PolySLGen: Online Multimodal Speaking–Listening Reaction Generation in Polyadic Interaction

Supplementary Material

A. Data Pre-processing

A.1. Transcription

As illustrated in Fig. 1, to segment utterances from the per-participant audio in the DnD Gesture Dataset [7], we first apply Voice Activity Detection (VAD) using Pyannote-audio [1]. We then perform Automatic Speech Recognition (ASR) using stable-ts [11] to obtain the transcription and corresponding timestamps for each utterance.

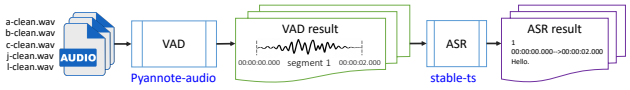

Figure 1. Transcription workflow. Utterances are extracted by pyannote-audio [1], and then transcribed using stable-ts [11].

A.2. Data Chunking

For each utterance, we extract a 64-frame segment (approximately 2.56 seconds) starting at the utterance onset. Segments from the target participant's utterances are assigned to the speaking subset, while all others are categorized as the listening subset. For dataset balance, the listening subset is downsampled to match the size of the speaking subset.

To make a chunk, each segment is paired with a speech history up to 512 frames (approximately 20.48 seconds) preceding the segment to support modeling long-form, engaging conversations. We select 512 frames because this window provides sufficient contextual information for generating coherent and contextually appropriate responses. Some examples can be found in Fig. 2. Unlike linguistic context, the motion observations include only 64 frames (approximately 2.56 seconds in duration) for all participants. In total, the processed dataset contains 8,926 chunks for training and validation, and 3,028 chunks for testing.

A.3. Speech Style Extraction

We first extract the utterance audio using the timestamps obtained from ASR, and then utilize StyleTTS 2 [6] to extract the corresponding speech style.

A.4. Motion Pre-processing

The captured motion data are originally in BVH format, containing a global translation and 54 joint rotations per frame. We convert the motion data into a 327-dimensional representation. The first three dimensions correspond to global translation, while the remaining 324 dimensions represent the 6D rotations [14] of the 54 body and hand joints.

B. Training Details

B.1. Loss Function

The full loss function is defined as:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{text}} \cdot \mathcal{L}_{\text{text}} + \lambda_{\text{style}} \cdot \mathcal{L}_{\text{style}} + \lambda_{\text{state}} \cdot \mathcal{L}_{\text{state}} + \mathcal{L}_{\text{motion}}, \quad (1)$$

where $\mathcal{L}_{\text{text}}$, $\mathcal{L}_{\text{style}}$, $\mathcal{L}_{\text{state}}$, and $\mathcal{L}_{\text{motion}}$ represent the losses associated with text token, speech style feature, speaking-state score, and motion, respectively. The coefficients λ_{text}, λ_{style}, and λ_{state} are the respective weighting factors. The motion loss $\mathcal{L}_{\text{motion}}$ is defined as:

$$\begin{aligned}\mathcal{L}_{\text{motion}} = \lambda_{\text{repr}} \cdot \mathcal{L}_{\text{repr}} + \lambda_{\text{keypoint}} \cdot \mathcal{L}_{\text{keypoint}} \\ + \lambda_{\text{root}} \cdot \mathcal{L}_{\text{root}} + \lambda_{\text{reg}} \cdot \mathcal{L}_{\text{reg}},\end{aligned} \quad (2)$$

where $\mathcal{L}_{\text{repr}}$, $\mathcal{L}_{\text{root}}$, and $\mathcal{L}_{\text{keypoint}}$ are L2 losses applied to the motion representation, the global translation, and the localized joint positions, respectively. The corresponding weighting coefficients are λ_{repr}, $\lambda_{\text{keypoint}}$, λ_{root}, and λ_{reg}. The regularization term, $\mathcal{L}_{\text{reg}}$, penalizes unnatural motion characteristics, including unrealistic joint velocities, excessive accelerations, and implausible foot-ground contact patterns.

