30. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-Or, D., Bermano, A.H.: Human motion diffusion model. arXiv preprint arXiv:2209.14916 (2022) 7
31. 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) 9

32. 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) 9
33. 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) 7
34. 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) 9