

# Supplementary Materials: Enabling Synergistic Full-Body Control in Prompt-Based Co-Speech Motion Generation

Anonymous Authors

## 1 IMPLEMENTATION DETAILS

### 1.1 Loss Function in Section 4.2

**Contrastive Losses.** We adopted the foundational set of losses basing on [5], which are expressed as the weighted sum of 4 losses  $\mathcal{L}_{\text{CON}} = \mathcal{L}_R + \lambda_{\text{KL}} \mathcal{L}_{\text{KL}} + \lambda_E \mathcal{L}_E + \lambda_{\text{NCE}} \mathcal{L}_{\text{NCE}}$ .

More specifically,

$$\mathcal{L}_R = \mathcal{L}_1(H_{1:F}, \hat{H}_{1:F}^M) + \mathcal{L}_1(H_{1:F}, \hat{H}_{1:F}^T) \quad (1)$$

$$\begin{aligned} \mathcal{L}_{\text{KL}} = & \text{KL}(\phi^T, \phi^M) + \text{KL}(\phi^M, \phi^T) \\ & + \text{KL}(\phi^T, \psi) + \text{KL}(\phi^M, \psi). \end{aligned} \quad (2)$$

$$\mathcal{L}_E = \mathcal{L}_1(z^T, z^M). \quad (3)$$

$$\mathcal{L}_{\text{NCE}} = -\frac{1}{2N} \sum_i \left( \log \frac{\exp S_{ii}/\tau}{\sum_j \exp S_{ij}/\tau} + \log \frac{\exp S_{ii}/\tau}{\sum_j \exp S_{ji}/\tau} \right), \quad (4)$$

Reconstruction loss  $\mathcal{L}_R$  quantifies the accuracy of motion reconstruction from text or motion inputs using a smooth L1 loss. The Kullback-Leibler (KL) divergence loss  $\mathcal{L}_{\text{KL}}$  includes four components: two to regularize each encoded distribution— $\mathcal{N}(\mu^M, \Sigma^M)$  for motion and  $\mathcal{N}(\mu^T, \Sigma^T)$  for text—to align with a standard normal distribution  $\mathcal{N}(0, I)$ . The other two components enforce distributional similarity across the two modalities. A cross-modal embedding similarity loss  $\mathcal{L}_E$  mandates that the latent codes for text  $z^T$  and motion  $z^M$  exhibit similarity (utilizing a smooth L1 loss). Additionally, a contrastive loss  $\mathcal{L}_{\text{NCE}}$  leverages negative motion-text pairs to enhance the structuring of the latent space, where  $\tau$  is the temperature hyperparameter and  $S$  denotes for similarity.

The coefficients  $\lambda_{\text{KL}}$  and  $\lambda_E$  were set to  $10^{-5}$ , and  $\lambda_{\text{NCE}}$  to  $10^{-1}$  in our experiments, consistent with the settings in [5].

### 1.2 Model Training Details in Section 4.3

During this training stage, the pre-trained model components are frozen. To ensure feature consistency, we consistently use implicit labeling, regardless of whether the current ground-truth motion has a corresponding text label. Audio signals are set to zero in the absence of input, and are randomly masked during training to implement classifier-free guidance.

### 1.3 Model Inference

In the actual inference process, users may enter text prompts that specify control over multiple body parts, such as "a person is sitting and raising their left hand." It is challenging for the model to automatically determine which body parts need to be controlled while avoiding unnecessary manipulation of other parts, often resulting in awkward generative outcomes [2]. To address this, in our separate-then-combine strategy, we use an effective method

for part-wise control. We preprocess the prompt using a language model (T5X-Large), which deconstructs the prompt into multiple sub-prompts targeting individual body parts. Part-wise blending is then applied based on the presence of a sub-prompt for each body part; in the absence of a sub-prompt, the part-wise blending mechanism effectively zeroes out the text feature for that part. This approach also allows users to manually assign text prompts to specific body parts, enabling more precise and granular control. In addition, during the blending process of inference signals, similar to many works using classifier-free guidance [6, 7], our model supports a finer-grained control by manually manipulating the part-wise blending weights to control the intensity of the signal.