The loss weights for the training objectives are set as follows: $\lambda_{\text{text}} = 1.0$, $\lambda_{\text{style}} = 100.0$, $\lambda_{\text{state}} = 0.4$, $\lambda_{\text{repr}} = 0.5$, $\lambda_{\text{keypoint}} = 1000.0$, $\lambda_{\text{root}} = 50$, and $\lambda_{\text{reg}} = 10$.

B.2. Multimodal Instruction Template

We adapt the instruction template to incorporate speech style, full-body motion, and social cues for polyadic interaction. An example template is shown in Fig. 3.

Adaptation for Multimodal Interaction. To support multimodal interaction, we extend the instruction template by assigning distinct roles to each participant and introducing additional special tokens to represent motion observations and social cues.

Adaptation for Multimodal Outputs. To accommodate variable-length text output while maintaining a fixed-length 64-frame motion sequence, text decoding is terminated either upon the generation of the first end-of-turn token, `<eot_id>`, or upon reaching a maximum of 64 generated words, whichever occurs first. Following this token, one additional embedding is decoded as the speech style, and the subsequent 64 embeddings are decoded as motion.

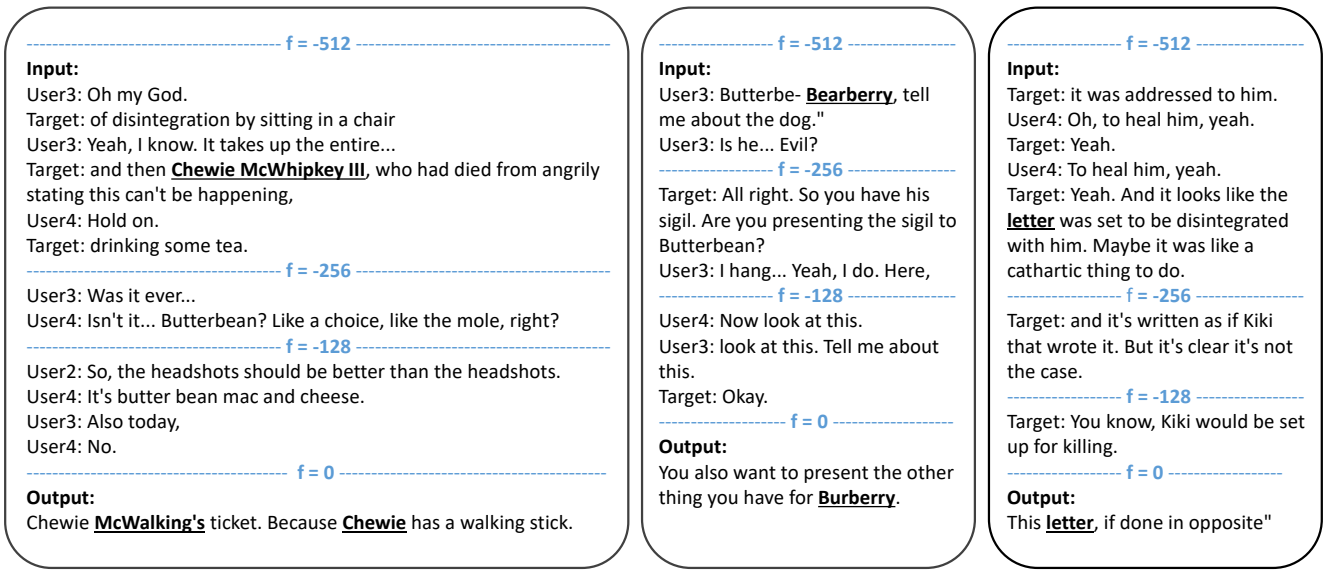

Figure 2. Three examples of the conversational chunks extracted from the DnD Gesture Dataset [7]. Keywords essential for generating in-context responses are underlined. f indicates the frame index. The output speaker in each example is the target participant.

