GestureHYDRA: Semantic Co-speech Gesture Synthesis via Hybrid Modality Diffusion Transformer and Cascaded-Synchronized Retrieval-Augmented Generation

Supplementary Material

A. Dataset

A.1. Dataset Details

Our curated Streamer dataset includes 18 categories of gestures with specific semantics, such as numbers, directions, Greet, and Deny. In addition, some parallel sentences in the dataset also contain specific semantic gestures. The details of the semantic gestures are shown in Table 1, and the distribution of different semantic gestures is illustrated in Fig. 2.

Table 1. The Types of Semantic Gestures in the Streamer Dataset

Types	Contents	Examples
Number	1-10	We'll give you the reduction of **four** hundred yuan
Direction	Upper, Lower, Upper left, Lower left, Upper right, Lower right	Click the **lower right** corner to add it to your collection now.
Greet	Hello (Hi)	**Hello** and welcome to our live room.
Deny	Don't (Doesn't, Not)	It's really **not** expensive.
Others	Parallel sentences	Whether it is **boys, girls, the elderly, or children**, can be used.

Our dataset includes 281 anchor actors, with a total video duration of about 58 hours. Each video is split into 10-second short clips, with a frame rate of 25 fps and an audio sampling rate of 22 kHz. The processed dataset contains 20,969 short clips, and the distribution of clip numbers for different anchors is shown in Fig. 3. The majority of actors have between 40 and 50 short clips.

After obtaining the clips, we follow SHOW [4] to reconstruct the SMPL-X [1] parameters from monocular videos. Our work focuses on body and gesture generation, so we do not reconstruct facial expression coefficients. Additionally, while SHOW uses PyMAF-X [5] to initialize hand poses, we replace PyMAF-X with the more powerful hamer [2] to improve the quality of initial hand reconstruction.

A.2. Automatic Annotation Approach

Our framework is also equipped with an automatic semantic gesture annotation procedure (shown in Fig. 1) that can extract semantic gestures from a video of each target person. Specifically, we first prepare 18 pre-defined standard semantic gestures in the form of 2D skeletons by using DWPose [3] detection, which can be regarded as semantic gesture templates. Given a target video, we first use automatic-speech-recognition (ASR) technique to convert its audio into text, and then locate the segments aligned with special trigger words (e.g., "first", "left" , and "hello"). By performing DWPose detection on these segments, we are able to calculate the similarity between the detected gestures and the corresponding pre-defined gesture template. It's worth noting that semantic gestures may not be triggered every time when a person speaks the same word, thus such a process is necessary to filter relevant segments for personalized gesture collection. Once a segment exhibits high consistency with the gesture template, we continue to perform the 3D human body reconstruction process, and the reconstructed result is accepted as a standard example of the corresponding gesture. Normally, we collect at least two examples for each gesture unless a specific gesture is quite rare or missing in the captured data.

Table 2. The statistical results of the user study

System	**Naturalness↑**	**Rhythm ↑**	**Semantic↑**
TalkSHOW	2.53	2.67	2.36
Probtalk	2.71	2.89	2.80
DSG	2.89	3.05	2.82
Ours	**3.82**	**3.84**	**4.29**

B. More Experimental Analysis

B.1. Human Evaluation.

We perform a user study including 17 examples from the Streamer dataset and 20 volunteers. For each example, 4 gesture sequences generated by our method and other SOTAs, accompanied by audio, are displayed in a random order. Each participant is asked to score the generated gestures based on the following three aspects: (a) **Naturalness:** Whether the generated gestures are natural and smooth; (b) **Rhythm:** Whether the rhythm of the generated gestures aligns with the audio rhythm; (c) **Semantics:** Whether the generated motions match the semantic information in the speech. The scoring range was from 1 to 5, with higher scores indicating better performance. The results in Table 2, demonstrate that our method significantly outperforms the comparison methods over all three metrics, highlighting the superiority of the gesture modeling ability of our system.

