RoboGesture: Real-Time Semantic-aligned Co-Speech Gestures Generation for Humanoid Interaction

Zifan Wang[*1,2], Ziang Ren[*2,3], Pengyang Shi[*1,2], Zirui Wang[4], Chenghuai Lin[2], Tianze Wang[2], Zekun Qi[1,2], Liangliang Zhao[4], He Wang[2,5], and Li Yi[✉1,2,6]

[1] Tsinghua University
[2] Galbot Inc.
[3] Beijing Institute of Technology
[4] Harbin Institute of Technology
[5] Peking University
[6] Shanghai Qi Zhi Institute
https://RoboGesture.github.io

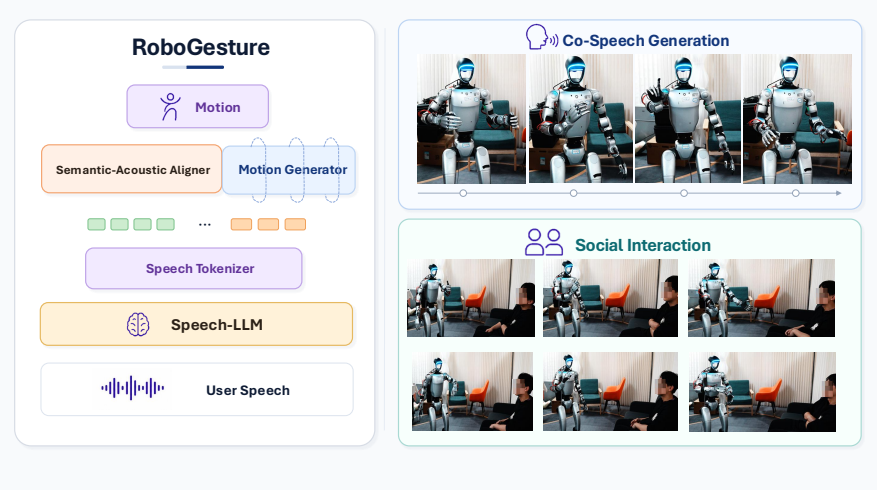

Fig. 1: RoboGesture generates expressive, semantically-aligned, and safety-aware co-speech gestures for humanoid robots from streaming speech. This framework enable robots to listen, respond, and gesture coherently in real-time social interaction.

Abstract. Enabling humanoid robots to respond to human speech with synchronized and semantically meaningful gestures is fundamental to natural human-robot interaction. However, this task faces three critical barriers: the scarcity of semantically rich datasets, the "modality eclipse" where models ignore audio cues in favor of kinematic inertia, and the sim-to-real gap regarding physical safety. We propose RoboGesture, a robot-centric framework that co-designs data, modeling, and control to power

[*] Equal contribution. [✉] Corresponding author.

a complete interactive human–humanoid system in which the robot listens, responds, and gestures in real time. We first establish the RoboGesture dataset featuring over 300 gesture categories and develop an automated pipeline to synthesize large-scale collision-free, robot-specific audio-motion pairs. Our architecture features a Hierarchical Semantic-Acoustic Aligner that extracts multi-granular prosodic and semantic cues directly from raw audio tokens. These cues drive a Streaming Conditional Motion Generator based on a diffusion transformer with conditional flow matching. To ensure high responsiveness, we introduce Anti-Inertia CFG Masking, which prevents the model from collapsing into repetitive historical patterns by compelling it to proactively mine control signals from the audio modality. Finally, an MPC-based safety filter ensures real-time, collision-free execution on physical hardware. Experiments on a Unitree G1 humanoid demonstrate that RoboGesture generates safer, more rhythmic, and more semantically appropriate responses compared to state-of-the-art baselines.

1 Introduction

The ability of a humanoid robot to *listen* to human speech and *respond* with synchronized, semantically meaningful audio and body motion lies at the heart of natural human-robot interaction (HRI). Such capability elevates robots from mere tools to genuine partners in assistive care, education, and collaboration. This work pursues a complete system that endows a humanoid with expressive, embodiment-safe, and real-time audio-motion responses—making face-to-face conversation with machines not only possible but intuitive and engaging.

Realizing this vision, however, confronts three entrenched barriers. First, the scarcity of large-scale, semantically rich audio-motion datasets with fine-grained annotations [14, 22] limits models' capacity to establish semantic alignment between speech and gestures. Existing methods therefore resort to post-processing heuristics or text-based intermediate representations [45, 48, 49], yet these compromises inherently sacrifice semantic and temporal coherence in the generated motion. Second, these limitations become critically exacerbated in online generation settings: without explicit modeling of audio semantics and controllable generation mechanisms, models tend to collapse into motion shortcuts—excessively replicating historical motion patterns rather than responding to incoming acoustic cues. Third, the sim-to-real gap between virtual avatars and physical humanoids remains substantial. Directly retargeting generated human motions to humanoid robots via online retargeting and PD control often yields unstable, unsafe behaviors, including jerky trajectories and self-collisions.

In this work, we aim to design a robot-centric, data-model co-designed framework by tackling the above challenges. As shown in Figure 1, we present a comprehensive solution that enables **real-time**, **semantically aligned**, and **safety-aware** humanoid co-speech gesture synthesis in an **end-to-end manner**, and serves as the motion core of a complete interactive human–humanoid system in which the robot listens, responds, and gestures.

To address the scarcity of large-scale, tightly aligned audio-motion data, we build a large-scale, semantically rich dataset by capturing a library of expressive body and hand gestures guided by an established gesture taxonomy. We then develop an automatic synthesis pipeline to generate vast quantities of paired audio and motion sequences. Critically, this pipeline incorporates an offline motion retargeting step to map all human motions to our target humanoid's kinematics, followed by an MPC-based filter that vets the entire dataset to remove any potential self-collisions. This process yields a clean, extensive, and inherently safe training corpus of robot-specific motions.