Interaction Template

Input:
User1: <TEXT>...<TEXT><eot_id><STYLE>
User3: <TEXT>...<TEXT><eot_id><STYLE>
Target: <TEXT>...<TEXT><eot_id><STYLE>
...
Motion: <POSE>...<POSE>
Social: <SCUE><SCUE>
Target:
Output:
<TEXT>...<TEXT><eot_id><STYLE><POSE>...<POSE>

Figure 3. Multimodal instruction template for polyadic interaction. <TEXT> denotes text embeddings, <STYLE> represents speech style embeddings, <POSE> corresponds to motion embeddings, and <SCUE> refers to social cue embeddings.

C. Evaluation Details

C.1. Speech Metrics

SIM. To assess voice similarity, we follow a controlled evaluation approach to isolate the effect of the generated text. Specifically, we synthesize audio using the ground-truth text and the generated speech style. This ensures comparability and avoids cases where no speech is generated. Voice similarity is evaluated only for the speaking subset.

BeatAlignDiff. We adopt the beat alignment metric as defined in [7], with the difference computed as:

$$\text{BeatAlignDiff} = |\text{BeatAlign}_{\text{pred}} - \text{BeatAlign}_{\text{GT}}|. \quad (3)$$

Audio beats are extracted from onset timestamps of synthesized speech, while motion beats are identified as local minima in the velocity magnitudes of selected body joints.

Beat alignment is computed only when both ground-truth and generated speech are available. To ensure that the evaluation does not benefit from skipping chunks with missing generated speech, we analyzed one evaluation run and found that 28 out of 1,514 speaking chunks (0.84%) generated by SOLAMI [5] lacked generated speech. In contrast, our method produces only 3 such cases (0.19%).

C.2. Motion Metric

Motion Evaluator. A motion evaluator is pre-trained for both FID and Diversity, since both are calculated in a latent feature space. Following prior motion generation works [3, 4, 8, 10], we adopt a convolutional encoder-decoder architecture and train it to reconstruct motions from the DnD Gesture Dataset. Hyperparameter settings are set following [3]. The resulting evaluator achieves a mean per-joint position error (MPJPE) of 20.89 mm and a root joint Euclidean distance of 21.07 mm.

Diversity. We compute diversity for both the generated and ground-truth motions to compare the variability between the two distributions. Diversity quantifies the average pairwise Euclidean distance among reactive motion features within a set. Let $\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_n$ denote the feature vectors of the generated motions. The diversity is computed as:

$$Diversity = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \|\mathbf{f}_i - \mathbf{f}_j\|_2 . \quad (4)$$

We follow the DnD Group Gesture dataset and baselines [7, 8], measuring diversity relative to ground-truth statistics, where values closer to ground-truth indicate variability

Table 1. Comparison of PolySLGen with baselines with standard deviation over five runs. †: Infer speaking state based on the generated text.*: Synthesize speech using prompts from test set audio. Diversity closer to the ground truth (117.09) indicates better performance.

Method	Motion					Speech			State
	Root↓ (mm)	MPJPE↓ (mm)	FID↓	Diversity→	BeatAlign Diff.↓	BERT Score↑	WER↓	SIM↑	AP↑
Random	$140.4^{\pm 0.90}$	$200.4^{\pm 0.90}$	$17.82^{\pm 0.035}$	$120.46^{\pm 0.406}$	$0.018^{\pm 0.001}$	$0.458^{\pm 0.0019}$	$1.699^{\pm 0.014}$	$0.494^{\pm 0.0048}$	$0.50^{\pm 0.003\dagger}$
NN cond.	$134.7^{\pm 0.00}$	$187.7^{\pm 0.00}$	$16.36^{\pm 0.000}$	$105.42^{\pm 0.000}$	$0.023^{\pm 0.000}$	$0.451^{\pm 0.0000}$	$2.075^{\pm 0.000}$	$0.520^{\pm 0.0000}$	$0.52^{\pm 0.000\dagger}$
LLM + ConvoFusion [7]	$125.7^{\pm 0.19}$	$170.1^{\pm 0.06}$	$18.57^{\pm 0.012}$	$72.45^{\pm 0.139}$	-	$0.388^{\pm 0.0007}$	$13.318^{\pm 0.073}$	-	$0.50^{\pm 0.000\dagger}$
LM-L2L Adapted [8]	$185.2^{\pm 0.00}$	$187.6^{\pm 0.00}$	$17.22^{\pm 0.000}$	**116.36**$^{\pm 0.000}$	-	-	-	-	-
SOLAMI [5]	$188.6^{\pm 1.22}$	$180.9^{\pm 2.08}$	$14.86^{\pm 0.122}$	$100.13^{\pm 1.575}$	$0.061^{\pm 0.005}$	$0.428^{\pm 0.0002}$	$1.854^{\pm 0.015}$	$0.745^{\pm 0.0001*}$	$0.50^{\pm 0.001\dagger}$
Ours	**108.7**$^{\pm 0.16}$	**144.9**$^{\pm 0.07}$	**12.18**$^{\pm 0.007}$	$113.32^{\pm 0.055}$	**0.007**$^{\pm 0.001}$	**0.508**$^{\pm 0.0011}$	**1.436**$^{\pm 0.008}$	**0.642**$^{\pm 0.0010}$	**0.67**$^{\pm 0.000}$

