Tencent AI Lab

IMPERIAL

ECCV
EUROPEAN CONFERENCE ON COMPUTER VISION
MILANO 2024

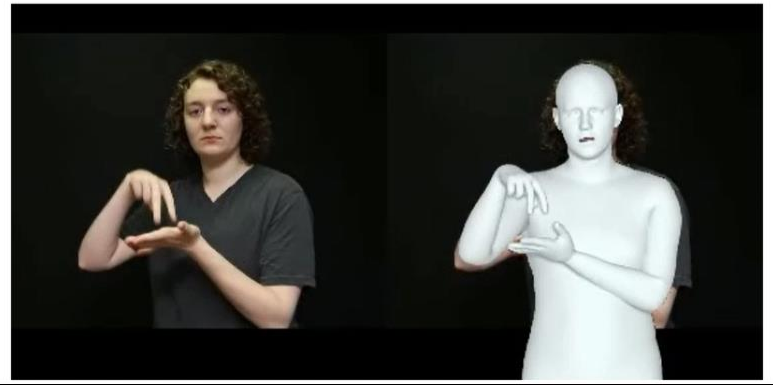
# SignAvatars: A Large-scale 3D Sign Language Holistic Motion Dataset and Benchmark

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

[1]Imperial College London, [2]Tencent AI Lab


# Co-articulated signs

# isolated signs






# Modality

- SignAvatars is the first large-scale 3D sign language holistic motion dataset with SMPL-X [1] and MANO [2] annotations.

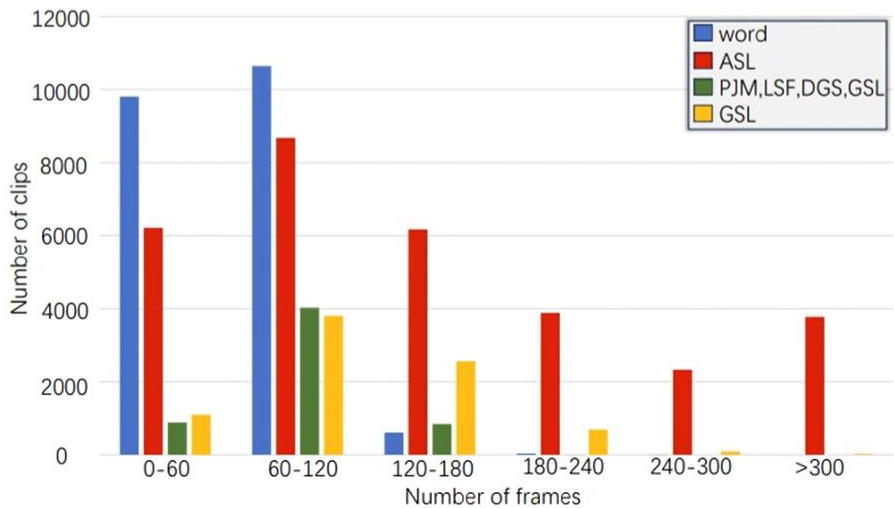
| Data | Video | Frame | Duration (hours) | co-articulated | Pose Annotation | Signer |
|---|---|---|---|---|---|---|
| RWTH-Phoenix-2014T (Camgoz et al., 2018) | 8.25K | 0.94M | 11 | C | - | 9 |
| DGS Corpus (Hanke et al., 2020) | - | - | 50 | C | 2D keypoints | 327 |
| BSL Corpus (Schembri et al., 2013) | - | - | 125 | C | - | 249 |
| MS-ASL (Joze & Koller, 2018) | 25K | - | 25 | I | - | 222 |
| WL-ASL (Li et al., 2020) | 21K | 1.39M | 14 | I | 2D keypoints | 119 |
| How2Sign (Duarte et al., 2021) | 34K | 5.7M | 79 | C | 2D keypoints, depth* | 11 |
| CSL-Daily (Huang et al., 2018) | 21K | - | 23 | C | 2D keypoints, depth | 10 |
| SIGNUM (Von Agris et al., 2008) | 33K | - | 55 | C | - | 25 |
| AUTSL (Sincan & Keles, 2020) | 38K | - | 21 | I | depth | 43 |
| Forte et al. (2023) | 0.05K | 4K | - | I | body mesh vertices | - |
| SignAvatars (Ours) | 70K | 8.34M | 117 | Both | SMPL-X, MANO, 2D&3D keypoints | 153 |