In evaluating single-modality signal metrics, we apply a single-source condition for fairness, omitting our proposed separate-then-combine strategy. For the quantitative assessment of speech-to-motion generation, we use the test set division from EMAGE [4], the speech-to-motion dataset employed. During this evaluation, part-wise blending is solely conditioned on the audio signal. For the quantitative assessment of text-to-motion results, we follow the test set division of HumanML3D [3]. Here, part-wise blending is solely conditioned on the text signal.

### 1.4 Implementation of Baselines in Section 6.3

**Baselines in Table 1.** For MDM [6] and T2M-GPT [9] due to the inherent seeding in our method, for a fair comparison during testing, with the MDM method, we infer 128 frames at once, incorporating the first 16 frames into the inference process via inpainting. For the T2M-GPT method, we treat the first four tokens as known and allow the model to autoregressively generate the subsequent tokens.

**Baselines in Table 2.** We directly report their results shown in their original papers. Since the original results are without variances, we keep this tradition in the table.

## 2 ADDITIONAL EXPERIMENTAL RESULTS

### 2.1 More Results for Main Paper Figure 4 and 5

To compensate main paper Figure 4, we provide more results of full-body synergistic generation in Figure 1, which illustrates the idealized generation quality of our methods under various of audio and prompt conditions. Furthermore, to compensate main paper Figure 5, we provide more results of ablation study in Figure 2, showing that the ablation results are solid so that our training and inference components are solid.

### 2.2 Comparison with FreeTalker and GestureDiffuCLIP

#### FreeTalker

FreeTalker [8] is a neural model designed to perform audio-to-motion and text-to-motion tasks simultaneously within the same

network. We experimented with feeding both speech and text inputs to the model simultaneously during inference. As shown in Figure 4, the model fails to generate synergistic co-speech motion when receiving both conditions simultaneously. This highlights that training the model directly for audio-to-motion and text-to-motion tasks does not automatically enable it to produce synergistic results under dual conditions.

### GestureDiffuCLIP

GestureDiffuCLIP [1] enables users to control the style of generated co-speech motion using text prompts. To illustrate the fundamental difference between our synergistic co-speech motion generation and text-based style control, we conducted a simple comparison shown in Figures 3. As the authors did not release the source code, we re-implemented the model for this comparison. As depicted in the figures, under the textual prompts "A person is sitting while talking" and "A person kneels down while talking," our results (left) accurately follow the prompts, while GestureDiffuCLIP's (right) only shows a tendency to squat, underscoring the fundamental difference between text-based style control and text-based motion control.

## REFERENCES

- [1] Tenglong Ao, Zeyi Zhang, and Libin Liu. 2023. GestureDiffuCLIP: Gesture Diffusion Model with CLIP Latents. *ACM Trans. Graph.* (2023), 18 pages. <https://doi.org/10.1145/3592097>

