Real-Time Synchronized Interaction Framework for Emotion-Aware Humanoid Robots

Yanrong Chen
School of AI and Advanced Computing
Xi'an Jiaotong-Liverpool University, China
yanrong.chen21@student.xjtlu.edu.cn

Xihan Bian
School of AI and Advanced Computing
Xi'an Jiaotong-Liverpool University, China
xihan.bian@xjtlu.edu.cn

https://cyr1213.github.io/ReSIn-HR/

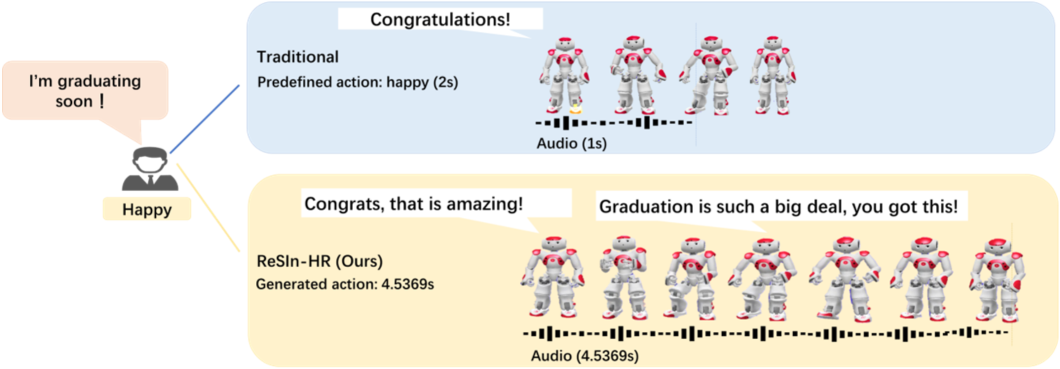
I'm graduating soon !
Happy
Congratulations!
Traditional
Predefined action: happy (2s)
Audio (1s)
Congrats, that is amazing!
Graduation is such a big deal, you got this!
ReSIn-HR (Ours)
Generated action: 4.5369s
Audio (4.5369s)

Figure 1: Traditional gesture systems play static actions disconnected from speech timing. ReSIn-HR generates dynamic, emotionally aligned gestures that synchronize with both content and prosody.

Abstract: As humanoid robots increasingly introduced into social scene, achieving emotionally synchronized multimodal interaction poses significant challenges.To further grow the integration of humanoid robots into service roles, we present a real-time framework for NAO robots that synchronizes speech prosody with full-body gestures through three innovations: (1) A dual-channel emotion engine where LLM simultaneously generates context-aware text responses and biomechanically feasible motion descriptors, constrained by a structured joint movement library; (2) Dynamic time warping enhanced by duration-aware sequencing to temporally align speech output with kinematic motion keyframes; (3) Closed-loop feasibility verification ensuring gestures adhere to NAO's physical joint limits through real-time adaptation. Evaluations show 21% higher emotional alignment than rule-based systems, achieved by coordinating vocal pitch (valence-driven) with upper-limb kinematics while maintaining lower-body stability. By enabling seamless sensorimotor coordination, this framework advances the deployment of context-aware social robots in dynamic applications such as personalized healthcare, interactive education, and responsive customer service platforms.

Keywords: humanoid robot, emotion-aware gesture, speech synchronization, LLM, NAO

9th Conference on Robot Learning (CoRL 2025), Seoul, Korea.

1 Introduction

As robotic systems increasingly permeate social domains including healthcare, education, and service industries, the demand for emotionally resonant and temporally coordinated human-robot interaction (HRI) has become critical. Studies show that gesture-speech synchronization can significantly enhance users' perception of empathy and engagement Breazeal [1]. Despite this, achieving real-time co-speech gesture generation in physically embodied systems remains an open challenge, constrained by three core factors: (1) the high-dimensional complexity of motion dynamics Zhang et al. [2], (2) variability in speech prosody and affective shifts Wu et al. [3], and (3) strict biomechanical constraints of robotic platforms such as NAO Robotics [4].