[1] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. Osman, D. Tzionas, and M. J. Black. Expressive body capture: 3d hands, face, and body from a single image. CVPR, 2019
[2] J. Romero, D. Tzionas, and M. J. Black. Embodied hands: Modeling and capturing hands and bodies together. SIGGRAPH Asia, 2017

# Data distribution

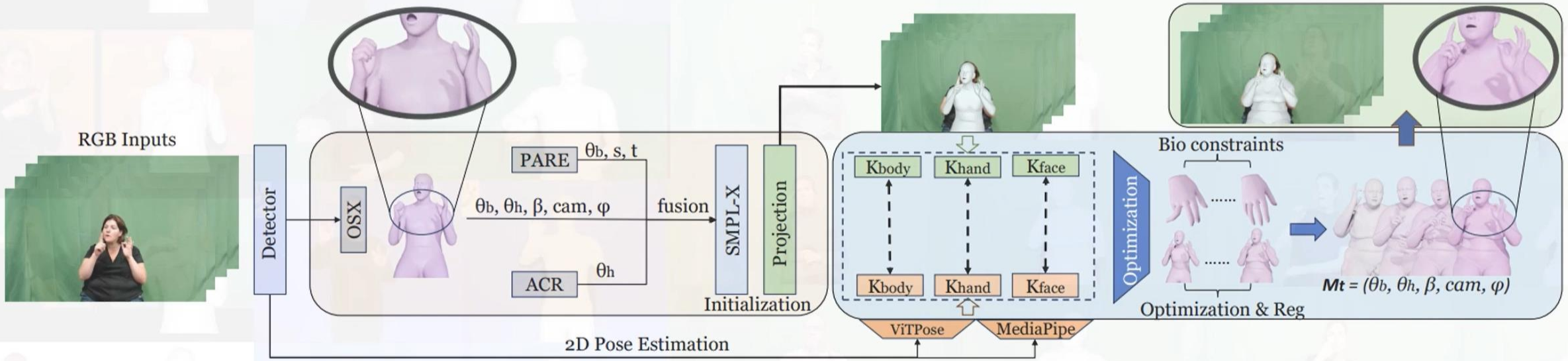
➢Our dataset consists of multiple subsets with different SL annotations.

- HamNoSys (**H**)
- Word (**W**)
- Spoken Language (**S**)
- Gloss (**G**)




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

# Data annotation pipeline





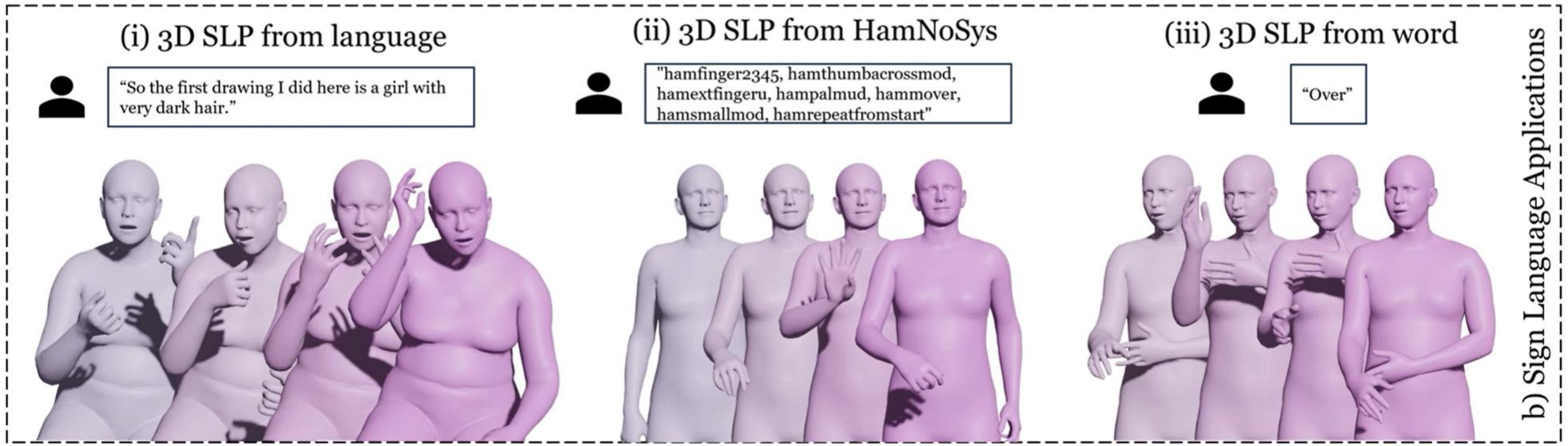
[3] J. Lin, A. Zeng, H. Wang, L. Zhang, Y. Li. OSX: One-Stage 3D Whole-Body Mesh Recovery with Component Aware Transformer. CVPR, 2023
[4] Z. Yu, S. Huang, C. Fang, T. P. Breckon, J. Wang. ACR: Attention Collaboration-based Regressor for Arbitrary Two-Hand Reconstruction. CVPR, 2023
[5] M. Kocabas, C. P. Huang, O. Hilliges M. J. Black. PARE: Part Attention Regressor for 3D Human Body Estimation. ICCV, 2021

