

ViBES: A Conversational Agent with Behaviorally-Intelligent 3D Virtual Body

Juze Zhang¹ Changan Chen¹ Xin Chen² Heng Yu¹ Tiange Xiang¹ Ali Sartaz Khan¹
Shrinidhi K. Lakshmikanth¹ Ehsan Adeli¹
¹Stanford University ²ByteDance

1. Supplementary Material

In this supplementary material, we provide additional details about:

1. Supplementary video for qualitative examples.
2. Application cases of ViBES.
3. Additional implementation details of ViBES.
4. Additional details on the YouTube data processing pipeline (referenced in Sec. 4).
5. Additional details on constructing conversational motion from the AMASS dataset (referenced in Sec. 4).
6. Additional qualitative example of talking head generation and text-to-motion.

1.1. Supplementary Video

We provide a supplementary video to illustrate our method and results. The video presents: 1) the background and motivation of this work; 2) an explanation of the overall framework; and 3) detailed qualitative comparisons across different tasks, including conversational behavior generation, talking-head generation, co-speech gesture generation, and text-to-motion generation. We recommend watching the video with headphones, as it offers a more comprehensive understanding of our approach.

1.2. Application Cases of ViBES

Benefiting from the 3D representation, our conversational agent naturally extends into the spatial domain. And also, it can be used to drive more realistic video avatars. As illustrated in Fig. 2, we use an off-the-shelf video generation tool powered by Runway AI [15] to synthesize avatar videos, conditioned on our generated head motions.

1.3. Additional Implementation Details of ViBES

The original GLM-4 model [3] consists of 40 transformer layers with a hidden size of 4096 and an FFN dimension of 13,696, for a total of roughly 9B parameters. For our face and motion branches, directly duplicating this architecture would be inefficient. Instead, we adopt a lightweight variant with 40 layers, a hidden size of 512, and an FFN dimension of 4096, resulting in approximately 430M parameters for

each branch. We train ViBES on 4 L40S GPUs for about one week with a learning rate of 1×10^{-4} .

For the motion tokenizer, as discussed in the main manuscript, two popular paradigms of motion representation are widely used. For most experiments in this paper, we adopt a compositional motion tokenizer [1, 10]. In addition, we train a separate HumanML3D representation [5] to enable fair comparison with existing text-to-motion models that rely on this convention.

1.4. Additional details on the YouTube data processing pipeline

In this section, we give a detailed explanation of the data processing pipeline of our Converse3D dataset. We summarize the acquisition, processing, and filtering of our Converse3D dataset into two main procedures: automatic and manual processing steps, as illustrated in Fig 1. To build a high-quality 3D co-speech gesture dataset with concurrent and interactive body dynamics, we collect a considerable number of videos. They are then processed using automated methods to extract both audio and motion information.

Raw Video Downloading and Processing We crawl in-the-wild conversational videos from YouTube, targeting couple-interview channels, sports conversations, autobiographical interviews, entertainment talk shows, and news discussions on social topics. Using keywords such as *talk show*, *conversation*, and *interview*, we retrieve videos together with their metadata (duration, frame resolution, audio sampling rate, etc.); the keyword distribution is visualized in Fig. 3. In total we collect over 2,000 hours of raw footage and retain 1,095 hours of conversational clips after automatic filtering and light manual cleaning. We discard videos that do not satisfy our requirements on category, visual quality, language, or body visibility. In particular, we only keep English dialogues and download each video at its maximum available resolution to better recover facial and body motion. Most retained clips contain a clearly visible face, fewer than half show the upper body, and only a small fraction contain the lower body. Owing to the scale of the data, the initial filtering is performed automatically rather than by inspecting each video individually.

