Signs as Tokens: A Retrieval-Enhanced Multilingual Sign Language Generator
Supplementary Materials

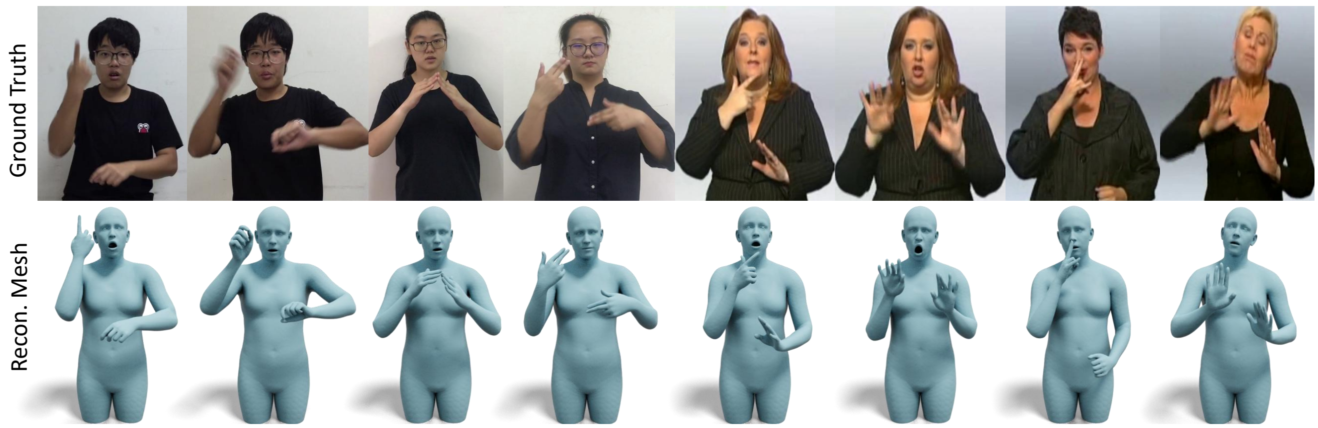

Figure S1. Qualitative comparisons between ground truth video frames and reconstructed meshes obtained from the proposed SMPL-X pose fitting pipeline on CSL-Daily (left) and Phoenix-2014T (right). Zoom-in for hand details.

A. Curating SMPL-X Poses

To curate a high-fidelity dataset with accurate 3D annotations, we rely on state-of-the-art performing methods for 3D hand [13] and body reconstruction [10]. Specifically, given a 2D video of a signer, we first detect the number of identities in the video using an off-the-shelf detector [11] and retain the most confident detection box. Following that we feed the tight human crop to OSX [10] to extract a rough human body pose estimation. Given that OSX often fails to accurately capture the arm positions and the hand poses, we follow a two-step approach that accurately refines the human pose. To accurately reconstruct the fine details of the hand poses, we utilize WiLoR [13], a state-of-the-art 3D reconstruction pipeline that can detect and reconstruct challenging hand poses with high fidelity. We acquire the hand poses of WiLoR along with the global orientation of the hand and directly substitute the hand parameters derived from OSX. In the second state, we employ Mediapipe body pose estimation [11] to extract 2D joint location $\mathbf{J}^{2D}$ for the shoulders and the arms. Using the derived joint locations, we employ an optimization scheme that refines the OSX poses of the upper body, while keeping the hand poses and orientation fixed:

$$\mathcal{L}_{rec} = ||\mathbf{J}^{2D} - \Pi_K(\hat{\mathbf{J}}^{3D})||_1, \tag{S1}$$

where $\hat{\mathbf{J}}^{3D}$ are the predicted 3D joints and Π_K is the weak-perspective projection. To further constrain the temporal coherence of the reconstructions, we include an additional temporal loss $\mathcal{L}_{temp}$:

$$\mathcal{L}_{temp} = ||\mathbf{X}_f - \mathbf{X}_{f-1}||_2 + ||\mathbf{J}_f - \mathbf{J}_{f-1}||_2, \tag{S2}$$

where $\mathbf{X}_f$ denotes the 3D mesh in frame f. Finally, to penalize irregular poses, we include a pose regularization:

$$\mathcal{L}_{reg} = ||\theta||_2 \tag{S3}$$

that constrains irregular upper body poses.

