

# The Language of Motion: Unifying Verbal and Non-verbal Language of 3D Human Motion

Anonymous CVPR submission

Paper ID 14677

## 001 7. Supplementary

002 In this supplementary material, we provide additional details about:

- 003 1. Supplementary video for qualitative examples (referenced in Sec. 1).
- 004 2. Additional details on post-training (referenced in Sec. 3.4).
- 005 3. Additional details on editable generation (referenced in Sec. 4.3).
- 006 4. Additional details on emotion prediction (referenced in Sec. 4.4).
- 007 5. Results for text-to-motion.
- 008 6. Additional implementation details.
- 009 7. Additional qualitative example of co-speech gesture generation.

### 016 7.1. Supplementary Video

017 We provide a supplemental video to illustrate our results. In the video, we present: 1) an overview of our overall framework, 2) detailed qualitative comparisons across four tasks: co-speech gesture generation, editable gesture generation, text-to-motion generation, and emotion understanding, and 3) examples of failure cases to inspire further research. We recommend watching this video with your headphone, as video results provide a more comprehensive understanding of our approach.

### 026 7.2. Additional Details on Post-training

027 Existing datasets primarily provide pair-wise motion data but lack corresponding instructions. Following [3], we construct paired data for downstream tasks such as co-speech gesture generation and text-to-motion generation, equipping the model with instruction-following capabilities. Building upon existing datasets [5–8], we develop an instructional multi-modal dataset comprising several core tasks. Unlike previous work [3], our approach explicitly distinguishes each body part by introducing specific part-specific keywords. As illustrated in Tab. 1, each core task includes dozens of carefully designed instruction prompts.

### 7.3. Additional Details on Editable Gesture Generation

As shown in Tab. 1, we prompt the model with part-specific keywords, enabling it to generate any body part based on either audio or text inputs. This approach allows us to easily edit specific body parts. In this paper, we demonstrate this by prompting the model twice: once to generate the upper body from audio and once to generate the lower body from a text description. We anticipate that with further training on larger datasets, the model will be able to simultaneously follow input prompts from multiple sources.

### 7.4. Additional Details on Emotion Understanding

Since we perform instruction tuning during the post-training stage, the model does not always guarantee precise single-emotion label predictions. Instead of using a classification accuracy metric, we adopt text embedding distance metrics to evaluate the similarity between the predicted emotion and the ground truth labels. Specifically, we use BLEU [9], ROUGE, CIDEr [4], and BERTScore [12] to assess the semantic distances between the predicted and reference texts.

### 7.5. Results for Text-to-motion Generation

In the main paper, we focused on demonstrating our model’s capability in co-speech gesture generation as well as editable gesture generation. Another task that our model is naturally good at is text-to-motion generation. To understand how good our model is at generating motion from instructions, we investigate the quality of generated motion given text descriptions.

We show some qualitative examples of our text-to-motion generation in Fig. 1, where we also compare with existing work [3, 10, 11]. We can see that our model produces smooth, natural, and sometimes better motions in comparison with other generation methods. We encourage watching the supplementary video to get a more comprehensive understanding of our model’s text-following ability.