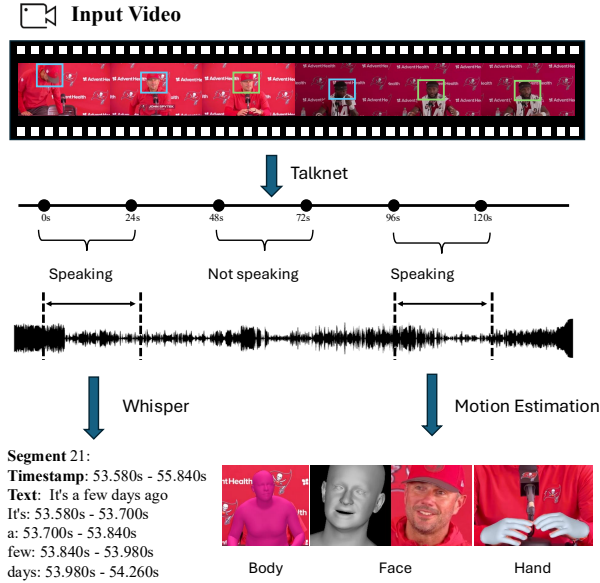


Figure 1. Overview of our YouTube data processing pipeline. The raw videos are processed to obtain high-quality 3D poses using automatic algorithms.

Audio Extraction and Filtering Audio and pose are the two core modalities in our Converse3D dataset. We first extract audio tracks from each video clip using FFmpeg. Since the person visible in the frame is not always the one speaking, we adopt TalkNet [17] to detect active speakers and remove non-speaking segments. Very short utterances (e.g., “yep”, “ok”) are merged into neighboring segments to avoid overly fragmented clips. After this stage, the dataset only contains segments where the visible speaker is actively talking. We then apply Whisper-large-v3 [14] to obtain transcripts with word-level timestamps for all retained segments.

Pose Estimation and Filtering As stated in the main manuscript, we use SMPL-X [12] and FLAME [9] as our parametric human models. Accordingly, we estimate three components from monocular video: body, hands, and face. We employ SPECTRE [2] to reconstruct FLAME parameters for facial motion, and use 4D-Humans [4] and HaMeR [13] to obtain SMPL-X body and hand parameters. We retain only sequences where the upper body is clearly visible to ensure reliable SMPL-X fitting. All motion sequences are transformed into a canonical world frame where the XZ-plane defines the ground, and resampled to 25 fps to ensure integer alignment with the audio tokens.

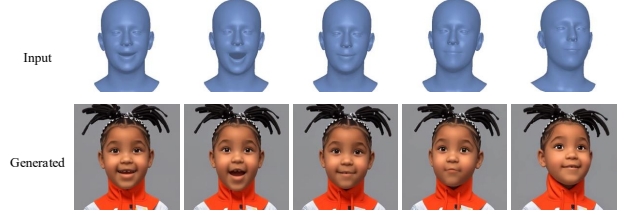


Figure 2. Application: driving video generation with ViBES. We use our generated 3D head motion as a behavioral condition to control an off-the-shelf video generation model (Runway AI [15]), producing realistic talking avatars.

1.5. Additional Details on Constructing Conversational Motion from the AMASS Dataset

Our Converse3D dataset aggregates multiple data sources. Existing motion datasets provide high-fidelity kinematics, but their modalities are often fragmented. Here we describe how we construct conversational data from the AMASS text-to-motion setting. All sequences are first retargeted to a unified SMPL-X/FLAME representation with consistent axes and then resampled to 25 fps. For AMASS [11], we synthesize the missing linguistic and acoustic modalities using the HumanML3D text annotations [5]. Given each motion description, we reformulate it as a natural-language question that explicitly requests the corresponding behavior (e.g., “Could you perform the motion of waving your hand?”). We then use a TTS system [7] to synthesize speech for this question, yielding a spoken query that would plausibly elicit the motion. Conditioned on the same question, GLM-4-Voice generates an answer, which we pair with the original motion to form a single conversational sequence (see Fig. 6 for the detailed prompt). The synthesized question audio, generated answer, and AMASS motion thus form aligned audio–text–motion triplets. All additions pass safety and style checks, near-duplicate filtering, and 25 fps timing validation, increasing tri-modal coverage without distorting the underlying motion distribution.