# Fitting examples

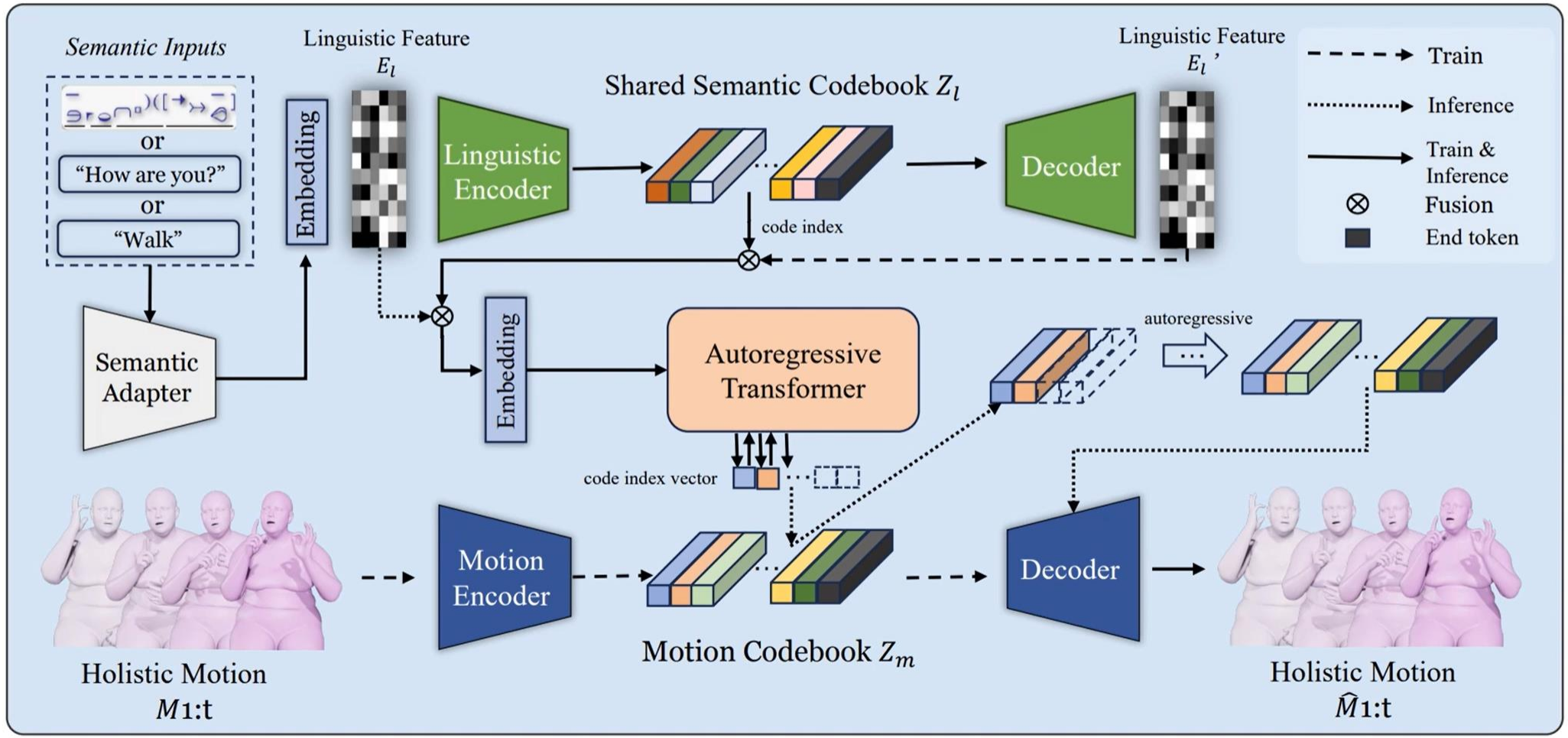

[6] H. Zhang, Y. Tian, Y. Zhang, M. Li, L. An, Z. Sun, and Y. Liu. Pymaf-x: Towards well-aligned full-body model regression from monocular images. T-PAMI, 2022.
[7] Y. Feng, V. Choutas, T. Bolkart, D. Tzionas, and M. J. Black. PIXIE: Collaborative regression of expressive bodies using moderation. 3DV, 2021