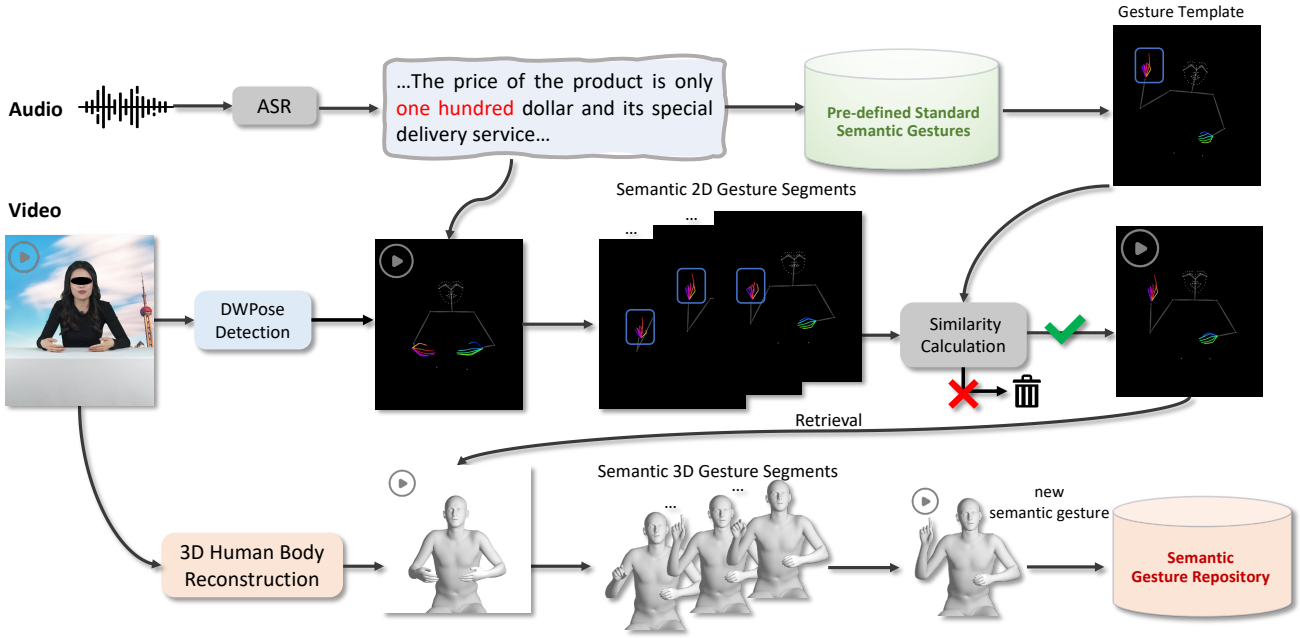

Figure 1. Automatic Annotation Pipeline for Personalized Semantic Gesture Repository.

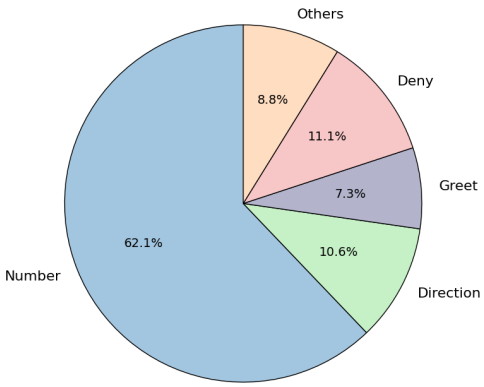

Figure 2. The distribution of the number of different semantic gestures.

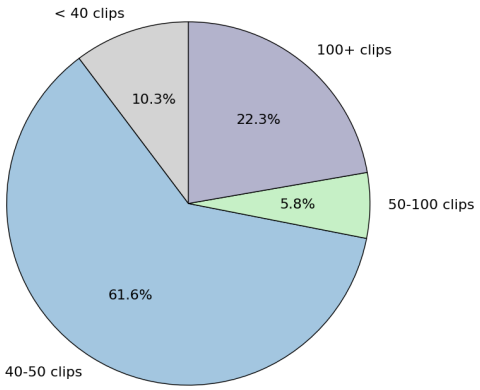

Figure 3. The distribution of the number of short clips for different anchors.