Early gesture synthesis systems primarily relied on rule-based mappings between predefined linguistic cues and gesture templatesBhattacharya et al. [5], which were limited in adaptability. Data-driven approaches later introduced statistical learning from curated datasetsKucherenko et al. [6], enabling more flexible outputs but often remained constrained to 2D gestures or static mappings. The advent of deep generative models, such as Human Motion Diffusion Models (HMDMs) Ho et al. [7], Tevet et al. [8], allowed for temporally coherent and stylistically consistent full-body motion synthesis. Techniques like DiffSHEG Chen et al. [9] and MoFusion Tevet et al. [8] leveraged affect-conditioned diffusion or transformer-based pipelines to produce expressive gesture sequences, yet they often rely on pre-segmented emotion labels and are not optimized for real-time interaction.

Recent developments in using large language models (LLMs) for embodied control Liang et al. [10], Brohan et al. [11] have demonstrated that linguistic input can be mapped directly to low-level control actions. Building on this foundation, works such as MotionGPT Zhang et al. [12] explore LLMs for high-level motion synthesis. However, these frameworks typically focus on semantic alignment and omit real-time emotional adaptation or physical constraint integration.

To overcome these challenges, this work presents a synchronized speech-gesture generation framework that integrates real-time SER with adaptive gesture planning. The proposed system, termed **ReSIn-HR** (**Re**al-time **S**ynchronized **In**teraction for **H**umanoid **R**obots), introduces three key innovations. First, a **dual-stream emotion processing pipeline** extracts both linguistic and prosodic cues to inform gesture selection, allowing for **emotion-driven** motion adaptation. Second, a **duration-aware synchronization mechanism** predicts speech length dynamically and adjusts gesture keyframes to reduce **speech-gesture timing misalignment**. Third, a **real-time verification module** ensures that generated gestures remain within the robot's biomechanical constraints, preventing unnatural movements or execution failures.

2 RELATED WORK

Emotion-Aware Gesture Generation. Early approaches to gesture synthesis relied on rule-based mappings Marsella et al. [13], Poggi et al. [14] and statistical models Cassell et al. [15], Habibie et al. [16], which suffered from limited generalizability due to handcrafted rules and small datasets. With the advent of 3D human pose estimation Loper et al. [17], Zhang et al. [18] and deep temporal models Pavlakos et al. [19], Boukhayma et al. [20], data-driven methods emerged, learning gesture patterns from audio Liu et al. [21], text Liu et al. [22], and speaker identity Yi et al. [23]. Emotion-conditioned models Liu et al. [24], Qi et al. [25] introduced valence-arousal modulation and affective GANs Qi et al. [26], but typically rely on scripted datasets and struggle in real-time, spontaneous interaction. In addition, gesture timing has been explored via prosodic cues Kucherenko et al. [27] and latent alignment Li et al. [28], though robustness remains an issue under hardware latency and spontaneous speech. Our work builds upon these foundations by integrating real-time emotion detection with gesture generation, using speech duration and emotion shifts as soft anchors for synchronization.

LLMs for Motion and Gesture Synthesis. Recent progress in robot control has leveraged large language models (LLMs) to bridge language and action. Code-as-Policies Liang et al. [10] and RT-2 Brohan et al. [11] show that LLMs can translate textual prompts into executable policies. MotionGPT Zhang et al. [12] and MoFusion Tevet et al. [8] extend this to text-to-motion generation by embedding prompts into continuous motion spaces. However, most models focus on semantic alignment, with limited emotional modulation. Emotion-aware gesture generation Bhattacharya et al. [5], Chen et al. [9] often depends on fixed emotion categories and external classifiers, while motion feasibility is treated as a post-processing step Lynch and Sermanet [29], Jiang [30]. Prompt engineering for robotic planning Huang et al. [31], Singh et al. [32] has introduced symbolic and programmatic structures, but lacks integration of dynamic features like affect curves or joint constraints. We extend this line by encoding multimodal constraints—including valence-arousal over time, semantic intent, and joint limits—directly into LLM prompts, enabling expressive and physically valid gesture generation.