To better capture temporal alignment between raw audio and semantics, we develop a hierarchical semantic-acoustic aligner. Unlike text-based methods that lose vital prosodic cues like intonation and emphasis, our aligner processes raw audio through a multi-level transformer architecture. By decoupling audio into streaming low-level rhythmic pulses and high-level semantic labels, the aligner is able to extract "pre-phonetic" anticipatory signals—such as the body's energy accumulation before a shout—ensuring the robot's movements are synchronized with both the beat and the intent of the speech.

To ensure these cues effectively drive the robot and prevent it from collapsing into history motion shortcut, we develop a streaming conditional motion generator. We observe that models often "cheat" by relying excessively on past kinematic inertia, effectively ignoring the weaker audio and semantic guidance. To break this dependency, we introduce a two-stage training strategy with anti-inertia masking, which randomly masks historical context to compel the diffusion-based motion generator to proactively mine control signals from the audio modality. By integrating FiLM for global semantic tone and Cross-Attention for local rhythmic alignment, our framework enables seamless, end-to-end streaming from raw audio to safe, high-fidelity robot trajectories.

We validate our framework through extensive quantitative metrics and qualitative user studies focusing on motion quality, synchrony, naturalness, and safety. Furthermore, we demonstrate the system's efficacy through real-world social interactions on a physical humanoid robot. Our approach consistently generates behaviors with superior semantic and rhythmic alignment compared to existing baselines. Our contributions are:(1) A robot-centric, data-model co-designed framework for an interactive human–humanoid system that enables end-to-end, streaming generation of real-time, semantically aligned, and safety-aware co-speech gestures, and that we deploy as a complete listen–respond–gesture loop on a physical humanoid. (2) A robust data synthesis pipeline that produces a large-scale, semantically rich, and collision-free training corpus by mapping expressive human gestures to robot-specific kinematics. (3) A continuous audio-driven motion policy to ensure tight temporal synchronization and prevent the model from collapsing into repetitive historical motion patterns.

Table 1: Comparison with related methods.*Streaming*: whether the model has a design for streaming support. *Semantics*: whether the model focuses on modeling special semantic gestures. *Text-free*: whether the model works without text as input or intermediate modality. *Hand*: whether hand modeling is considered. *Motion*: motion generation approach.

Model	Streaming?	Semantics?	Text-free?	Hand?	Motion?
LivelySpeaker [49]	✗	✓	✗	✗	Continuous
DiffSHEG [8]	✓	✗	✓	✓	Continuous
Semantic Gesticulator [48]	✗	✓	✗	✓	Discrete
SemTalk [45]	✗	✓	✗	✓	Discrete
Ours	✓	✓	✓	✓	Continuous

2 Related Works

2.1 Human-Humanoid Social Interaction

Social interactions between humans and humanoid robots have gained increasing attention [7,11,13,25,35–37], driven by advances in both robotic embodiment and social intelligence modeling. Many works [2,5,16,39] have addressed social interaction problems through virtual avatars. For instance, SOLAMI [16] enables immersive interaction with 3D autonomous characters and aspires to bridge toward robotic embodiments. However, a significant domain gap exists between avatar environments and real robots. From the embodiment perspective, physical humanoids face hardware, safety, and real-time constraints, whereas virtual-avatar environments do not suffer from these limitations. A core challenge in this direction is retargeting human motions to humanoid robots: classical pipelines [41] first learn models in human representation and then retarget to the robot, or learns human-to-robot mappings directly [24]. Despite progress in high-fidelity retargeting [3, 38], balancing real-time performance, fine-grained gesture detail, and safety remains difficult in social scenarios. Motivated by these limitations, we instead learn directly in the robot's representation, addressing cross-embodiment alignment as a pre-processing rather than post-processing problem.
For social intelligence modeling on real-world robots, traditional systems rely on rule-based or policy-structured controllers, including behavior trees and handcrafted preconditions [4, 33, 34]. While interpretable, such pipelines struggle to capture the open-ended complexity of real social behavior, and often decouple modalities—such as speech and gesture that humans naturally coordinate. In contrast, we develop a multimodal end-to-end model trained directly in robot space, enabling natural, coherent social interaction on a real humanoid platform.

2.2 Co-speech Gesture Generation

Gestures are integral nonverbal and non-manipulative body movements that enhance human communication [26]. Research in co-speech gesture generation has transitioned from early rule-based [17] and statistical machine learning methods [18] to modern deep learning frameworks [8, 9, 23, 40, 43, 44, 46], fueled by

the availability of large-scale audio-gesture datasets [14, 22]. As summarized in Table 1, recent efforts prioritize bridging gestures with linguistic content through diverse alignment strategies. LivelySpeaker [49] utilizes CLIP [32] to fuse rhythmic and semantic cues, yet it often fails to maintain semantic consistency. DiffSHEG [8] employs global semantic descriptors for alignment but lacks the precision required for fine-grained contexts. While Semantic Gesticulator [48] achieves high-fidelity semantic correspondence via LLM-based retrieval, its dependence on post-hoc alignment hinders real-time streaming applications. Similarly, SemGes [23] introduces coherence and relevance losses to ground semantics; however, its performance is constrained by sparse semantic annotations. SemTalk [45] attempts to decouple general and sparse motions, but its reliance on text-based features makes capturing precise temporal alignment with audio cues challenging. To address these limitations, we propose an automatic, semi-synthetic data generation pipeline coupled with a two-stage training strategy. Our approach enables end-to-end synchronization between speech and expressive gestures by learning semantic cues directly from audio. This text-free design inherently supports low-latency streaming settings while ensuring robust cross-modal alignment.