While our model shows strong text-to-motion genera-

Task	Input	Output
Audio-to-Full Motion	Based on [audio], generate a synchronized movement sequence involving both face, hands, upper and lower body. Listen to [audio] and produce movements that involve both the upper and lower body in harmony.	[face][hands] [upper][lower]
Audio-to-Full Motion	Based on [audio], generate a synchronized movement sequence involving both face, hands, upper and lower body. Listen to [audio] and produce movements that involve both the upper and lower body in harmony.	[face][hands] [upper][lower]
Audio&Transcript-to-Full Motion	Generate a set of movements for face, hand, upper, and lower body that correspond to the timestamped alignment in [audio&transcript] Using the precise timestamp match in [audio&transcript], generate corresponding face, hand, upper, and lower body movements.	[face][hands] [upper][lower]
Audio-to-Upper Body Motion	Using [audio], produce upper body movements that capture the tone and energy. From [audio], create a series of gestures that use the upper body to reflect its flow.	[upper]
Audio-to-Lower Body Motion	Interpret [audio] with lower body gestures that reflect its tempo. Create leg and foot movements that align with the intensity shifts in [audio].	[lower]
Audio-to-Hands Body Motion	Develop a set of hand movements that respond dynamically to [audio]. Generate expressive hand gestures that reflect the cues in [audio].	[hand]
Audio-to-Face Body Motion	Create expressions that correspond to the varying sentiments in [audio]. Listen to [audio] and generate a sequence of facial expressions that match its energy.	[face]
Emotion-to-Motion	Generate a movement sequence that fully embodies the emotion of [emotion] using the face, hands, upper body, and lower body. Express the emotion [emotion] through a series of actions involving the face, hands, upper, and lower body.”	[face][hands] [upper][lower]
Motion-to-Emotion	What emotion is conveyed by the movements in the face, hands, upper body, and lower body within [face][hands][upper][lower]? Examine the face, hand, upper, and lower body movements in [face][hands] [upper][lower] to interpret the emotional tone.	[emotion]
Text-to-Full Motion	Give me gestures involving the face, hands, upper body, and lower body that correspond to [caption] Show me gestures involving the face, hands, upper body, and lower body that capture the essence of Input: [caption].	[face][hands] [upper][lower]
Text-to-Upper Body Motion	Create an upper body gesture that aligns with the sentiment of [caption]. Develop an upper body action sequence that mirrors the tone in [caption].	[upper]
Text-to-Lower Body Motion	Illustrate the message in [caption] with lower body motions. Translate [caption] into a lower body movement sequence.	[lower]
Text-to-Lower Body Motion	Describe the motion represented by [upper][lower] using plain English. What does the [upper][lower] communicate? Please describe it in words.	[caption]

Table 1. Examples of instruction prompt templates during post-training. For each task, we show two examples of the input prompts and the output format.

tion on par or even better than existing models, we observe that the common text-to-motion metrics (e.g., FID [2]) are strongly coupled with the motion representation that existing work adopts, i.e., HumanML3D [2] (H3D-Format), because the VAEs are trained using that format. While the H3D-Format focuses predominantly on skeletal movements, such as swinging motions, it under-represents twisting rotations and other nuanced body dynamics. In contrast, our method prioritizes expressive motion with a composi-

tional representation, capturing a broader range of movements. Because these metrics are heavily entangled with specific motion representations, we find them not suitable to evaluate our method. We encourage readers to refer to the qualitative results in Fig. 1 and the supplementary video for a more comprehensive understanding. Future work is necessary to develop evaluation approaches that assess the quality of generated motion independently of the motion representation used.

## 7.6. Additional Implementation Details

**Model training.** Our model employs a two-stage training process: Generative Pre-training and Post-training. During the first stage of modality alignment, we trained the full model using  $8 \times$  NVIDIA H100-80GB GPUs and the AdamW optimizer with a learning rate of  $2e-4$ . Each configuration of the pre-trained model was trained until convergence. For the post-training stage, we used  $8 \times$  NVIDIA 3090-24G GPUs with the AdamW optimizer and a learning rate of  $1e-4$ . To ensure fair comparisons in ablation studies, each configuration of the post-trained model was trained for a fixed 350 epochs.

**Global Translation Prediction.** Benefiting from the compositional body representation, our approach generates high-quality expressive motions, particularly for gestures and emotion understanding. However, the holistic motion is divided into several body parts for local frames, as noted in [6]. To address this, we follow [6] and train a VAE module with a 4-layer TCN structure. This module takes the lower body as input and estimates the global translations  $T_{trans} \in \mathbb{R}^{T \times 3}$ .

## 7.7. Additional Qualitative Example of Co-speech Gesture Generation

To show the effectiveness of our model on co-speech gesture generation, we provide one more qualitative example in Fig. 2. We can see that our model generates gestures that are synchronized with the speech and expressive of the emotion, outperforming two state-of-the-art methods.

## References

- [1] Bohong Chen, Yumeng Li, Yao-Xiang Ding, Tianjia Shao, and Kun Zhou. Enabling synergistic full-body control in prompt-based co-speech motion generation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 6774–6783, 2024. 5
- [2] 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. 2
- [3] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. In *NeurIPS*, 2023. 1, 4
- [4] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 1
- [5] Haiyang Liu, Zihao Zhu, Naoya Iwamoto, Yichen Peng, Zhengqing Li, You Zhou, Elif Bozkurt, and Bo Zheng. Beat: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis. In *ECCV*, 2022. 1
- [6] 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 holis-