# Application

- High-quality sign language applications, such as Sign Language Production (SLP) and Sign Language Translation (SLT). We also support SLP from various prompts.

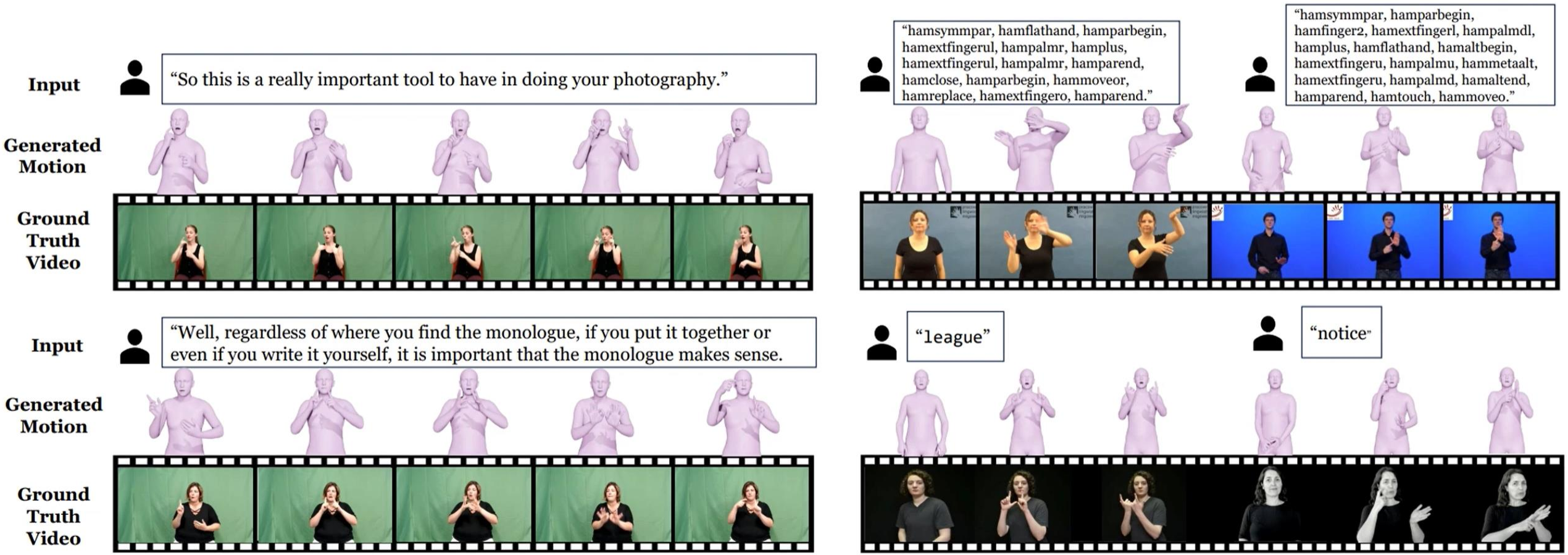


# Application: SignVAE




[8] A. Van Den Oord, O. Vinyals, et al. Neural discrete representation learning. NeurIPS, 2017.