3 Method

The primary objective of this work is to achieve **real-time, semantically aligned, and safety-aware** co-speech gesture generation that drives an interactive human–humanoid system on physical hardware. Our framework operates within a streaming-to-streaming paradigm: it processes incoming audio chunks and generates corresponding motions in real-time. This architecture ensures seamless integration with high-performance streaming Large Language Models and audio feeds, thereby facilitating natural human-robot social interaction. For the physical embodiment, we employ the Unitree G1 humanoid robot integrated with BrainCo dexterous hands. Our motion generation focuses on upper-body dynamics, utilizing a kinematic state space of 41 degrees of freedom (DoFs): 17 DoFs for the upper torso and arms, and 24 DoFs for the dexterous hands. This configuration provides the requisite mechanical flexibility for generating expressive, human-like gestures.

Our solution is a two-stage training framework structured into three core modules as shown in Figure 2: (i) a hierarchical semantic-acoustic aligner, (ii) a streaming conditional motion generator, and (iii) an MPC-based kinematic safety filter. Once trained, the framework enables seamless, end-to-end streaming from raw audio to robot trajectories. Section 3.1, 3.2 and 3.3 describe the architectural details, while Section 3.4 outlines the training objectives.

3.1 Hierarchical Semantic-Acoustic Aligner

While recent approaches attempt to extract semantic cues from intermediate modalities such as text [45,48,49], they often struggle to achieve precise temporal

alignment between motion and subtle acoustic nuances. On one hand, converting audio to text inevitably leads to the loss of essential prosodic information, such as intonation and emphasis. For example, the word 'Really' conveys skepticism with a rising intonation but confirmation with a falling one—crucial distinctions embedded in raw audio that text-based methods fail to capture. On the other hand, co-speech gestures often exhibit pre-phonetic anticipation; for example, before articulating an emphatic verb like 'Stop!', the body typically initiates energy accumulation, such as leaning back or raising a hand. Such preparatory signals are hidden in the subtle acoustic precursors immediately preceding vocalization. To address these limitations, we propose to fully leverage multi-granular audio features through a semantic-acoustic aligner.

To get multi-level acoustic features, we utilize the Mimi codec [12] to tokenize the streaming audio. Unlike traditional codecs, Mimi explicitly decouples audio chunks into a hierarchical representation: the first quantizer provides high-level semantic tokens; while the subsequent residual quantizers capture fine-grained acoustic details such as prosody and intonation.

To harness these multi-granular cues, we introduce a transformer-based semantic-acoustic aligner optimized via multi-task auxiliary learning. Shallow layers are designed to capture rapid acoustic energy transients. We extract low-level features from these layers and route them to a beat head for auxiliary rhythmic supervision. This ensures the generated motion is precisely synchronized with acoustic onsets. Conversely, deep layers distill macroscopic semantics via a 300-class discrete classification task powered by RoboGesture dataset in Section 4. By processing high-level features through a semantic head, the architecture acts as an information bottleneck, compressing acoustic nuances into actionable tokens that retain prosodic accuracy often lost in raw text.

Ultimately, this hierarchical disentanglement allows the aligner to provide distinct yet complementary control signals. The low-level rhythmic pulses and high-level semantic labels work in tandem to guide the downstream motion generator.

3.2 Streaming Conditional Motion Generator

Existing *Motion Tokenizer + Autoregressive (AR)* frameworks often struggle to capture high-fidelity kinematics, as discrete tokens lack the precision required for expressive hand gestures. Furthermore, AR models are prone to recursive error accumulation, compromising long-term semantic coherence. To overcome these limitations, we employ a Diffusion Transformer (DiT) [29] operating in a continuous state space, parameterized by Conditional Flow Matching (CFM) [21].

However, a critical challenge in this pipeline is the modality eclipse: the generator may rely excessively on historical kinematic inertia, producing motions that are plausible but audio-unaware. To ensure the gestures remain strictly driven by the streaming audio while maintaining physical continuity, we propose a dual condition injection mechanism: **(1) Framewise Micro-alignment:** We utilize cross-attention to fuse multi-scale acoustic cues with historical motion context. In this process, the shallow features that capture rhythmic transients

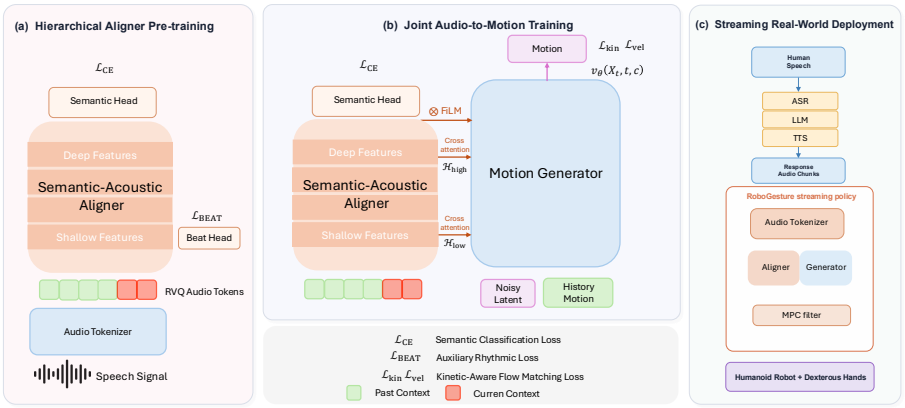

Fig. 2: Overview of RoboGesture. **(a)** The Semantic-Acoustic Aligner is pre-trained to decouple multi-granular audio cues from streaming tokens. **(b)** The Motion Generator synthesizes motions via Conditional Flow Matching, jointly conditioned on hierarchical audio features and historical motion to maintain autoregressive consistency. **(c)** During inference, the pipeline seamlessly integrates with upstream Speech-LLMs and employs a safety filter to drive physical humanoid robots in real-time.

and the deep features that provide local semantic guidance are extracted from the aligner. These multi-level sequences, alongside the *Past Motion* chunk, are projected as Key-Value pairs. By explicitly injecting the historical kinematic state and granular acoustic signals through the attention mechanism, the model ensures seamless transitions and precise temporal synchronization across consecutive generation windows. **(2) Global Macro-modulation:** In contrast, the highly compressed semantic instructions derived from the aligner serve as macroscopic behavioral baselines. To prevent these global intents from being overshadowed by local kinematic noise, we employ Feature-wise Linear Modulation (FiLM) [30]. FiLM applies affine transformations to the DiT's intermediate feature maps, robustly establishing the overarching emotional and semantic tone without interfering with intricate local alignments.