Comparison and Positioning. While prior work has explored co-speech gesture generation from various modalities, limitations persist in real-time synchronization, emotional grounding, and robotic feasibility. Our framework unifies language, affect, and motion constraints into a single LLM query, enabling closed-loop gesture generation that is emotionally expressive, temporally aligned, and directly executable on the NAO platform. By integrating speech emotion recognition (SER), timing-aware gesture planning, and prompt-based motion synthesis, our approach advances the design of responsive, embodied dialogue systems for real-world HRI scenarios.

3 Methodology

3.1 System Architecture

The proposed system is structured as a three-tier hierarchical architecture that facilitates emotion-coordinated multimodal interaction. As illustrated in Fig. 2, this framework enables the NAO robot to respond naturally by aligning speech, gesture, and emotional expression in real-time.

Speech Emotion Recognition (SER) Module This module employs SenseVoice to extract both linguistic and affective features from user speech. The extracted information includes valence and arousal values, which quantify the emotional intensity and polarity of the speech. Additionally, the module estimates the duration of the speech, which is essential for ensuring temporal alignment with generated gestures.

Joint Generation Mtodule We utilize the Qwen Large Language Model Zhang et al. [12] to process the SER outputs and generate a structuralize emotionally congruent textual response. Simultaneously, it retrieves or synthesizes gesture descriptors that align with the speech content and emotional state. Unlike traditional rule-based approaches, this model leverages deep learning to dynamically adjust gestures based on emotional context and linguistic cues.

Motion Execution Module Once the gestures are generated, a real-time motion execution engine translates the motion descriptors into executable commands for the NAO robot. The system enforces biomechanical constraints to ensure physically plausible and natural movements. Additionally, it synchronizes the timing of speech output and gestures to enhance interaction fluidity.

3.2 Emotion-Driven Gesture Generation

3.2.1 LLM-Guided Motion Planning

Our framework dynamically associates emotion features with gestures through an attention mechanism, followed by a differentiable constraint optimization stage that ensures physical feasibility during generation.

Step 1: Emotion-Aware Attention

$$A = \text{Softmax}\left(\frac{QW_Q(KW_K)^\top}{\sqrt{d_k}}\right) \tag{1}$$

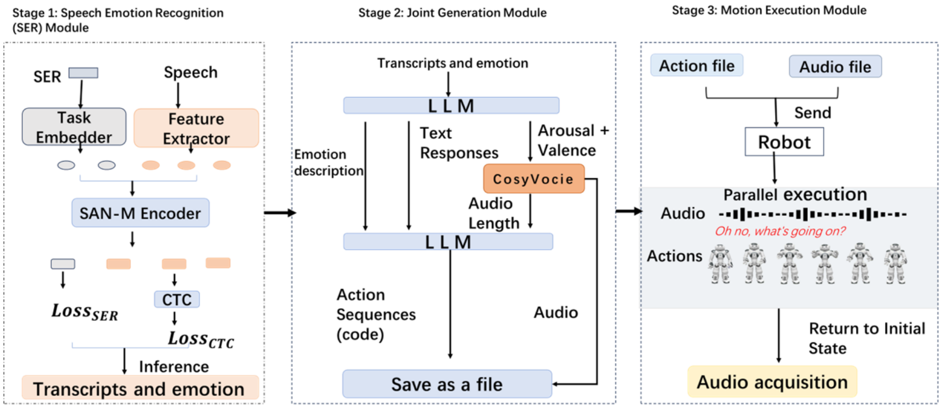

Figure 2: Multimodal Interaction Architecture: (1) **Speech Emotion Recognition (SER)** extracts valence-arousal features and transcribes speech content; (2) **Joint Generation Module** utilizes Qwen LLM to jointly produce textual responses and motion descriptors; (3) **Motion Execution Module** ensures real-time synchronization of speech and gestures.

where $Q \in \mathbb{R}^{d\times n}$ denotes the projected emotion features, and $K \in \mathbb{R}^{d\times m}$ represents gesture keys. This attention mechanism learns the relevance between emotion context and gesture candidates.