- tic co-speech gesture generation via expressive masked audio gesture modeling. In *CVPR*, 2024. 3, 5
- [7] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. Amass: Archive of motion capture as surface shapes. In *ICCV*, 2019. 146
- [8] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE, 2015. 150
- [9] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 158
- [10] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2023. 1, 4
- [11] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Yong Zhang, Hongwei Zhao, Hongtao Lu, Xi Shen, and Ying Shan. Generating human motion from textual descriptions with discrete representations. In *CVPR*, pages 14730–14740, 2023. 1, 4
- [12] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019. 170

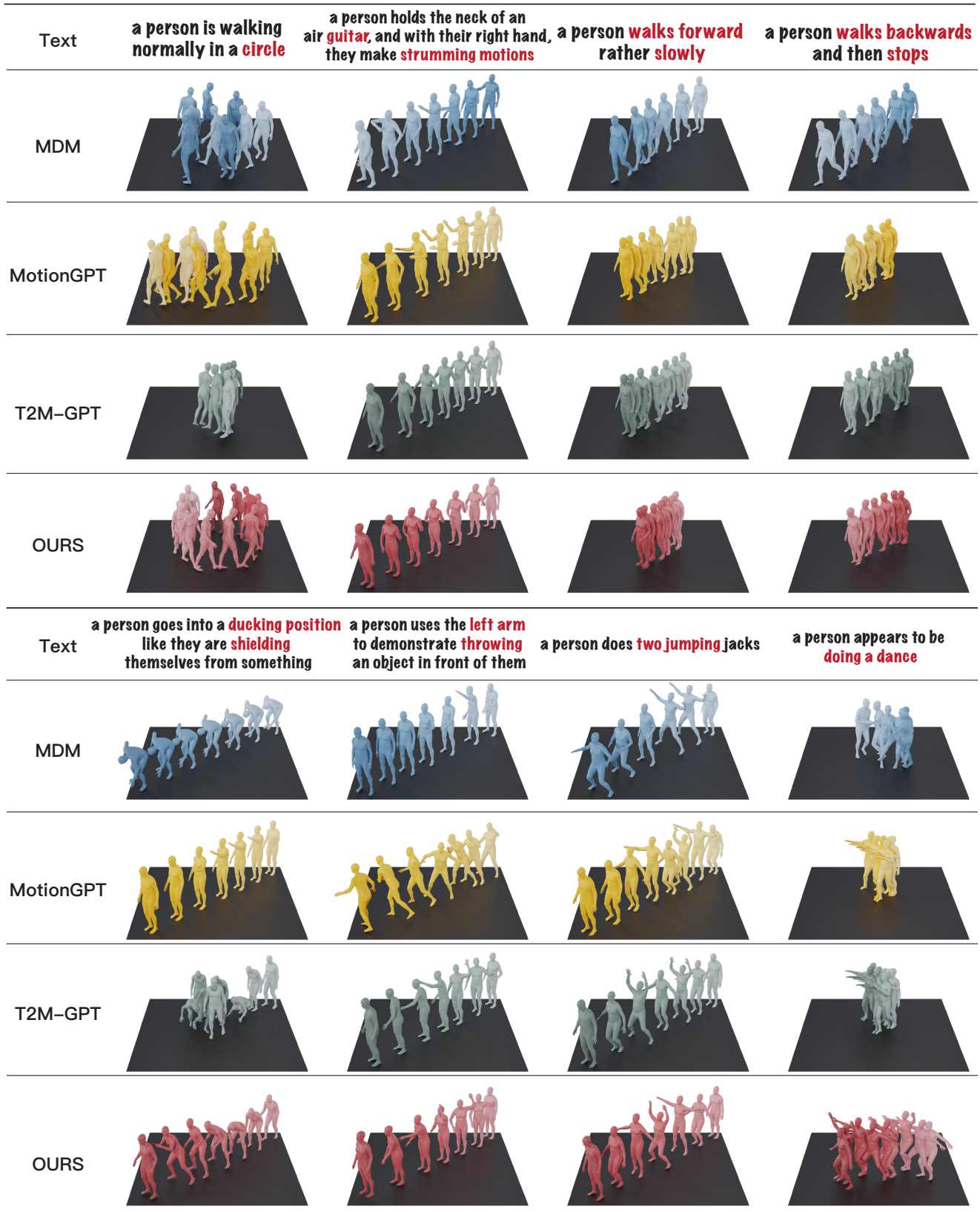


Figure 1. Qualitative examples for text-to-motion generation. Given a text caption, we compare the 3D motion generated by our method with those generated by state-of-the-art methods, including MDM [10], T2M-GPT [11], and MotionGPT [3]. Our model produces smooth, natural, and sometimes better motion in comparison with existing methods, which do not model the audio modality.