3.3 MPC-based Collision-avoidance Filter

While training our framework directly within the robot's action space significantly mitigates the artifacts typical of manual retargeting, the stochastic nature of generative models cannot inherently guarantee absolute physical safety.
To ensure collision-free and physically feasible deployment, we pass the synthesized motion chunks through an MPC-based optimization module in the inference time. This module acts as a safety filter by solving a constrained optimization problem in real-time. It enforces collision constraints together with velocity and position tracking constraints, which we cast as a convex quadratic program

with velocity, trajectory-tracking, and smoothing costs, solved per frame with OSQP in a streaming manner. This filter reduces the self-collision frame ratio of the generated motions from 4.16% to 0.13%. We provide the full formulation and statistics in the supplementary material.

3.4 Training Recipes and Objectives

To develop a robust aligner and a highly responsive motion generator, we adopt a hierarchical two-stage training strategy as shown in Figure 2(a)(b). Specifically, we define a temporal horizon of $T = 30$ frames for each one-second action chunk, operating in a robot action space of $D = 41$ dimensions. To incorporate sufficient context, we introduce a historical window of $T_{hist} = 2T$ frames.

Stage 1: Representation Disentanglement Pre-training In the initial stage, the DiT generator remains frozen. We exclusively optimize the aligner using 1,000 hours of semantic-annotated chunk-level semi-synthetic data, as detailed in Section 4. To capture temporal context, the input consists of a $3T$-frame audio window, comprising the current chunk and a historical window. Intermediate features from the shallow layers are routed to a beat head. The rhythmic target $B_{target} \in \mathbb{R}^{T\times 2}$ is a concatenation of the normalized physical velocity magnitude from the ground-truth motion and the acoustic onset strength:

$$B_{target} = [\text{Norm}(\text{Resample}(\|\dot{q}\|_2)); \text{Norm}(\text{Onset}(wav))] \tag{1}$$

where $\dot{q}$ represents the joint velocities. The deep layers are routed to a semantic head to predict discrete action categories Y_{target}. The Stage 1 objective is:

$$\mathcal{L}_{Stage1} = \lambda_{beat} \cdot \text{MSE}(\hat{B}, B_{target}) + \text{CE}(\hat{Y}, Y_{target}) \tag{2}$$

where λ_{beat} balances the regression and classification gradients.

Stage 2: Joint Generative Training with CFG Masking In the second stage, we unfreeze the DiT to jointly optimize motion synthesis and semantic preservation. To balance expressive semantics with rhythmic alignment, we utilize a composite dataset blending the 1000-hour semantic-centric semi-synthetic data with an upsampled (3×) beat-centric dataset (BEAT [22], about 76 hours). To break the "modality eclipse", we introduce Anti-Inertia CFG Masking: the *Past Motion* condition is randomly masked with a 15% probability, compelling the network to perform "cold-starts" driven by aligner features.

To prevent fine-grained hand kinematics from being submerged, we define a spatial weight $W_s \in \mathbb{R}^D$ ($w_{hand} = 4.0$ for hand joints) and a temporal weight $W_t \in \mathbb{R}^T$ ($w_{frame} \in \{5, 10\}$ for frames within semantic intervals). The core generative process is supervised by a spatio-temporally weighted Velocity Matching Loss:

$$\mathcal{L}_{vel} = \mathbb{E}_{\tau,M_1,M_0}\left[\sum_{t=1}^{T}\sum_{d=1}^{D} W_t^{(t)} W_s^{(d)} \left(v_\theta(M_\tau, \tau, C)^{(t,d)} - (M_1 - M_0)^{(t,d)}\right)^2\right] \tag{3}$$

where $\tau \sim \mathcal{U}(0,1)$ is the flow matching time step, $M_0 \sim \mathcal{N}(0, I)$ is the standard Gaussian noise, M_1 is the target motion, and $M_\tau = (1-\tau)M_0 + \tau M_1$ is the intermediate interpolated state. To suppress high-frequency jittering, we explicitly penalize the first-order temporal difference. Using a single-step Euler approximation $\hat{M}_1 = M_\tau + (1-\tau)v_\theta(M_\tau, \tau, C)$, we introduce a Kinetic Consistency Loss ($\mathcal{L}_{kin}$):

$$\mathcal{L}_{kin} = \mathbb{E}_{\tau, M_1, M_0}\left[\sum_{t=1}^{T-1}\sum_{d=1}^{D} \bar{W}_t^{(t)} W_s^{(d)} \left(\Delta\hat{M}_1^{(t,d)} - \Delta M_1^{(t,d)}\right)^2\right] \tag{4}$$

where $\Delta M^{(t)} = M^{(t+1)} - M^{(t)}$ denotes the inter-frame displacement, and $\bar{W}_t^{(t)}$ is the averaged temporal weight of adjacent frames. Ultimately, the final joint objective for Stage 2 is:

$$\mathcal{L}_{Stage2} = \mathcal{L}_{vel} + \lambda_{kin}\mathcal{L}_{kin} + \lambda_{sem}\mathrm{CE}(\hat{Y}, Y_{target}) \tag{5}$$

where λ_{kin} and λ_{sem} are hyper-parameters that balance the trade-off.