better matching real interactions.

C.3. Social Semantics Metrics

We use Mean Angular Error MAE_{head} and social cue score error to evaluate the social semantics of the generated reaction. MAE_{head} is the rotation angle between the generated and ground-truth head orientations. The social cue score error semantically represents the difference in the attention that each non-target participant receives from the target participant. It is computed as the difference between the non-target participant's social cue scores derived from the generated and ground-truth head positions and orientations. Specifically, the social cue score for a non-target participant i at frame k is defined as:

$$s_i^k = \cos(\mathbf{u}_{\text{P}}^k, \mathbf{v}_{\text{P}\to i}^k), \tag{5}$$

, where $\mathbf{u}_{\text{P}}^k$ is the head orientation of the target participant, and $\mathbf{v}_{\text{P}\to i}^k$ is the relative head vector from the target participant to the non-target participant i.

C.4. Baseline Comparison with Standard Deviation

We run all evaluations five times, and only the average results are reported in the main paper for clarity. In Tab. 1, we present the baseline comparisons along with standard deviations. Compared to SOLAMI, PolySLGen achieves not only better performance but also lower standard deviations across most metrics, indicating improved stability and consistency. The LM-L2L Adapted baseline shows no variation across runs, as its mapping from embeddings to pose space is deterministic. In contrast, the motion generated by PolySLGen varies based on the preceding generated text due to the autoregressive nature of LLMs.

D. Baseline Implementation

D.1. NN Condition

For each chunk, we use the embeddings of the input observations to retrieve the Nearest-Neighbor (NN) from the training set, and take the corresponding response as the output. Text embeddings are obtained using the embedding layer of Llama3–8B-Instruct [2], while motion embeddings are from the motion evaluator introduced in Sec. C.2.

D.2. LLM + ConvoFusion

We utilize Llama3-8B-Instruct [2] as the LLM backbone and disable the audio condition of ConvoFusion [7] by setting it to unconditional tokens, following its original design. As ConvoFusion is already trained on the DnD Group Gesture dataset, finetuning is not performed for this baseline.

D.3. SOLAMI

Motion Tokenizer. In SOLAMI [5], motions are first encoded into tokens. Following this approach, we train a VQ-VAE-based motion tokenizer using the network from [10] for DnD Gesture Dataset. Consistent SOLAMI, we decompose the pose into three groups: root position, hand rotations, and body rotations. The training loss includes L2 loss on global translation, pose representation, keypoint positions and velocities, and a commitment loss. We assign one codebook per group, each containing 512 codewords with a hidden size of 256 and a temporal depth of 2.

The resulting tokenizer achieves an MPJPE of 90.5 mm and an Euclidean distance of 7.9 mm for the root joint position, which are comparable to the 88 mm MPJPE reported by SOLAMI on their dataset.

Direct Training. Pre-training is not performed since the DnD Group Gesture dataset lacks motion captions. This also ensures a fair comparison with our PolySLGen.