Since neither CSL-Daily [16] nor Phoenix-2014T [3] provides 3D annotations, we perform qualitative evaluations, as illustrated in Figure S1. The results clearly demonstrate that the proposed pose fitting pipeline can accurately reconstruct 3D hands and is robust across various handshapes. To quantitatively assess the pipeline, we further apply it to the SGNify mocap dataset [6], which includes 57 signs with annotated meshes. The results presented in Table S1 indicate that our method achieves the lowest hand reconstruction errors and comparable body errors to the previous best method [2], establishing our approach as a powerful tool for curating more sign language datasets in the future.

Figure S2. Qualitative comparisons of generated signs between our proposed method, SOKE, with the SOTA method, S-MotionGPT [7], on the test sets of How2Sign (left), CSL-Daily (middle), and Phoenix-2014T (right).

Method	Body↓	Left Hand↓	Right Hand↓
FrankMoCap [14]	78.07	20.47	19.62
PIXIE [5]	60.11	25.02	22.42
PyMAF-X [15]	68.61	21.46	19.19
SMPLify-X [12]	56.07	22.23	18.83
SGNify [6]	55.63	19.22	17.50
OSX [10]	47.32	18.34	18.12
NSA [2]	**46.42**	16.17	15.23
Ours	46.73	**10.55**	**8.94**

Table S1. Reconstruction errors on SGNify mocap dataset [6]. We report mean per vertex errors in mm.

B. Additional Qualitative Results

Please refer to our project page for video demonstrations of generated signs. These demos include ground truth sign videos, as well as generations from the SOTA method, S-MotionGPT [7], and our proposed SOKE. Additionally, we provide several qualitative results to showcase the generated signs (Figure S2) and highlight the effectiveness of our retrieval-enhanced SLG approach (Figure S3).

C. Additional Quantitative Results

Codebook Size. We perform a hyper-parameter analysis on the codebook sizes for the body (N_Z^B) and hands (N_Z^{LH}, N_Z^{RH}) in our decoupled tokenizer. As shown in

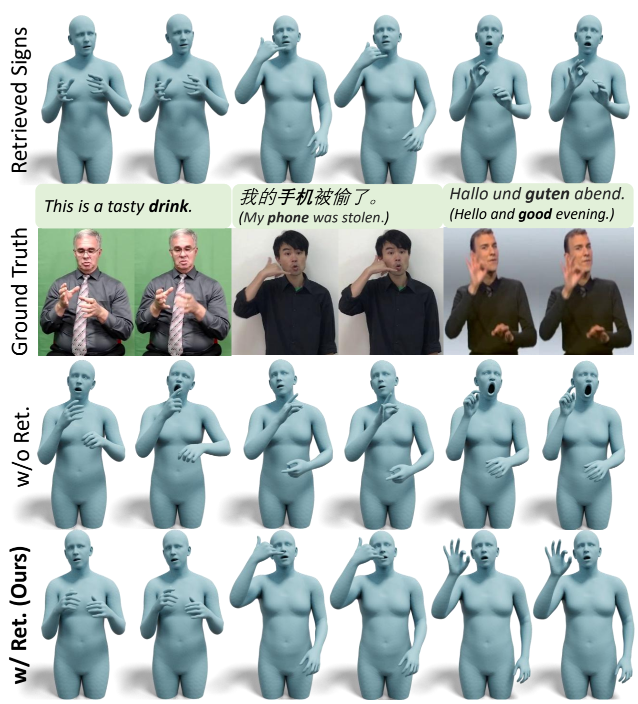

Figure S3. Qualitative ablation study for retrieval-enhanced SLG. (Left: How2Sign; Middle: CSL-Daily; Right: Phoenix-2014T.)