# Generation examples



"hamsymmlr,hamfinger23,hamthumbacrossmod,hamextfingerul,hampalmud,hammovel,hamarcd,hamfingerpad,hamtouch,hamrepeatfromstart"

Our Generation Results and Character Driven


Ground Truth Video

* The character is modified based on https://www.aplaybox.com/details/model/gN7MAyKwiYoS

"about"

Our Generation Results and Character Driven

Ground Truth Video

"bless"

Our Generation Results and Character Driven


Ground Truth Video

"complex"

Our Generation Results and Character Driven

Ground Truth Video

"notice"

Our Generation Results and Character Driven

Ground Truth Video

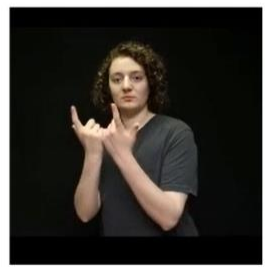
"every"

Our Generation Results and Character Driven

Ground Truth Video

"league"

Our Generation Results and Character Driven



Ground Truth Video

* The characters are obtained from TADA [9].

[9] T. Liao, H. Yi, Y. Xiu, J. Tang, Y. Huang, J. Thies, M. J. Black. TADA! Text to Animatable Digital Avatars. 3DV, 2024

“If you don't go to the far corner, it is going to bow your ceil.

Our Generation Results and Character Driven

Ground Truth Video

* The character is obtained from TADA [9].

[9] T. Liao, H. Yi, Y. Xiu, J. Tang, Y. Huang, J. Thies, M. J. Black. TADA! Text to Animatable Digital Avatars. 3DV, 2024

"And other optional equipment is your chest plate for a man."

Our Generation Results and Character Driven

Ground Truth Video

* The character is obtained from TADA [9].

[9] T. Liao, H. Yi, Y. Xiu, J. Tang, Y. Huang, J. Thies, M. J. Black. TADA! Text to Animatable Digital Avatars. 3DV, 2024

# Experimental evaluation

| Data Type | | R-Precision (↑) | | | FID (↓) | Diversity (→) | MM (→) | MM-dist (↓) | MR-Precision (↑) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | top 1 | top 3 | top 5 | | | | | top 1 | top 3 | top 5 |
| **Real motion** | Language | $0.398^{\pm.005}$ | $0.612^{\pm.007}$ | $0.709^{\pm.008}$ | $0.017^{\pm.153}$ | $9.565^{\pm.075}$ | - | $2.655^{\pm.057}$ | - | - | - |
| | HamNoSys | $0.455^{\pm.002}$ | $0.689^{\pm.006}$ | $0.795^{\pm.004}$ | $0.007^{\pm.062}$ | $8.754^{\pm.028}$ | - | $2.113^{\pm.023}$ | - | - | - |
| | Word-300 | $0.499^{\pm.003}$ | $0.811^{\pm.002}$ | $0.865^{\pm.003}$ | $0.006^{\pm.054}$ | $8.656^{\pm.035}$ | - | $1.855^{\pm.019}$ | - | - | - |
| Holistic | Language | $0.375^{\pm.007}$ | $0.535^{\pm.008}$ | $0.661^{\pm.0059}$ | $1.238^{\pm.389}$ | $11.56^{\pm.101}$ | $1.129^{\pm.107}$ | $3.156^{\pm.067}$ | $0.441^{\pm.007}$ | $0.675^{\pm.007}$ | $0.731^{\pm.009}$ |
| | HamNoSys | $0.429^{\pm.004}$ | $0.657^{\pm.005}$ | $0.756^{\pm.002}$ | $0.884^{\pm.035}$ | $9.451^{\pm.087}$ | $0.941^{\pm.056}$ | $2.651^{\pm.027}$ | $0.552^{\pm.002}$ | $0.745^{\pm.010}$ | $0.813^{\pm.034}$ |
| | Word-300 | $0.475^{\pm.002}$ | $0.731^{\pm.003}$ | $0.815^{\pm.005}$ | $0.756^{\pm.021}$ | $8.956^{\pm.091}$ | $0.815^{\pm.059}$ | $2.101^{\pm.024}$ | $0.615^{\pm.005}$ | $0.797^{\pm.006}$ | $0.875^{\pm.002}$ |
| Gesture | Language | $0.341^{\pm.008}$ | $0.491^{\pm.009}$ | $0.671^{\pm.010}$ | $0.975^{\pm.315}$ | $10.08^{\pm.121}$ | $1.156^{\pm.135}$ | $3.491^{\pm.089}$ | $0.475^{\pm.011}$ | $0.689^{\pm003}$ | $0.751^{\pm.004}$ |
| | HamNoSys | $0.435^{\pm.005}$ | $0.649^{\pm.004}$ | $0.745^{\pm.006}$ | $0.851^{\pm.033}$ | $8.944^{\pm.097}$ | $0.913^{\pm.036}$ | $2.876^{\pm.015}$ | $0.581^{\pm.004}$ | $0.736^{\pm.006}$ | $0.825^{\pm.008}$ |
| | Word-300 | $0.465^{\pm.001}$ | $0.711^{\pm.003}$ | $0.818^{\pm.003}$ | $0.715^{\pm.016}$ | $8.235^{\pm.055}$ | $0.801^{\pm.021}$ | $2.339^{\pm.027}$ | $0.593^{\pm.006}$ | $0.814^{\pm.005}$ | $0.901^{\pm.006}$ |