4 Automatic Semi-Synthetic Data Generation

The size and quality of paired audio-motion data are critical for training multimodal models. Recent studies [8, 23, 48] typically rely on the BEAT dataset [22] which contains about 76 h of motion-capture recordings. While BEAT is excellent for rhythm-synchronous learning, gestures that convey explicit semantics are extremely sparse, lying in the long-tail of natural human motion. Semantic Gesticulator [48] alleviates this issue through the SeG dataset, a curated collection of more than 200 semantic gesture classes. However, SeG still omits many everyday scenarios and contains interpenetration artifacts, which limits its suitability for real-world robot motion synthesis.

To address these limitations, we introduce an automatic semi-synthetic data generation pipeline that expands the scale and improves the quality of semantic gesture data paired with speech and text. Notably, the whole data generation pipeline is designed on humanoid motion representations, generating smooth and collision-free humanoid motions for native support of humanoid control.

4.1 Retargeting and Collision-Avoidance

To obtain robot data that is feasible in the real world, two issues are central: transferring human motion to the robot while preserving accurate human-likeness, and ensuring that the transferred motions remain collision-free and executable for safe operation.

For accuracy, we use enhanced GMR [42] and Dex-Retargeting [31] to retarget body and hand motions to a G1 humanoid equipped with a customized BrainCo hand. We can handle human motions originating from different skeletal structures and enable accurate transfer of full-body and hand kinematics. For collision avoidance, we adopt the MPC-based filter in Section 3.3 to clean the data.

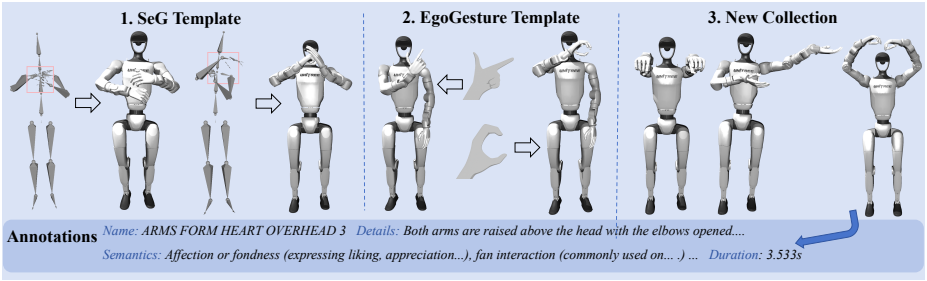

Fig. 3: Visualization of RoboGesture Dataset. RoboGesture refines the SeG dataset by correcting motion penetration issues (e.g., avoiding arm–body collisions), augments the EgoGesture dataset with full-body motion, and additionally incorporates newly collected ones. Each sample is annotated with detailed semantic labels.

4.2 RoboGesture Dataset

As shown in Figure 3, we build RoboGesture, a high-quality semantic gesture dataset for humanoid robots comprising over 300 gesture categories. The class list is derived from SeG [48] and EgoGesture [6,47] templates and expanded through user questionnaires. Each gesture is recorded with a marker-based motion capture system to preserve fine-grained kinematics, and manually annotated with detailed motion descriptions and its possible meanings across different scenarios. Each gesture is retargeted and optimized, and then replayed on the physical robot to verify reachability and control fidelity. The resulting RoboGesture dataset preserves human-like expressiveness while remaining fully executable on the robot, providing a reliable foundation for downstream gesture synthesis.

4.3 Semi-Synthetic Data Generation Pipeline

Based on RoboGesture, our retargeting pipeline and collision-avoidance tools, we generate audio–motion data enriched with semantic gestures.
Our synthesis pipeline comprises five stages: (1) Scenario Generation: We leverage LLMs [10, 27] to produce diverse daily scenarios and narrative scripts enriched with emotional descriptions. (2) Gesture Tagging: LLMs select semantically appropriate gestures from RoboGesture and determine their optimal insertion points, adjusting the wording to ensure natural linguistic integration. (3) Multimodal Synthesis: We generate emotion-aware TTS audio with word-level timestamps alongside rhythmic beat gestures via a model trained on the BEAT dataset. (4) Temporal Blending: Following [15, 48], semantic gestures are scheduled to initiate 0.4s before their corresponding spoken keywords—mimicking human anticipatory behavior—and are then fused with rhythmic motions using fifth-order interpolation. (5) Safety Optimization: Finally, we apply collision-avoidance refinement to the blended trajectories. This automated, low-cost pipeline enables the generation of millions of high-quality, semi-synthetic audio-motion pairs across diverse conversational contexts by simply varying the input prompts or text libraries.

5 Experiments

To assess gesture quality, we first evaluate on the standard co-speech gesture generation task. This task, typically conducted in simulation, requires the model to produce natural and synchronized gesture motions conditioned on speech and serves as a widely-used benchmark with established metrics.
To further evaluate the model's interactive capabilities and potential for real-world deployment, we conduct experiments on a social interaction task in a physical setting. In this task, given an audio input, the model is required to generate expressive and physically safe multimodal responses that encompass both audio and motion in a streaming manner.