Table S2, we find that using either larger or smaller codebooks results in degraded reconstruction performance. Our

N_Z^B	$N_Z^{LH} = N_Z^{RH}$	H2S (JPE↓)		CSL (JPE↓)		Ph-T (JPE↓)	
		Body	Hand	Body	Hand	Body	Hand
96	128	19.37	7.07	23.52	5.80	25.79	7.35
96	256	19.37	6.86	23.52	5.52	25.79	7.11
64	192	20.04	6.65	24.13	5.13	26.02	6.78
128	192	19.95	6.65	23.91	5.13	26.27	6.78
96	192	**19.37**	**6.65**	**23.52**	**5.13**	**25.79**	**6.78**

Table S2. Study on the codebook sizes for the body (N_Z^B) and hands (N_Z^{LH}, N_Z^{RH}). We use procrustes-aligned mean per joint position error (PA-MPJPE) to assess the reconstruction performance of the decoupled tokenizer.

λ	H2S (DTW↓)		CSL (DTW↓)		Ph-T (DTW↓)	
	Body	Hand	Body	Hand	Body	Hand
0.1	7.95	2.82	7.46	2.13	5.47	2.04
0.2	7.28	2.76	6.91	1.95	5.08	1.68
1/3	**6.82**	**2.35**	**6.24**	**1.71**	**4.77**	**1.38**
0.4	7.34	2.62	7.11	1.91	6.39	1.96

Table S3. Study on the impact of λ, a hyper-parameter used for fusing part-wise token embeddings in our multi-head decoding method.

Method	Multi ling.	H2S (DTW↓)			CSL (DTW↓)			Phoenix (DTW↓)		
		Avg	Body	Hand	Avg	Body	Hand	Avg	Body	Hand
S-MotionGPT	×	5.91	11.23	4.39	5.34	10.81	3.78	4.75	9.45	3.41
Ours	×	4.14	7.92	3.07	4.18	8.18	3.04	3.83	7.25	2.85
Ours	✓	**3.34**	**6.82**	**2.35**	**2.72**	**6.24**	**1.71**	**2.13**	**4.77**	**1.38**

Table S4. Performance of our method on monolingual datasets.

default configuration ($N_Z^B = 96$, $N_Z^{LH} = N_Z^{RH} = 192$) delivers the best performance among all settings.

Impact of λ on SLG. In our multi-head decoding method, we introduce a hyper-parameter, λ, to control the weight of hand tokens during embedding fusion. The results in Table S3 demonstrate that $\lambda = 1/3$, *i.e.*, assigning equal weights to the body and hands, yields the best performance. This further underscores the importance of each body part in conveying the semantics of sign languages.

Monolingual Performance. As shown in Table S4, our method still outperforms the SOTA method, S-MotionGPT, when training on monolingual SL datasets, while the best results are achieved by the multilingual version of our method.

D. Illustration of Decoupled Tokenizer

As shown in Figure S4, we provide an illustration of our decoupled tokenizer for better understanding. It utilizes three VQ-VAEs to model the key regions of a signer: the upper body, left hand, and right hand.

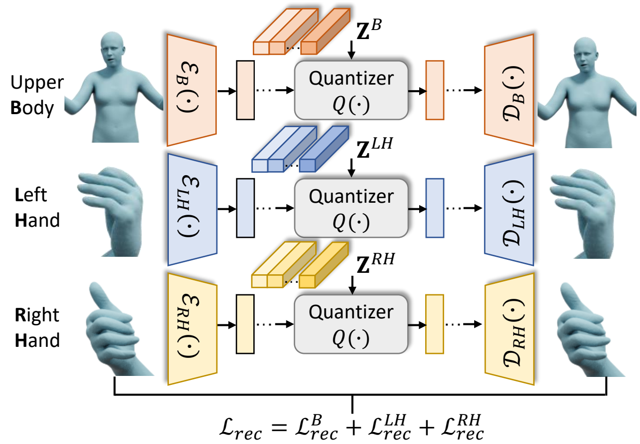

Figure S4. Workflow of our decoupled tokenizer. It is composed of three parallel VQ-VAEs, each dedicated to generating motion tokens for a different part of the signer's body: the upper body, left hand, and right hand.