E. Additional Experiments

E.1. Comparison to Extended Baselines

All baselines are originally designed for generating speaking behaviors. We further investigate how well they can handle listening reactions when provided with both speaking and listening data. Note that some adaptations require changes that may deviate from the original approaches.

For LLM+ConvoFusion, we finetune the language model to generate both speaking and listening reactions. Listening

Table 2. Comparison of PolySLGen with further extended baselines. SOLAMI is performed with LoRA finetuning and without pre-training. †: Infer speaking state based on the generated text.*: Synthesize speech with prompts from test set audio. Diversity closer to the ground truth (117.09) indicates better performance. Underline denotes the second-best results.

Method	Training Data		Motion					Speech			State
	Listening	Speaking	Root↓ (mm)	MPJPE↓ (mm)	FID↓	Div.→	BeatAlign Diff.↓	BERT Score↑	WER↓	SIM↑	AP↑
LLM + ConvoFusion [7]	w/o finetune		125.7	<u>170.1</u>	18.57	72.45	-	0.388	13.318	-	0.50†
	✓	✓	<u>121.5</u>	170.5	17.43	67.18	-	**0.511**	**1.396**	-	0.56†
LM-L2L Adapted [8]	✓		185.2	187.6	17.22	<u>116.36</u>	-	-	-	-	-
	✓	✓	167.3	186.7	17.05	**116.64**	-	-	-	-	-
SOLAMI [5]		✓	188.6	180.9	<u>14.86</u>	100.13	<u>0.061</u>	0.428	1.854	0.745*	0.50†
	✓	✓	170.9	181.3	15.21	103.87	0.065	0.503	1.548	0.744*	0.52†
Ours	✓	✓	**108.7**	**144.9**	**12.18**	113.32	**0.007**	<u>0.508</u>	<u>1.436</u>	**0.642**	**0.67**

Table 3. Impact of different motion representations. For controlled experiments on motion tokens, the pose fusion module is disabled. Diversity closer to the ground truth (117.09) indicates better performance. Root and MPJPE are reported in millimeters (mm).

Pose Fusion	Motion Representation	Root↓	MPJPE↓	FID↓	Div.→
✓	3D keypoints transl.+rotations	126.9 **108.7**	150.8 **144.9**	13.82 **12.18**	123.05 113.32
✗	motion tokens transl.+rotations	316.4 **124.6**	177.8 **152.5**	16.00 **13.53**	80.60 **123.19**

responses are represented using textual placeholders without spoken content, such as "...", "(...listening...)", or no text output. For LM-L2L Adapted, we further include chunks with speaking reactions, but disregard textual responses. For the speech-based SOLAMI, we further include chunks with listening reactions and constrain the model to not generate any speech tokens for listening reactions.

As shown in Tab. 2, LLM+ConvoFusion achieves significant improvements on speech-related metrics, as expected due to its re-grounding on the topic of the dataset. However, motion-related metrics show only marginal gains, a pattern also observed in LM-L2L Adapted. Since both methods rely solely on past conversation in text, these results again underscore the importance of incorporating group motion observations to generate contextually aligned motion reactions in polyadic interactions. SOLAMI shows degradation in MPJPE, FID, and BeatAlignDiff, with only minor improvements in speech metrics, suggesting that architectures designed for simpler settings cannot be directly extended and applied to address our task.

Compared to all baselines and their stronger variants, PolySLGen remains the most competitive method, consistently ranking first or second across all speech and motion metrics, except for Diversity.

E.2. Motion Representation

In PolySLGen, poses are represented using global translation and 6D rotations [14] of body and hand joints. We further investigate the impact of alternative motion representations such as 3D keypoint positions and motion tokens.

3D Keypoint Positions. While 3D keypoint positions are a commonly used pose representation [7, 12], the lack of constraints between joints introduces inconsistencies in body shapes, which results in temporal instability. As shown in Tab. 3, using 3D keypoint positions (first row) results in a slightly worse performance compared to PolySLGen (second row).