B.2. More Gesture Editing

Thanks to our hybrid-modality design, masking training strategy and motion style injection, the proposed GestureHYDRA not only possesses the ability to insert keyframes (i.e. Semantic gesture injection) but also includes other meaningful features, such as motion style transfer, motion inpainting, and motion erase. Our model allows the users to perform robust and flexible editing operations on the generated gesture sequence for distinct real-world demands.

Motion Style Transfer. Given the same speech input and different style embeddings for various characters, our proposed GestureHYDRA can generate body gestures in different styles, all while synchronizing with the given speech.(See Fig. 4) **Motion Inpainting.** Given the input speech, starting frame, and ending frame, our proposed GestureHYDRA can automatically generate the motion sequence between the starting and ending frames by inserting them as keyframes at the corresponding positions. (See

Fig. 5) **Motion Segment Replacement.** Our GestureHYDRA replaces unwanted frame segments in the generated motion with target gesture fragments as keyframes, while also treating the remaining usable frames as keyframes, thus enabling the replacement of undesired generated gestures.(See Fig. 6)

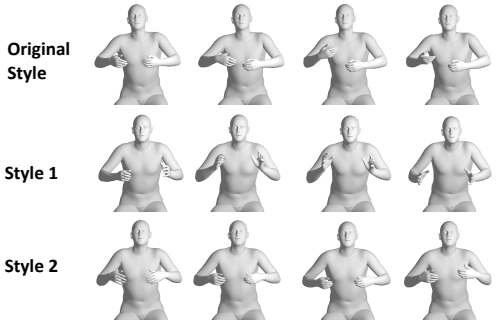

Figure 4. Motions generated in different styles under the same speech.

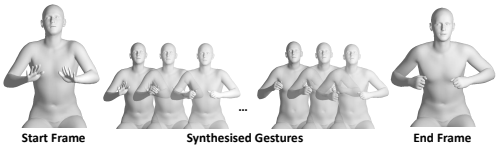

Figure 5. Motion inpainting given start and end frames.

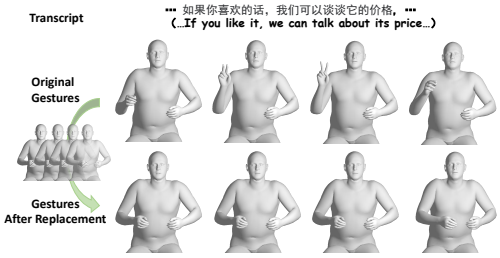

Figure 6. Replace undesired segments in generated motion.

Please watch the demo video we provided to view the generated results.

C. Limitations

Although our work paves the way for generating gestures with strong semantics, it still has certain limitations: 1) Our Streamer dataset currently excludes face reconstruction and digital human creation without vivid facial expressions partly degrades user experiences. 2) The designed Retrieval-Augmented Generation strategy relies on correct automatic speech recognition results. Ambiguous or incorrect recognition results will lead to an evident mismatch between speech audio and human gestures. Our future exploration will advance from these two directions.

References

[1] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *CVPR*, pages 10975–10985. Computer Vision Foundation / IEEE, 2019.

[2] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3d with transformers. In *CVPR*, pages 9826–9836. IEEE, 2024.

[3] Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective whole-body pose estimation with two-stages distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4210–4220, 2023.

[4] Hongwei Yi, Hualin Liang, Yifei Liu, Qiong Cao, Yandong Wen, Timo Bolkart, Dacheng Tao, and Michael J. Black. Generating holistic 3d human motion from speech. In *CVPR*, pages 469–480. IEEE, 2023.

[5] Hongwen Zhang, Yating Tian, Xinchi Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun. Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop. In *ICCV*, pages 11426–11436. IEEE, 2021.