E. Discussion

Broader Impacts. Sign language is the primary mode of communication for the deaf communities. Due to significant grammatical differences from spoken languages, a notable communication gap exists between the deaf and hearing individuals. In this work, we propose an autoregressive sign language model, which is capable of generating multilingual sign language avatars from text inputs within a single unified framework. Extensive quantitative and qualitative results suggest the potential of our method to form a practical deaf-hearing communication system.

Limitations. Our method employs 3D avatars to represent signers, enabling high-fidelity motion representations. However, there is a lack of 3D annotations in existing sign language datasets. While our proposed SMPL-X pose fitting pipeline can accurately reconstruct 3D meshes from 2D keypoints, some reconstruction errors are inevitable. In the future, the release of more sign language datasets with annotated meshes is anticipated, which could significantly enhance avatar-based sign language generation models.

Future Works. We have validated the proposed multilingual sign language generator on three widely-adopted sign languages, Chinese, American, and German sign language [4, 9, 16]. As the scalability of our approach has been demonstrated in Table 3 of the main paper, in the future, we plan to extend our method to support more sign languages, such as British Sign Language [1] and Indian Sign Language [8].

References

[1] 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 *ECCV*, pages 35–53, 2020. 3

[2] Vasileios Baltatzis, Rolandos Alexandros Potamias, Evangelos Ververas, Guanxiong Sun, Jiankang Deng, and Stefanos Zafeiriou. Neural sign actors: A diffusion model for 3d sign language production from text. In *CVPR*, pages 1985–1995, 2024. 1, 2

[3] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Neural sign language translation. In *CVPR*, 2018. 1

[4] 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 *CVPR*, pages 2735–2744, 2021. 3

[5] Yao Feng, Vasileios Choutas, Timo Bolkart, Dimitrios Tzionas, and Michael J Black. Collaborative regression of expressive bodies using moderation. In *3DV*, pages 792–804, 2021. 2

[6] Maria-Paola Forte, Peter Kulits, Chun-Hao P Huang, Vasileios Choutas, Dimitrios Tzionas, Katherine J Kuchenbecker, and Michael J Black. Reconstructing signing avatars from video using linguistic priors. In *CVPR*, pages 12791–12801, 2023. 1, 2

[7] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *NeurIPS*, 36:20067–20079, 2023. 2

[8] Abhinav Joshi, Romit Mohanty, Mounika Kanakanti, Andesha Mangla, Sudeep Choudhary, Monali Barbate, and Ashutosh Modi. isign: A benchmark for indian sign language processing. In *Findings of ACL*, pages 10827–10844, 2024. 3

[9] Oscar Koller, Jens Forster, and Hermann Ney. Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *CVIU*, 141:108–125, 2015. 3

[10] Jing Lin, Ailing Zeng, Haoqian Wang, Lei Zhang, and Yu Li. One-stage 3d whole-body mesh recovery with component aware transformer. In *CVPR*, pages 21159–21168, 2023. 1, 2

[11] 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. 1

[12] 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 *CVPR*, pages 10975–10985, 2019. 2

[13] Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, and Stefanos Zafeiriou. Wilor: End-to-end 3d hand localization and reconstruction in-the-wild. In *CVPR*, 2025. 1

[14] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration. In *ICCVW*, pages 1749–1759, 2021. 2

[15] 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 TPAMI*, 45(10):12287–12303, 2023. 2

[16] Hao Zhou, Wengang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li. Improving sign language translation with monolingual data by sign back-translation. In *CVPR*, 2021. 1, 3