Motion Tokens. Language-model-based motion generation often relies on pre-trained motion tokenizers [5, 8–10, 13], which facilitate adaptation of LLMs to the newly added motion modality. However, the learned codebooks often struggle to generalize to unseen motions due to the limited codebook vocabularies. This limitation reduces both performance and motion diversity, as seen in the third row of Tab. 3 and for SOLAMI in Tab. 1. In contrast, PolySLGen directly learns motion embeddings from global translation and 6D joint rotations, avoiding information loss caused by constrained motion vocabularies.

E.3. Modality Order

For an auto-regressive model such as Llama3-8B-Instruct [2], data ordering introduces modality dependencies. We hypothesize that speech provides a stronger grounding for body movements. Therefore, the text and speech style are processed and generated before body and hand motions. To test this, we also investigate the reverse order, where body and hand motions are processed and generated before speech. As shown in Tab. 4, this alternative (first row) performs worse than PolySLGen (second row), supporting our hypothesis that conditioning motion on accompanying speech improves motion quality.

Table 4. Impact of modality ordering. Diversity closer to the ground truth (117.09) indicates better performance. Root and MPJPE are reported in millimeters (mm). Motion, social cue, and speech are denoted as **M**, **C**, and **S**, respectively.

Modality Order	Motion					Speech			State
	Root↓ (mm)	MPJPE↓ (mm)	FID↓	Div.→	BeatAlign Diff.↓	BERT Score↑	WER↓	SIM↑	AP↑
M → **C** → **S** ‖ **M** → **S**	135.1	154.9	14.23	123.94	0.018	0.502	**1.403**	0.638	0.59
S → **M** → **C** ‖ **S** → **M** (Ours)	**108.7**	**144.9**	**12.18**	**113.32**	**0.007**	**0.508**	1.436	**0.642**	**0.67**

Table 5. Impact of social cue embedding lengths n. Diversity closer to the ground truth (117.09) indicates better performance. Root and MPJPE are reported in millimeters (mm).

	Root↓	MPJPE↓	FID↓	Div.→
n = 1	118.8	152.3	13.56	123.84
n = 4	113.7	149.8	13.04	**120.43**
n = 2 (ours)	**108.7**	**144.9**	**12.18**	113.32

Table 6. Comparison of PolySLGen with two retrieval-based baselines on a non-DM participant.

Non-DM	MPJPE↓	FID↓	BERTScore↑	WER↓	State AP↑
Random	348.4	24.93	0.513	1.423	0.50
NN	329.6	21.27	0.526	1.299	0.50
PolySLGen	**288.6**	**13.54**	**0.543**	**1.251**	**0.74**

Table 7. Long-duration generation via recursive inference over 4 and 8 rounds (∼10.24s and ∼20.48s).

	MPJPE↓	FID↓	BERTScore↑	WER↓	State AP↑
1-round (ours)	**144.9**	**12.18**	**0.508**	**1.436**	**0.67**
4-round	172.5	15.41	0.472	1.468	0.61
8-round	179.5	15.60	0.472	1.487	0.57

Table 8. User study questions for evaluating motion, speech, and overall experience.

Aspect	Statements
Motion Coherence	The motion aligns well with the speech.
Motion Continuity	The movement looks continuous and real.
Speech Semantics	The response is relevant to the current topic of the group.
Speech Tone	The tone of voice is natural.
Overall	The character reacts naturally.

E.4. Social Cue Embedding Length

In PolySLGen, the social cues observed from the past H^m frames are compressed into embeddings with length of two. In Tab. 5, we evaluate the effect of different embedding lengths. The results show that a length of $n = 2$ achieves the best performance on motion quality, while only causing a minor reduction in Diversity.

E.5. Interaction Role Generalization

To evaluate PolySLGen 's generalization across roles, we train it on a non-DM participant that exhibits lower activity levels. As shown in Tab. 6, PolySLGen outperforms the baselines across all metrics, demonstrating its effectiveness for participants with different roles and activity patterns.