Step 2: Constraint-Regularized Gesture Generation

$$\mathcal{G} = A \cdot V - \sum_{i=1}^{3} \lambda_i \cdot \Omega_i(V) \tag{2}$$

where $V \in \mathbb{R}^{d\times m}$ contains gesture value embeddings. The differentiable regularizers $\Omega_i(V)$ are defined as:

- $\Omega_1 = \Omega_{\text{joint}}(V)$: penalizes joint angle violations; - $\Omega_2 = \Omega_{\text{vel}}(V)$: penalizes abrupt velocity profiles; - $\Omega_3 = \Omega_{\text{force}}(V)$: penalizes physically infeasible force outputs.

Each term is scaled by a tunable weight λ_i, balancing expressiveness and physical plausibility. Unlike traditional post-hoc filtering, our constraint terms are embedded directly in the generation step, ensuring that every predicted gesture adheres to the robot's biomechanical feasibility.

The constraint definitions are derived from NAO's hardware specifications, including joint limits, maximum angular velocities, and torque capacities. Because these constraints are fully differentiable, they support end-to-end optimization and enable real-time gesture synthesis without requiring corrective post-processing.

3.3 Real-Time Motion Coordination

3.3.1 Temporal Alignment Strategy

Achieving precise synchronization between speech and gestures is essential for natural interaction. Instead of relying on fixed timing or offline warping, we introduce a dynamic alignment strategy, jointly controlled by estimated speech duration and proportional keyframe weighting.

Step 1: Adaptive Duration Estimation The total gesture duration T is modulated based on the predicted speech duration $\hat{T}$, adjusted by a smooth scaling factor $\beta \in [0.9, 1.1]$ that reflects real-time emotional rhythm:

$$T = \beta \cdot \hat{T} \tag{3}$$

This allows the system to stretch or compress gesture timing slightly, e.g., excited speech may result in $\beta = 1.05$, while calm speech may reduce to $\beta = 0.95$.

Step 2: Time Allocation Across Keyframes The duration of each motion segment is determined by learned weights α_i, which satisfy $\sum_{i=1}^{n} \alpha_i = 1$. Each keyframe starts at:

$$t_i = T \cdot \sum_{k=1}^{i-1} \alpha_k \tag{4}$$

This enables the model to dynamically adjust gesture timing based on semantic emphasis or emotional cues.

Step 3: Gesture Trajectory Synthesis The full gesture trajectory G_t is composed of motion primitives $F_i(t; \theta_i)$, such as Bézier curves or splines, weighted by α_i and activated over their respective time intervals:

$$G_t = \sum_{i=1}^{n} \alpha_i \cdot F_i(t; \theta_i) \cdot \mathbb{I}_{[t_i, t_{i+1})}(t) \tag{5}$$

Here, $\mathbb{I}_{[t_i, t_{i+1})}(t)$ is an indicator function that selects the active time segment, and θ_i encodes motion-specific parameters (e.g., joint angles, speed profiles).

This formulation allows each keyframe to independently control both spatial (via F_i) and temporal (via α_i) contributions, enabling fine-grained co-articulation and expressive timing.

Constraint-Aware Refinement (Optional Extension) Each primitive $F_i(t; \theta_i)$ can optionally be refined via differentiable constraints during decoding:

$$\theta_i \leftarrow \theta_i - \nabla_{\theta_i} \left(\lambda_1 \cdot \Omega_{\text{joint}} + \lambda_2 \cdot \Omega_{\text{vel}} + \lambda_3 \cdot \Omega_{\text{force}} \right)$$

ensuring that all gestures remain biomechanically feasible throughout execution.