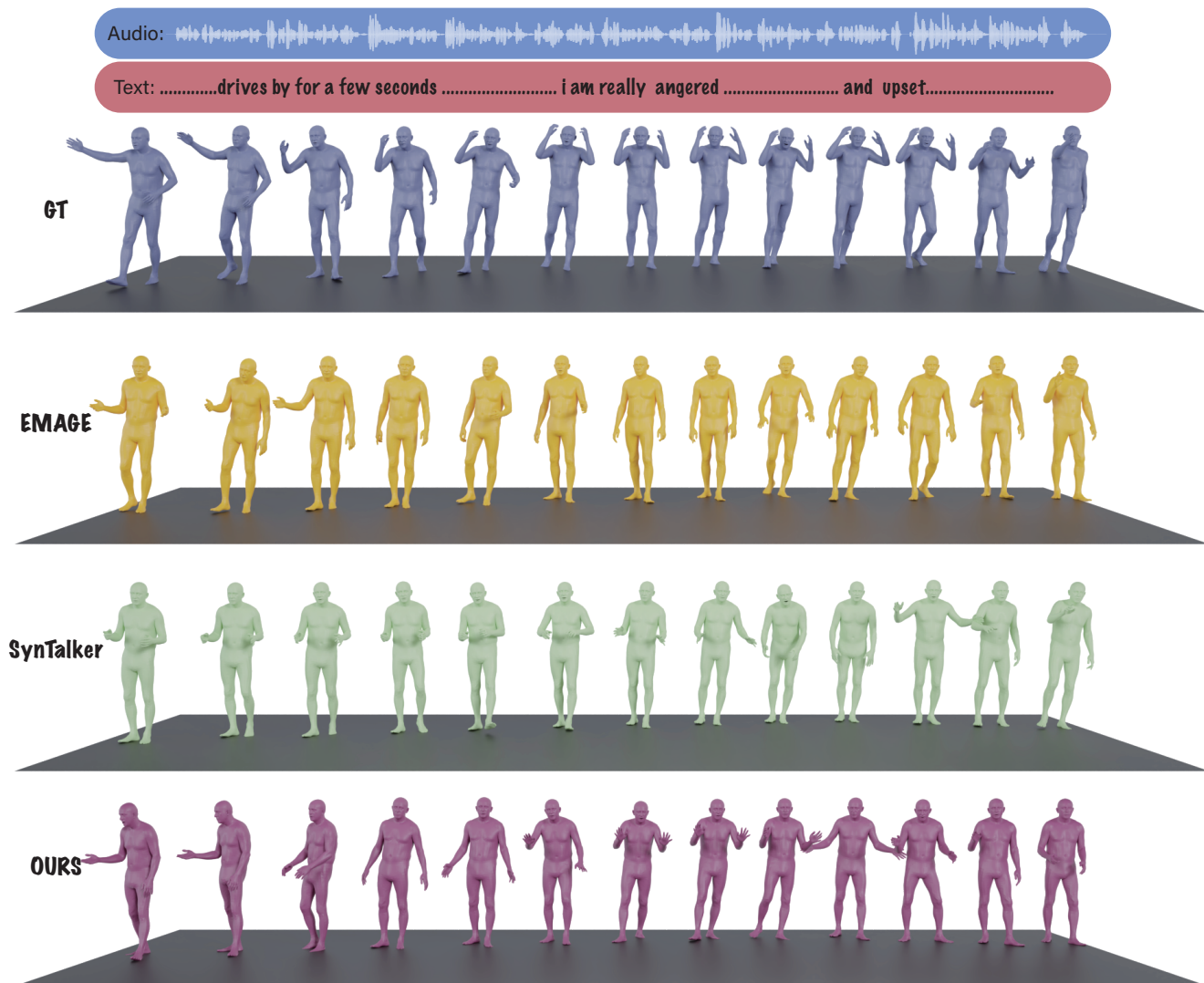


Figure 2. Additional qualitative example on co-speech gesture generation. Given an input speech, we visualize the ground truth 3D motion accompanying the audio, the motion generated by two baselines: EMAGE [6], SynTalker [1], and our method. Our model generates more diverse and expressive motion compared to existing methods, especially when the speaker emphasizes on words such as “angered” and “upset”.