5.1 Evaluating on Co-speech Gesture Generation

Datasets. (1) **BEAT.** The BEAT dataset [22] comprises 76 hours of multimodal recordings from 30 participants, encompassing speech audio, transcriptions, and high-quality motion capture data. It serves as a standard benchmark for training and evaluating co-speech gesture generation models and includes semantic labels for quantitative evaluation. (2) **SemanticBEAT.** The BEAT dataset exhibits limited domain coverage, and its semantic gesture annotations are often vague. These factors render it insufficient for the rigorous evaluation of semantic-aware co-speech gesture generation. To address these limitations, we introduce *SemanticBEAT*, a new test set comprising 1,000 speech videos compiled from online sources and user surveys. For this dataset, we performed manual temporal annotation of semantic gestures, precisely marking the intervals during which speakers should perform meaningful gestures.
Baselines. We compare our method with recent state-of-the-art baselines that provide pretrained checkpoints (trained on BEAT). (1) **LivelySpeaker** (ICCV 2023) [49]: generates semantically and rhythmically aware gestures using an MLP-based diffusion model. Since it produces fixed-length motion clips (34 frames), we follow the authors' strategy to smoothly concatenate consecutive clips. (2) **Diffsheg** (CVPR 2024) [8]: aligns gestures with audio/text using global semantic cues. (3) **SemTalk** (ICCV 2025) [45]: separately learn general motions and sparse motions with adaptive fusion. (4) **Semantic Gesticulator(SG)** (SIGGRAPH 2024) [48]: first produces rhythmic gestures and then refines semantic gestures via an LLM-based retrieval module. All baseline outputs are originally human motions; thus, we apply meticulous retargeting tuned to the robot to ensure a fair comparison of motion quality. We further verify in the supplementary material that re-training representative baselines directly in the robot joint space yields nearly identical results, confirming that our gains stem from the model rather than from retargeting artifacts.

Quantitative Evaluation Metrics. Following [8, 23, 48, 49], we evaluate our method using five key metrics: (1) **Fréchet Gesture Distance (FGD)** [40] evaluates the fidelity of generated gestures to the real motion distribution by

Table 2: Quantitative comparison of our model with other methods on the BEAT and SemanticBEAT datasets. Since the SemanticBEAT dataset does not provide ground-truth motion or semantic annotations, we omit FGD and MSE for this dataset. Best results are shown in **bold**, and second-best results are underlined.

	BEAT					SemanticBEAT		
Method	**FGD ↓**	**BC ↑**	**MSE ↓**	**DIV ↑**	**Col. ↓**	**BC ↑**	**DIV ↑**	**Col. ↓**
LivelySpeaker [49]	3.0350	0.1811	0.1861	0.1485	21.41	0.2852	0.1384	13.36
DiffSHEG [8]	2.2316	0.1851	0.1753	0.1218	**0.85**	0.2859	0.1209	1.20
SemTalk [45]	7.9328	0.1828	0.3931	0.1781	52.82	0.2913	0.1440	42.65
Semantic Gesticulator [48]	3.0147	0.1771	0.2375	**0.2818**	13.11	0.2906	**0.2831**	13.84
Ours	**0.8452**	**0.1866**	**0.1347**	0.2075	0.88	**0.2950**	0.2041	**0.13**

Table 3: Human Evaluation on BEAT and SemanticBEAT datasets. We evaluate models across four key dimensions: Rhythmic Alignment (RA), Semantic Accuracy (SA), Physical & Hand Consistency (PHC), and Overall Preference (OP). All values are reported as mean ± standard error. Higher values indicate better performance.

	BEAT				SemanticBEAT			
Model	RA↑	SA↑	PHC↑	OP↑	RA↑	SA↑	PHC↑	OP↑
LivelySpeaker [49]	-0.2223	-0.3994	-0.4304	-0.3064	-0.2126	-0.3015	-0.3466	-0.3497
Diffsheg [8]	-0.0058	-0.2691	-0.1440	-0.0992	-0.0090	-0.0817	-0.0458	-0.0626
SemTalk [45]	0.0128	0.0151	0.1679	-0.0132	0.0650	-0.0012	0.1207	0.0970
SemanticGesticulator [48]	0.0819	0.2264	-0.0273	0.1137	0.0296	0.1605	-0.0347	0.0640
Ours	**0.1334**	**0.4270**	**0.4338**	**0.3050**	**0.1269**	**0.2239**	**0.3066**	**0.2514**

embedding sequences into a latent space via a pre-trained autoencoder. (2) **Beat Consistency (BC)** [20] quantifies speech-motion synchronization by measuring the alignment between audio onsets and motion beats (velocity minima in upper-body joints). (3) **Diversity (DIV)** [19] assesses the variability of generated motions, computed as the average L_1 distance between pairs of N generated clips. (4) **Collision Rate (Col.)** measures the physical plausibility of the motion; we employ GMR [42] to detect self-collisions between the robot's mesh components. (5) **Mean Squared Error (MSE)** [39] calculates the average Euclidean distance between the generated joint positions and the ground-truth sequences to measure structural reconstruction accuracy. Metrics (1)(5) were omitted for SemanticBeat as it is an out-of-domain test set lacking ground truth.

Results. Table 2 demonstrates that our method outperforms baselines across most metrics. We achieve state-of-the-art FGD/BC/MSE on the BEAT benchmark, indicating superior alignment with the ground-truth distribution and accurate structural reconstruction. Furthermore, our model ensures strict physical plausibility with a near-zero Collision Rate (0.88%/0.13%). While SG yields higher DIV scores, numerical diversity can be artificially inflated by unstable or jittery motions [8]. Overall, our approach strikes an optimal balance between kinematic realism, speech-motion synchronization, and safe gesture diversity.