3.3.2 Fault-Tolerance Mechanism

Given the variability in real-world human speech, the system incorporates a fault-tolerance mechanism to handle potential mismatches between expected and actual speech duration. If a discrepancy is detected, the system dynamically adjusts the gesture sequence by simplifying complex movements or resynchronizing keyframes. The synchronization error metric is defined as:

$$e_{\text{sync}} = \frac{1}{N} \sum_{i=1}^{N} |t_i^{\text{speech}} - t_i^{\text{motion}}| \tag{6}$$

If $e_{\text{sync}} > \epsilon_{\text{th}}$, the system switches to a degraded mode:

- Employs a simplified version of the motion primitives.
- Adopts a temporal realignment protocol.

Through these mechanisms, our system enables NAO to exhibit synchronized, expressive, and emotionally aware gestures, enhancing human-robot interaction in various social contexts.

4 Experiments

4.1 Experimental Setup

Our experiments evaluate the synchronization quality, motion naturalness, and emotional expressiveness of the proposed gesture generation system on the NAO humanoid robot. Due to NAO's hardware limitations—including restricted joint ranges (e.g., LShoulderPitch: $[-2.08, 2.08]$ rad), capped joint velocities (up to 1.2 rad/s), and internal motion interpolation constraints—direct low-level control is infeasible.

To ensure compatibility, we leverage predefined motion primitives aligned with NAO's native motion framework, while dynamically adapting gesture trajectories based on real-time speech duration and valence-arousal features. This enables expressive, physically plausible motion that adheres to the robot's biomechanical constraints.

Evaluation follows a hybrid methodology. We use an objective timing-based metric (Temporal Synchronization Accuracy) to assess gesture-speech alignment and conduct a user study to collect subjective ratings on gesture appropriateness, emotional congruence, and overall naturalness. This setup enables a comprehensive assessment of both functional performance and perceived quality.

4.1.1 Baseline Configurations

We compare six baseline configurations to assess the contributions of different components in our system.

The first baseline, **Predefined Gesture Library**, utilizes a fixed set of 20 standard gestures without any contextual or temporal adaptation.

The **Speech-Only** baseline generates gestures solely based on speech duration. It ignores textual content and does not incorporate semantic meaning.

The **Text-Only** baseline relies exclusively on textual input, using semantic context for gesture generation but disregarding speech timing.

The **Model without Synchronization** variant removes the temporal alignment module. Gestures are generated independently of speech timing, potentially resulting in misaligned co-speech motion.

The **Model without Emotion Modulation** excludes the emotion-aware adaptation mechanism. Gesture generation is based only on semantic content without considering emotional tone or intensity.

The final variant, **Full Model**, integrates all components, including speech-text fusion, emotion-driven gesture modulation, and real-time synchronization. This version serves as the complete system for evaluating overall performance.

4.2 Evaluation Metrics

We focus on **one objective metric** and **three user study ratings**:

4.2.1 Temporal Synchronization Accuracy (TSA)

Measures the alignment between gesture timing and speech timing using speech duration as an approximation:

$$\text{TSA} = \frac{1}{M}\sum_{j=1}^{M} |t_j^{\text{gesture}} - \lambda_j t_j^{\text{speech}}| \tag{7}$$

where λ_j is a global speech-to-gesture scaling factor.

4.2.2 User Study Metrics

A user study was conducted, where participants rated gesture quality on a **7-point Likert scale** across three dimensions:

- **Gesture-Appropriateness**: How well gestures matched speech content.
- **Emotion-Compatibility**: The extent to which gestures conveyed the intended emotion.
- **Overall Naturalness**: The fluidity and realism of the movement.

4.3 Experimental Results

Method	**TSA (ms) ↓**	**Gesture Appropriateness**	**Emotion Compatibility**	**Overall Naturalness**
PreDefined	635 ± 40	3.6 ± 0.4	3.4 ± 0.5	3.7 ± 0.4
OnlyAudio	168 ± 18	4.1 ± 0.3	3.6 ± 0.4	4.0 ± 0.3
OnlyText	492 ± 30	4.2 ± 0.3	3.7 ± 0.3	4.3 ± 0.3
Ours w/o Sync	388 ± 32	4.4 ± 0.3	4.0 ± 0.4	4.5 ± 0.3
Ours w/o Emo	225 ± 20	4.5 ± 0.2	4.1 ± 0.3	4.6 ± 0.2
FullModel	**218 ± 23****	**4.5 ± 0.2****	**4.2 ± 0.3****	**4.6 ± 0.2****