- [2] Nikos Athanasiou, Mathis Petrovich, Michael J. Black, and Gül Varol. 2023. SINC: Spatial Composition of 3D Human Motions for Simultaneous Action Generation. *ICCV* (2023).
- [3] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. 2022. Generating Diverse and Natural 3D Human Motions From Text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5152–5161.
- [4] Haiyang Liu, Zihao Zhu, Giorgio Becherini, Yichen Peng, Mingyang Su, You Zhou, Naoya Iwamoto, Bo Zheng, and Michael J. Black. 2024. EMAGE: Towards Unified Holistic Co-Speech Gesture Generation via Masked Audio Gesture Modeling. *arXiv:2401.00374 [cs.CV]*
- [5] Mathis Petrovich, Michael J. Black, and Gül Varol. 2023. TMR: Text-to-Motion Retrieval Using Contrastive 3D Human Motion Synthesis. In *International Conference on Computer Vision (ICCV)*.
- [6] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. 2023. Human Motion Diffusion Model. In *The Eleventh International Conference on Learning Representations*. <https://openreview.net/forum?id=SJ1kSyO2jwu>
- [7] Sicheng Yang, Zhiyong Wu, Minglei Li, Zhensong Zhang, Lei Hao, Weihong Bao, Ming Cheng, and Long Xiao. 2023. DiffuseStyleGesture: Stylized Audio-Driven Co-Speech Gesture Generation with Diffusion Models. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*. International Joint Conferences on Artificial Intelligence Organization, 5860–5868. <https://doi.org/10.24963/ijcai.2023/650>
- [8] Sicheng Yang, Zunnan Xu, Haiwei Xue, Yongkang Cheng, Shaoli Huang, Mingming Gong, and Zhiyong Wu. 2024. Freetalker: Controllable Speech and Text-Driven Gesture Generation Based on Diffusion Models for Enhanced Speaker Naturalness. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.
- [9] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. 2023. T2M-GPT: Generating Human Motion from Textual Descriptions with Discrete Representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