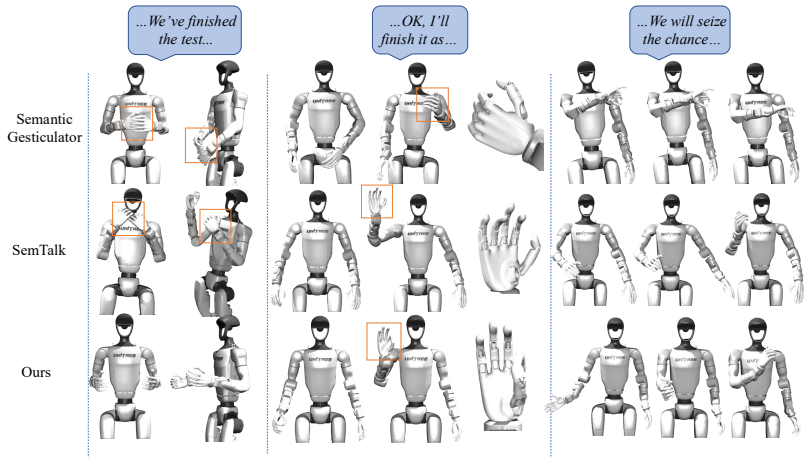

Fig. 4: Qualitative comparison of generated co-speech gestures. We present three representative scenarios to highlight the advantages of our method. **Left:** Baseline methods (e.g., SG) occasionally suffer from severe self-penetration (highlighted by the orange box), whereas our model ensures physical safety and strictly avoids self-collisions. **Middle:** Given the explicit semantic cue "OK", our model successfully synthesizes the precise, fine-grained "OK" hand gesture, while baselines fail to capture this specific semantic alignment. **Right:** For emphatic speech contexts ("seize the chance"), our model generates highly expressive body language, such as confidently patting the chest, demonstrating superior contextual richness and overall expressiveness.

Human Evaluation. To achieve a more comprehensive assessment of semantic-aligned, safe, and human-like motion qualities, we conducted a subjective user study based on pairwise comparisons. Following [1,28,48], We randomly sampled 20 generated gesture clips for each baseline, yielding a total of 200 comparison pairs. Participants were asked to evaluate the paired motions across four key dimensions: Rhythmic Alignment, Semantic Accuracy, Physical & Hand Consistency, and Overall Preference . The pairwise responses were then aggregated into merit scores to reflect relative performance. Further details regarding the user study setup and score calculation are provided in the supplementary material.

Results. The subjective evaluation results are summarized in Table 3. Our method consistently outperforms all baselines across all dimensions. Specifically, our model achieves the highest scores in Physical & Hand Consistency, verifying that our approach successfully suppresses high-frequency jittering and generates safer, physically plausible motions. Furthermore, our substantial lead in Rhythmic Alignment and Semantic Accuracy demonstrates that our framework produces gestures that are not only temporally synchronized with audio beats but also highly contextually appropriate. Consequently, our method secures the highest Overall Preference, confirming that the generated trajectories are perceived as significantly more natural and human-like compared to existing baselines.

Table 4: Ablation study on the key components of our method. SA: Semantic Action Score, HD: Hand Detail Score, HN: Human-likeness & Naturalness, BM: Beat Matching Score. Bold indicates the best performance among all configurations.

Configuration	SA ↑	HD ↑	HN ↑	BM ↑
Data Scale				
Ours w/o Semi-data	4.281	2.321	4.945	4.628
Ours (1/4 Semi-data)	6.316	5.980	6.882	6.892
Architecture & Strategy				
Ours w/o Context Motion	4.237	4.263	2.192	1.928
Ours w/o FiLM Injection	5.181	4.389	4.506	4.589
Ours w/o Semantic Classification	5.007	4.747	4.750	5.268
Ours w/o CFG	4.628	4.885	6.453	6.212
AR Strategy	4.843	5.031	5.358	6.850
Loss & Refinement				
Ours w/o Kinetic-Aware (KA) Loss	5.573	3.947	4.763	5.587
Ours w/o Filter	6.541	6.913	6.891	7.387
Ours (Full Model)	**7.175**	**7.203**	**7.394**	**7.529**

5.2 Ablation Study

Experimental Setup. To thoroughly evaluate the contribution of each proposed component in our framework, we conduct a comprehensive ablation study. We recruited 200 participants to evaluate the generated motions across four key dimensions: Semantic Action Score, Hand Detail Score, Human-likeness & Naturalness, and Beat Matching Score. The quantitative results of these configurations are reported in Table 4. We group our analysis into three main perspectives: data scale, architecture and strategy, and loss refinement.

Data Scale: The semi-synthetic dataset plays a vital role in our system. Removing it entirely (*Ours w/o Semi-data*) causes a catastrophic drop in SA and HD, as the model loses the ability to generate rich, directionally explicit semantic hand gestures, degrading to producing merely rhythmic beat motions. Even when trained on a subset (*Ours 1/4 Semi-data*, approx. 250 hours), the overall expressiveness and semantic richness remain noticeably sub-optimal.

Architecture & Strategy: Discarding the past motion context in the cross-attention module (*Ours w/o Context Motion*) leads to severe motion discontinuities and jumps, reflected in the lowest HN and BM scores. Removing the FiLM injection (*Ours w/o FiLM Injection*) impairs the model's expressiveness and semantic accuracy. Furthermore, omitting the semantic classification task (*Ours w/o Semantic Classification*) prevents the masked model from explicitly learning semantic cues from audio tokens, resulting in a significant accuracy drop. Classifier-Free Guidance (*Ours w/o CFG*) is also crucial; without it, the model tends to infer motions solely from past states rather than maintaining a balanced integration of cross-modal features.

Loss & Refinement: The Kinetic-Aware Loss provides essential supervision. Removing it (*Ours w/o KA Loss*) weakens the constraints on hand movements, semantic intervals, and kinematic smoothness, leading to a profound degradation in expressiveness. Finally, disabling the post-processing filter (*Ours w/o Filter*)

results in a slight but noticeable penalty in HN, confirming its necessity for ensuring ultimate motion smoothness.