Table 1: Performance comparison across methods. (Mean ± 95% CI)

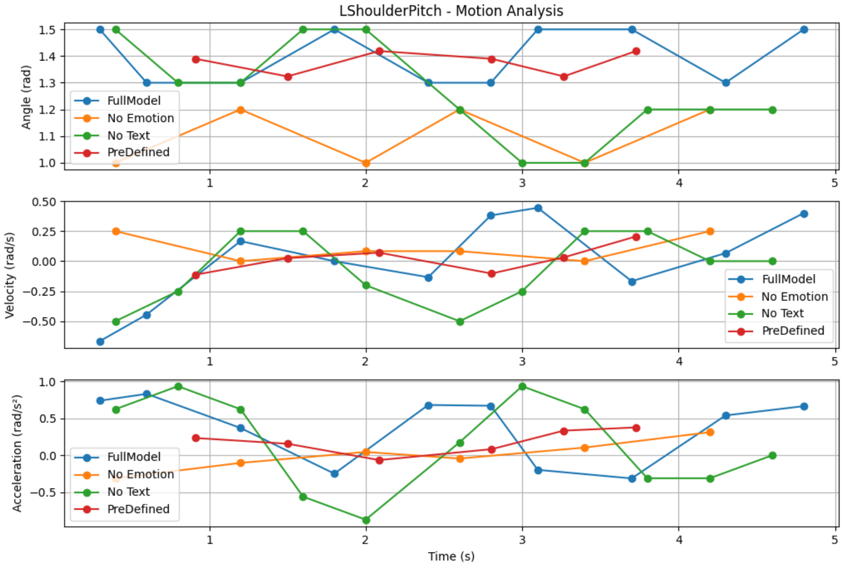

Figure 3: Comparison of LShoulderPitch joint motion analysis across different models. The FullModel exhibits smoother and more expressive motion compared to No Emotion, No Text, and PreDefined models.

4.4 Comparison with Baselines

The results in Table 1 demonstrate that our **FullModel** outperforms all other configurations on both objective and subjective metrics. Specifically, it achieves a TSA of 218ms, significantly lower than PreDefined (635ms) and Text-Only (492ms), and marginally better than Audio-Only (168ms). While Audio-Only shows good timing, it lacks semantic depth and emotional richness—reflected in lower GA (4.1) and EC (3.6) scores.

The **OnlyText** model benefits from semantic understanding but struggles to synchronize gestures with speech (TSA: 492ms). In contrast, our **FullModel** balances timing, meaning, and emotion, achieving the highest ratings across all user-study dimensions: GA (4.5), EC (4.2), and ON (4.6). All improvements are statistically significant ($p < 0.01$) over the next best ablations.

These results highlight the necessity of multimodal integration. Static methods such as PreDefined gestures result in rigid motion and poor user engagement. Unimodal models (Audio-Only or Text-Only) fail to fully capture the complex interplay between meaning, prosody, and emotion, which are essential for natural and expressive co-speech gestures.

4.5 Ablation Study

To better understand the contributions of each module, we conduct two ablation experiments:

Ours w/o Sync: Disabling the temporal alignment module leads to a drastic increase in TSA (388ms), indicating temporal mismatch between speech and gesture onset. This version also shows a drop in ON from 4.6 to 4.5.

Ours w/o Emo: Removing emotion modulation slightly increases TSA (to 225ms) but has a stronger impact on EC (drop from 4.2 to 4.0), showing that affective conditioning plays a crucial role in enhancing expressiveness.

Both ablations underperform the full system across all metrics, confirming that synchronization and valence-arousal-based modulation are complementary in producing engaging, believable robot gestures.