Table 3: Quantitative evaluation results for the 3D holistic SL motion generation. *Real motion* is the sampled motions from the original holistic motion annotation in the dataset. *Holistic* represents the results for generated holistic motion. *Gesture* stands for the evaluation conducted on two arms.

| Method | DTW-MJE Rank (↑) | | |
|---|---|---|---|
| | top 1 | top 3 | top 5 |
| Ham2Pose*[10] | $0.092^{\pm.031}$ | $0.197^{\pm.029}$ | $0.354^{\pm.032}$ |
| Ham2Pose-3d | $0.253^{\pm.036}$ | $0.369^{\pm.039}$ | $0.511^{\pm.035}$ |
| SignVAE (Ours) | $0.516^{\pm.039}$ | $0.694^{\pm.041}$ | $0.786^{\pm.035}$ |

Table 4: Comparison with state-of-the-art SLP method from HamNoSys. * represents using only 2D information.

| Method | R-Precision (↑) | | | MM-dist (↓) |
|---|---|---|---|---|
| | top 1 | top 3 | top 5 | |
| Ham2Pose-3d | $0.291^{\pm.006}$ | $0.386^{\pm.005}$ | $0.535^{\pm.005}$ | $3.875^{\pm.086}$ |
| SignDiffuse[11] | $0.285^{\pm.003}$ | $0.415^{\pm.005}$ | $0.654^{\pm.003}$ | $3.866^{\pm.054}$ |
| SignVAE (Base) | $0.385^{\pm.008}$ | $0.613^{\pm.009}$ | $0.745^{\pm.007}$ | $3.056^{\pm.108}$ |
| SignVAE (Ours) | $0.429^{\pm.009}$ | $0.657^{\pm.008}$ | $0.756^{\pm.008}$ | $2.651^{\pm.119}$ |

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

[10] R. Shalev-Arkushin, A. Moryossef, O. Fried. Ham2Pose: Animating Sign Language Notation into Pose Sequences. CVPR, 2023
[11] G. Tevet, S. Raab, B. Gordon, Y. Shafir, D. Cohen-Or, A. H. Bermano. MDM: Human Motion Diffusion Model. ICLR, 2023

# Conclusion

- **1**. SignAvatars is the first large-scale multi-prompt 3D holistic motion SL dataset, containing diverse forms of semantic input.
- **2**. A multi-objective optimization capable of dealing with the complex **interacting hands scenarios**, while respecting the biomechanical hand constraints.
- **3.** A new 3D sign language production (SLP) benchmark for our dataset, considering multiple prompts and full-body meshes.
- **4.** A VQVAE-based strong 3D SLP network significantly outperforming the baselines.