1.6. Additional Qualitative Examples of Talking-Head Synthesis and Text-to-Motion

To further demonstrate the effectiveness of our model on talking-head synthesis and text-to-motion, we provide an additional qualitative example for the speech-driven talking-head task in Fig. 4. Our method produces facial expressions that are well synchronized with the speech, particularly in the lip movements, and outperforms state-of-the-art baselines. We also present additional text-to-motion qualitative examples in Fig. 5. These results show that our model generates motions that more faithfully align with the textual descriptions, indicating a strong understanding of the input text.

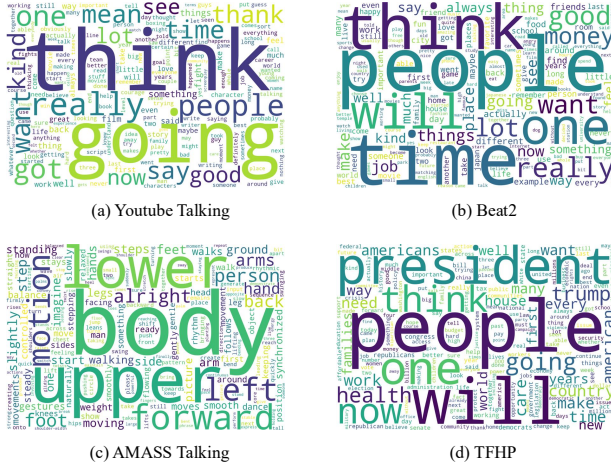


Figure 3. Word cloud visualization of our Converse3D data from different dataset sources.

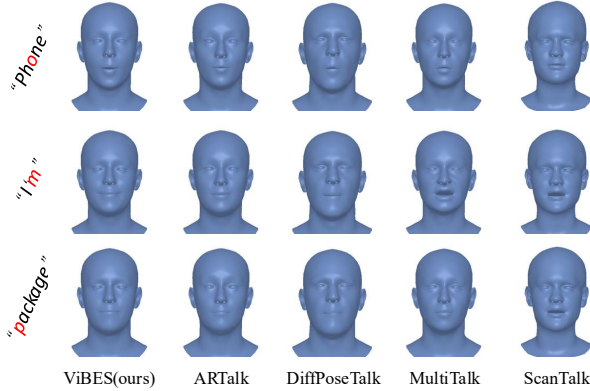


Figure 4. Additional qualitative comparisons with prior methods on the speech-driven 3D head animation task (with head pose held fixed). Since most existing models are trained on TFHP [16], we evaluate on this dataset to ensure a fair comparison.

References

- [1] Changan Chen, Juze Zhang, Shrinidhi K Lakshmikanth, Yusu Fang, Ruizhi Shao, Gordon Wetzstein, Li Fei-Fei, and Ehsan Adeli. The language of motion: Unifying verbal and non-verbal language of 3d human motion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6200–6211, 2025. 1, 4
- [2] Panagiotis P. Filntisis, George Retsinas, Foivos Paraperas-Papantoniou, Athanasios Katsamanis, Anastasios Roussos, and Petros Maragos. Visual speech-aware perceptual 3d facial expression reconstruction from videos, 2022. 2
- [3] Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*, 2024. 1
- [4] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4D: Reconstructing and tracking humans with transformers. In *ICCV*, 2023. 2
- [5] 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 *CVPR*, pages 5152–5161, 2022. 1, 2
- [6] Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. Momask: Generative masked modeling of 3d human motions. In *CVPR*, pages 1900–1910, 2024. 4
- [7] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 2
- [8] 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. 4
- [9] Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6):194:1–194:17, 2017. 2
- [10] Haiyang Liu, Zihao Zhu, Giorgio Becherini, Yichen Peng, Mingyang Su, You Zhou, Xuefei Zhe, Naoya Iwamoto, Bo Zheng, and Michael J. Black. Emage: Towards unified holistic co-speech gesture generation via expressive masked audio gesture modeling. In *CVPR*, 2024. 1
- [11] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision*, pages 5442–5451, 2019. 2
- [12] 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 *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2
- [13] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3D with transformers. In *CVPR*, 2024. 2
- [14] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023. 2
- [15] Runway AI, Inc. Runway: Ai video generation platform. <https://runwayml.com>, 2025. Accessed: 2025-11-20. 1, 2
- [16] Zhiyao Sun, Tian Lv, Sheng Ye, Matthieu Lin, Jenny Sheng, Yu-Hui Wen, Minjing Yu, and Yong-jin Liu. Diffposetalk: Speech-driven stylistic 3d facial animation and head pose generation via diffusion models. *ACM Transactions on Graphics (TOG)*, 43(4):1–9, 2024. 3
- [17] Ruijie Tao, Zexu Pan, Rohan Kumar Das, Xinyuan Qian, Mike Zheng Shou, and Haizhou Li. Is someone speaking?