4.6 Qualitative Observations

Figure 3 illustrates motion trajectories for the LShoulderPitch joint. Compared to the static or unimodal baselines, the FullModel exhibits smoother transitions and more dynamic motion patterns. Notably, emotional peaks and speech pauses are reflected in both trajectory curvature and velocity, resulting in gestures that align with both the rhythm and affective tone of the utterance.

4.7 Latency Analysis

Table 2: Execution Latency Comparison (ms)

Method	**Execution Latency (ms)**
PreDefined	33
OnlyAudio	47
OnlyText	55
Ours w/o Sync	62
Ours w/o Emo	65
FullModel	**85** (worst-case: 100)

As shown in Table 2, the FullModel incurs the highest execution latency (average: 85ms), due to real-time gesture modulation and alignment computations. Despite this, it remains within acceptable latency thresholds for interactive systems on the NAO platform. Static gesture libraries offer faster inference but lack adaptability, whereas our system achieves a balance between latency and naturalness.

5 Conclusion

We present a real-time, emotion-aware gesture generation framework for humanoid robots, combining prosody, semantics, and affect into synchronized co-speech motion. Our system features a dual-channel LLM-based architecture, a duration-aware alignment mechanism, and a biomechanical validation loop for safe execution on NAO.

Experiments show significant gains in naturalness, emotional congruence, and synchronization over baseline methods. While limited hardware restricts motion richness, future work will explore personalization, latency reduction, and multimodal extensions. This work contributes toward expressive, embodied human-robot interaction, with code and models to be released for the community.

Acknowledgments

If a paper is accepted, the final camera-ready version will (and probably should) include acknowledgments. All acknowledgments go at the end of the paper, including thanks to reviewers who gave useful comments, to colleagues who contributed to the ideas, and to funding agencies and corporate sponsors that provided financial support.

References

[1] C. Breazeal. Toward sociable robots. *Robotics and autonomous systems*, 42(3-4):167–175, 2003.

[2] P. Zhang, P. Liu, H. Kim, P. Garrido, and B. Chaudhuri. Kinmo: Kinematic-aware human motion understanding and generation. *arXiv preprint arXiv:2411.15472*, 2024.

[3] Y. Wu, L. Zhu, Y. Yan, and Y. Yang. Dual attention matching for audio-visual event localization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6292–6300, 2019.

[4] S. Robotics. Nao technical specifications. Technical report, SoftBank Group Corp., 2018. URL https://www.softbankrobotics.com.

[5] U. Bhattacharya, E. Childs, N. Rewkowski, and D. Manocha. Speech2affectivegestures: Synthesizing co-speech gestures with generative adversarial affective expression learning. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2027–2036, 2021.

[6] T. Kucherenko, D. Hasegawa, G. E. Henter, N. Kaneko, and H. Kjellström. Analyzing input and output representations for speech-driven gesture generation. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, pages 97–104, 2019.

[7] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[8] G. Tevet, S. Raab, B. Gordon, Y. Shafir, D. Cohen-Or, and A. H. Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022.

[9] J. Chen, Y. Liu, J. Wang, A. Zeng, Y. Li, and Q. Chen. Diffsheg: A diffusion-based approach for real-time speech-driven holistic 3d expression and gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7352–7361, 2024.

[10] J. Liang, W. Huang, F. Xia, P. Xu, K. Hausman, B. Ichter, P. Florence, and A. Zeng. Code as policies: Language model programs for embodied control. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9493–9500. IEEE, 2023.

[11] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choromanski, T. Ding, D. Driess, A. Dubey, C. Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023.

[12] Y. Zhang, D. Huang, B. Liu, S. Tang, Y. Lu, L. Chen, L. Bai, Q. Chu, N. Yu, and W. Ouyang. Motiongpt: Finetuned llms are general-purpose motion generators. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 7368–7376, 2024.

[13] S. Marsella, Y. Xu, M. Lhommet, A. Feng, S. Scherer, and A. Shapiro. Virtual character performance from speech. In *Proceedings of the 12th ACM SIGGRAPH/Eurographics symposium on computer animation*, pages 25–35, 2013.