E.6. Long-term Generation

We assess long-duration generation through recursive inference over 4 and 8 rounds (∼10.24s and ∼20.48s), where the model's generated motion and speech are fed back as inputs to produce successive reactions. As shown in Tab. 7, the model preserves temporal and social coherence without motion collapse, while exhibiting gradual degradation in MPJPE, FID, and State AP over successive iterations. No catastrophic failure or physically implausible motion is observed in the visualization. We note that this evaluation is conservative, as the behaviors of other participants are kept fixed rather than fully interactive. Nevertheless, the results demonstrate the feasibility of multi-round generation and highlight a promising direction for future work.

E.7. Speaking State and Outputs Alignment

We analyze whether the model learns correlations between speaking states and generated content through a shared latent space without explicit constraints. Empirically, PolySLGen generates an average of 5.07 words in speaking states (ground truth: 6.12) and 0.6 words on average in listening states (ground truth: 0), indicating strong alignment between predicted states and multimodal outputs. For this analysis, we use a threshold of 0.5 to distinguish between speaking and listening states.

E.8. User Study

In Tab. 8, we list the questionnaire items used to evaluate different aspects of human perception. The turn-taking metric for user study was omitted because pilot tests showed it added cognitive load without reliable judgments. Thus, we demonstrate turn-taking in the supplementary video.

References

[1] Hervé Bredin and Antoine Laurent. End-to-end speaker segmentation for overlap-aware resegmentation. *arXiv preprint arXiv:2104.04045*, 2021. 1

[2] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 3, 4

[3] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5152–5161, 2022. 2

[4] Chuan Guo, Xinxin Zuo, Sen Wang, and Li Cheng. Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In *European Conference on Computer Vision*, pages 580–597. Springer, 2022. 2

[5] Jianping Jiang, Weiye Xiao, Zhengyu Lin, Huaizhong Zhang, Tianxiang Ren, Yang Gao, Zhiqian Lin, Zhongang Cai, Lei Yang, and Ziwei Liu. Solami: Social vision-language-action modeling for immersive interaction with 3d autonomous characters. *arXiv preprint arXiv:2412.00174*, 2024. 2, 3, 4

[6] Yinghao Aaron Li, Cong Han, Vinay Raghavan, Gavin Mischler, and Nima Mesgarani. Styletts 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models. *NeurIPS*, 36:19594–19621, 2023. 1

[7] Muhammad Hamza Mughal, Rishabh Dabral, Ikhsanul Habibie, Lucia Donatelli, Marc Habermann, and Christian Theobalt. Convofusion: Multi-modal conversational diffusion for co-speech gesture synthesis. In *CVPR*, pages 1388–1398, 2024. 1, 2, 3, 4

[8] Evonne Ng, Sanjay Subramanian, Dan Klein, Angjoo Kanazawa, Trevor Darrell, and Shiry Ginosar. Can language models learn to listen? In *ICCV*, 2023. 2, 3, 4

[9] Evonne Ng, Javier Romero, Timur Bagautdinov, Shaojie Bai, Trevor Darrell, Angjoo Kanazawa, and Alexander Richard. From audio to photoreal embodiment: Synthesizing humans in conversations. In *CVPR*, pages 1001–1010, 2024.

[10] Jose Ribeiro-Gomes, Tianhui Cai, Zoltán A Milacski, Chen Wu, Aayush Prakash, Shingo Takagi, Amaury Aubel, Daeil Kim, Alexandre Bernardino, and Fernando De La Torre. Motiongpt: Human motion synthesis with improved diversity and realism via gpt-3 prompting. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5070–5080, 2024. 2, 3, 4

[11] stable-ts. stable-ts. https://github.com/jianfch/stable-ts. Accessed: 2025-05-15. 1

[12] Julian Tanke, Linguang Zhang, Amy Zhao, Chengcheng Tang, Yujun Cai, Lezi Wang, Po-Chen Wu, Juergen Gall, and Cem Keskin. Social diffusion: Long-term multiple human motion anticipation. In *ICCV*, pages 9601–9611, 2023. 4

[13] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. T2m-gpt: Generating human motion from textual descriptions with discrete representations. In *CVPR*, 2023. 4

[14] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5745–5753, 2019. 1, 4