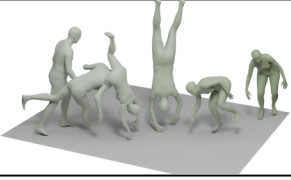
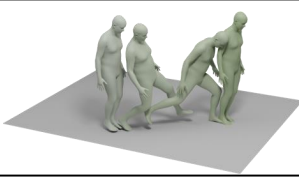
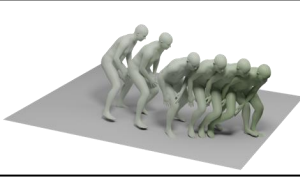
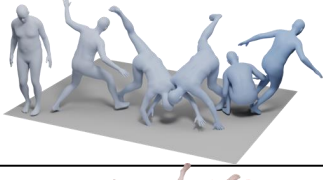
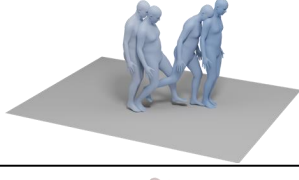
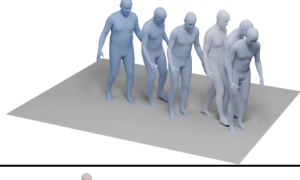
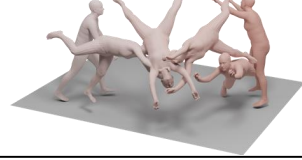
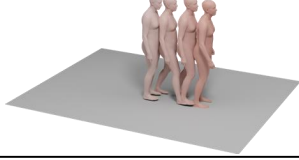
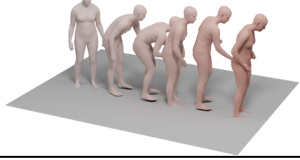

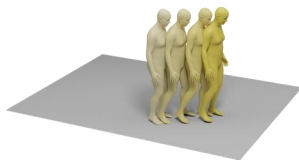

Text	A person walks forward then turns completely around and does a cartwheel.	A person taking a huge diagonal step	A person tries to clean the floor
Ours			
MotionGPT			
LoM			
MoMask			

Figure 5. Additional qualitative examples for text-to-motion generation. Given a text caption, we compare the 3D motion generated by our method with those generated by state-of-the-art methods, including MotionGPT [8], LoM [1], and MoMask [6]. Our model produces smooth, natural, and sometimes better motion in comparison with existing methods, which do not model the conversation behavior.

System Prompt for Motion-conditioned Conversational Answer Generation

System Prompt.

< |system| > User will provide you with a text instruction. Do it step by step. First, think about the instruction and respond in an interleaved manner, with 13 text token followed by 26 audio tokens. Please follow these steps carefully: Think about the instruction first. Respond in an interleaved manner: output 13 text tokens followed by 26 audio tokens. In your reply, imagine that you have a body and are already moving, pretending to perform 'the motion required by the question. Make sure your answer aligns with both the question and the motion being asked. Remember: the motion is imaginary (pretend), not real. If you describe the motion, use the first-person perspective (e.g., 'my hand,' 'my body,' 'my movement'). Please reply as if you are experiencing and expressing the motion yourself."

Figure 6. System prompt for motion-conditioned conversational answer generation. We instruct the model to generate answers as if the avatar had its own body and were responding through both speech and body movement.

exploring long-term temporal features for audio-visual active speaker detection. In *ACMMM*, page 3927–3935, 2021.