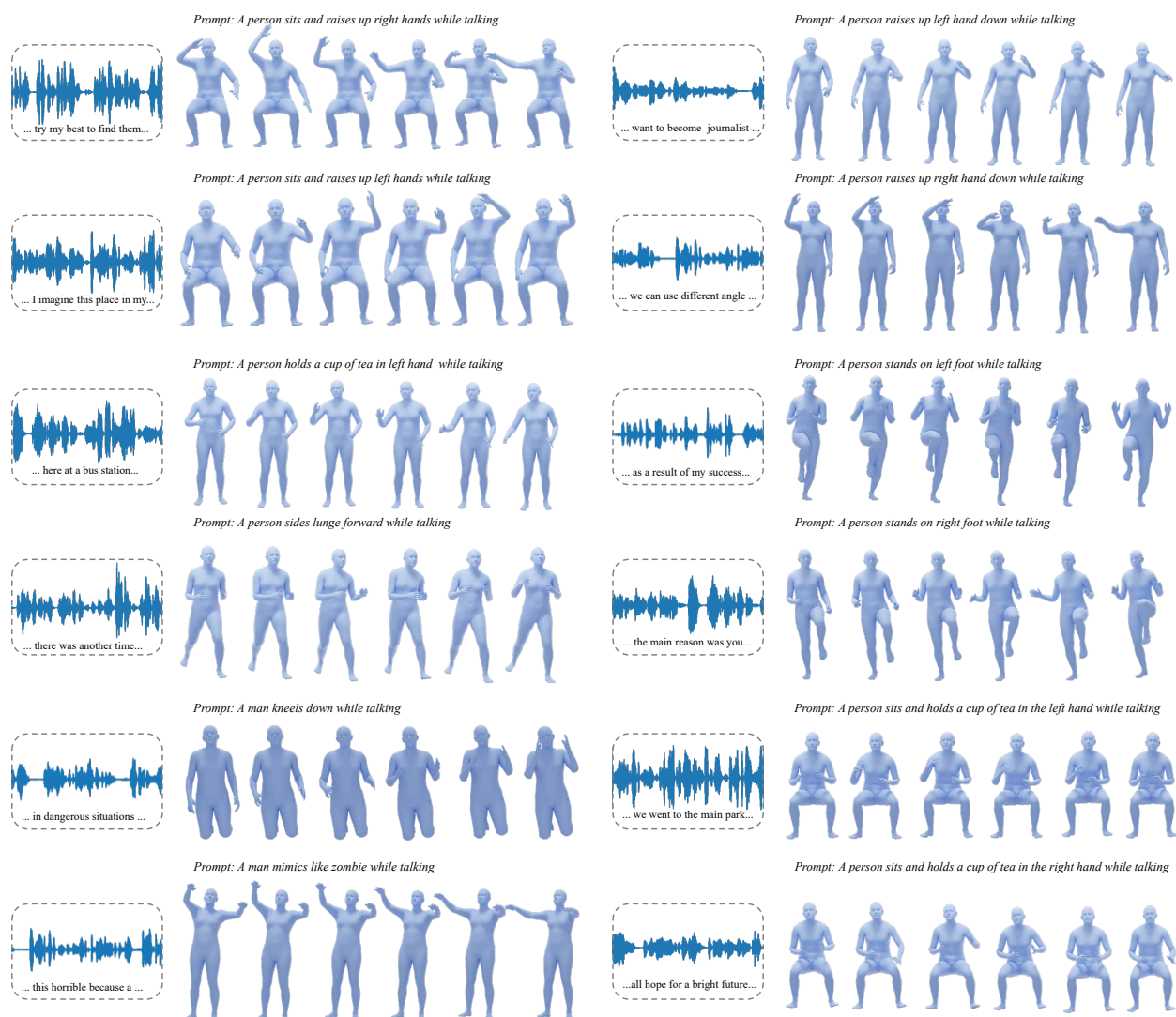


Figure 1: More results for main paper Figure 4 (full-body synergistic generation).



Figure 2: More results for main paper Figure 5 (ablation study)



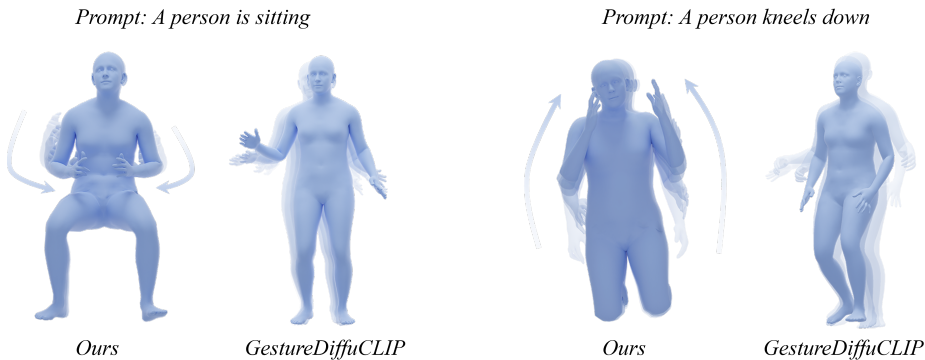


Figure 3: Qualitative comparison with GestureDiffuCLIP conditioned on audio and text features simultaneously.

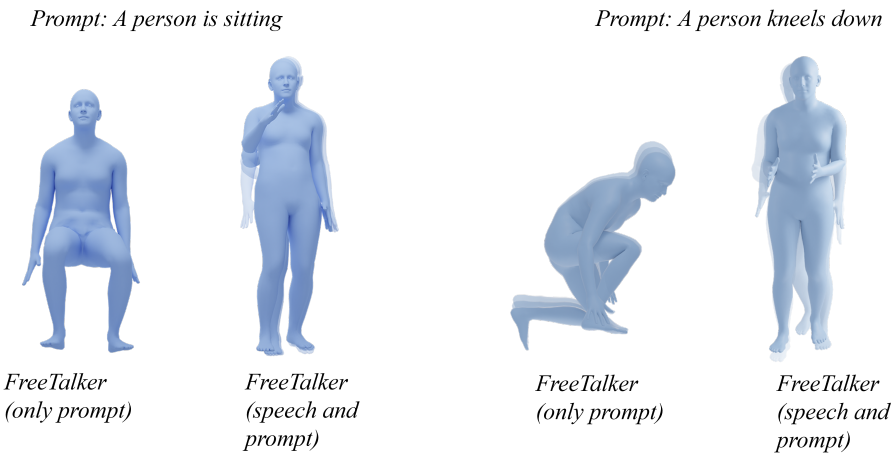


Figure 4: Qualitative results of FreeTalker conditioned on audio and text features simultaneously