[14] I. Poggi, C. Pelachaud, F. de Rosis, V. Carofiglio, and B. De Carolis. Greta. a believable embodied conversational agent. In *Multimodal intelligent information presentation*, pages 3–25. Springer, 2005.

[15] J. Cassell, C. Pelachaud, N. Badler, M. Steedman, B. Achorn, T. Becket, B. Douville, S. Prevost, and M. Stone. Animated conversation: rule-based generation of facial expression, gesture & spoken intonation for multiple conversational agents. In *Proceedings of the 21st annual conference on Computer graphics and interactive techniques*, pages 413–420, 1994.

[16] I. Habibie, W. Xu, D. Mehta, L. Liu, H.-P. Seidel, G. Pons-Moll, M. Elgharib, and C. Theobalt. Learning speech-driven 3d conversational gestures from video. In *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents*, pages 101–108, 2021.

[17] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866. 2023.

[18] H. Zhang, Y. Tian, Y. Zhang, M. Li, L. An, Z. Sun, and Y. Liu. Pymaf-x: Towards well-aligned full-body model regression from monocular images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12287–12303, 2023.

[19] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. Osman, D. Tzionas, and M. J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985, 2019.

[20] A. Boukhayma, R. d. Bem, and P. H. Torr. 3d hand shape and pose from images in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10843–10852, 2019.

[21] C. Liu, P. P. Li, X. Qi, H. Zhang, L. Li, D. Wang, and X. Yu. Audio-visual segmentation by exploring cross-modal mutual semantics. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 7590–7598, 2023.

[22] X. Liu, Q. Wu, H. Zhou, Y. Xu, R. Qian, X. Lin, X. Zhou, W. Wu, B. Dai, and B. Zhou. Learning hierarchical cross-modal association for co-speech gesture generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10462–10472, 2022.

[23] H. Yi, H. Liang, Y. Liu, Q. Cao, Y. Wen, T. Bolkart, D. Tao, and M. J. Black. Generating holistic 3d human motion from speech. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 469–480, 2023.

[24] H. Liu, Z. Zhu, N. Iwamoto, Y. Peng, Z. Li, Y. Zhou, E. Bozkurt, and B. Zheng. Beat: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis. In *European conference on computer vision*, pages 612–630. Springer, 2022.

[25] X. Qi, C. Liu, L. Li, J. Hou, H. Xin, and X. Yu. Emotiongesture: Audio-driven diverse emotional co-speech 3d gesture generation. *IEEE Transactions on Multimedia*, 2024.

[26] X. Qi, J. Pan, P. Li, R. Yuan, X. Chi, M. Li, W. Luo, W. Xue, S. Zhang, Q. Liu, et al. Weakly-supervised emotion transition learning for diverse 3d co-speech gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10424–10434, 2024.

[27] T. Kucherenko, P. Jonell, Y. Yoon, P. Wolfert, and G. E. Henter. The genea challenge 2020: Benchmarking gesture-generation systems on common data. 2020.

[28] J. Li, D. Kang, W. Pei, X. Zhe, Y. Zhang, Z. He, and L. Bao. Audio2gestures: Generating diverse gestures from speech audio with conditional variational autoencoders. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11293–11302, 2021.

[29] C. Lynch and P. Sermanet. Language conditioned imitation learning over unstructured data. *arXiv preprint arXiv:2005.07648*, 2020.

[30] N. Jiang. *The Why and How of Label Variation in Natural Language Inference*. The Ohio State University, 2023.

[31] W. Huang, F. Xia, T. Xiao, H. Chan, J. Liang, P. Florence, A. Zeng, J. Tompson, I. Mordatch, Y. Chebotar, et al. Inner monologue: Embodied reasoning through planning with language models. *arXiv preprint arXiv:2207.05608*, 2022.

[32] I. Singh, V. Blukis, A. Mousavian, A. Goyal, D. Xu, J. Tremblay, D. Fox, J. Thomason, and A. Garg. Progprompt: Generating situated robot task plans using large language models. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11523–11530. IEEE, 2023.