5.3 Evaluating Real-world Social Interaction

Beyond gesture generation alone, we instantiate a complete interactive human–humanoid system in which the robot listens, responds, and gestures in real time. As detailed in the supplementary material, the deployed system comprises three modules: a language interaction module (ASR, a LoRA-tuned LLM, and TTS), our streaming speech-to-motion module (the audio aligner and motion generator), and a robot execution module. By leveraging our streaming aligner and motion generator, this pipeline achieves low-latency, end-to-end social interaction between humans and humanoid robots, as shown in Figure 2(c), which is then processed by our model to generate synchronized motion chunks.
The total system latency is primarily determined by the duration of the first generated audio chunk (typically $< 1.5\,\mathrm{s}$) and the first-chunk inference time ($\approx 0.25\,\mathrm{s}$). Crucially, the motion stack itself runs well above real time: the streaming motion generator sustains $\approx 120\,\mathrm{FPS}$ ($\approx 0.25\,\mathrm{s}$ per 1-second chunk) and the MPC safety filter adds only 5.6 ms per frame, both comfortably exceeding the 30 Hz robot control rate. The dominant latency therefore stems from the upstream speech pipeline (ASR, LLM, and TTS) rather than from our motion model; a full per-module latency breakdown is provided in the supplementary material. Our experimental platform consists of a Unitree G1 robot equipped with dual BrainCo dexterous hands, utilizing a Proportional-Derivative (PD) control scheme to translate high-level trajectories into low-level motor commands. As illustrated in Figure 1, the experimental results validate the framework's capability to generate semantically coherent and safety-constrained gestures.

6 Conclusion

In this work, we present a robot-centric framework for enabling humanoid robots to engage in natural, real-time multimodal interaction through synchronized speech and expressive gestures. Our approach co-designs data curation, model architecture, and safety mechanisms with three main contributions: (1) a holistic framework integrating data, modeling, and control; (2) a data synthesis pipeline that produces a large-scale, semantically rich training corpus by mapping expressive human gestures to robot-specific kinematics; and (3) A continuous audio-driven motion policy to ensure tight temporal synchronization and prevent the model from collapsing into repetitive historical motion patterns. Through experiments on a humanoid robot, we demonstrate that our system generates more semantically appropriate and safety-aware responses compared to baselines.

7 Acknowledgments

We thank Galbot for its generous support of this project. We are also grateful to Benhao Qin and colleagues for their assistance with the hardware deployment.

References

1. Alexanderson, S., Nagy, R., Beskow, J., Henter, G.E.: Listen, denoise, action! audio-driven motion synthesis with diffusion models. ACM Transactions on Graphics (TOG) **42**(4), 1–20 (2023)
2. Ao, T., Zhang, Z., Liu, L.: Gesturediffuclip: Gesture diffusion model with clip latents. ACM Trans. Graph. . https://doi.org/10.1145/3592097
3. Araujo, J.P., Ze, Y., Xu, P., Wu, J., Liu, C.K.: Retargeting matters: General motion retargeting for humanoid motion tracking. arXiv preprint arXiv:2510.02252 (2025)
4. Bettosi, C., Baillie, L., Ross, M.K., Broz, F.: A systematic approach to modeling structured behavior in social robots. In: Proceedings of the 2024 International Symposium on Technological Advances in Human-Robot Interaction. pp. 29–37 (2024)
5. Cai, Z., Jiang, J., Qing, Z., Guo, X., Zhang, M., Lin, Z., Mei, H., Wei, C., Wang, R., Yin, W., Pan, L., Fan, X., Du, H., Gao, P., Yang, Z., Gao, Y., Li, J., Ren, T., Wei, Y., Wang, X., Loy, C.C., Yang, L., Liu, Z.: Digital life project: Autonomous 3d characters with social intelligence. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 582–592 (June 2024)
6. Cao, C., Zhang, Y., Wu, Y., Lu, H., Cheng, J.: Egocentric gesture recognition using recurrent 3d convolutional neural networks with spatiotemporal transformer modules. In: Proceedings of the IEEE international conference on computer vision. pp. 3763–3771 (2017)
7. Cao, L.: Ai robots and humanoid ai: Review, perspectives and directions. arXiv preprint arXiv:2405.15775 (2024)
8. Chen, J., Liu, Y., Wang, J., Zeng, A., Li, Y., Chen, Q.: 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. pp. 7352–7361 (2024)
9. Cheng, Y., Huang, S.: Holegest: Decoupled diffusion and motion priors for generating holisticly expressive co-speech gestures. In: 2025 International Conference on 3D Vision (3DV). pp. 748–757. IEEE (2025)
10. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
11. Dafarra, S., Pattacini, U., Romualdi, G., Rapetti, L., Grieco, R., Darvish, K., Milani, G., Valli, E., Sorrentino, I., Viceconte, P.M., et al.: icub3 avatar system: Enabling remote fully immersive embodiment of humanoid robots. Science Robotics **9**(86), eadh3834 (2024)
12. Défossez, A., Mazaré, L., Orsini, M., Royer, A., Pérez, P., Jégou, H., Grave, E., Zeghidour, N.: Moshi: a speech-text foundation model for real-time dialogue. Tech. rep. (2024), https://arxiv.org/abs/2410.00037
13. Galatolo, A., Winkle, K.: Simultaneous text and gesture generation for social robots with small language models. Frontiers in Robotics and AI **12**, 1581024 (2025)
14. Ghorbani, S., Ferstl, Y., Holden, D., Troje, N.F., Carbonneau, M.A.: Zeroeggs: Zero-shot example-based gesture generation from speech. In: Computer Graphics Forum. vol. 42, pp. 206–216. Wiley Online Library (2023)
15. Indefrey, P., Levelt, W.J.: The spatial and temporal signatures of word production components. Cognition **92**(1-2), 101–144 (2004)

16. Jiang, J., Xiao, W., Lin, Z., Zhang, H., Ren, T., Gao, Y., Lin, Z., Cai, Z., Yang, L., Liu, Z.: Solami: Social vision-language-action modeling for immersive interaction with 3d autonomous characters. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 26887–26898 (2025)
17. Kipp, M.: Gesture generation by imitation: From human behavior to computer character animation. Universal-Publishers (2005)
18. Levine, S., Krähenbühl, P., Thrun, S., Koltun, V.: Gesture controllers. In: Acm siggraph 2010 papers, pp. 1–11 (2010)
19. Li, J., Kang, D., Pei, W., Zhe, X., Zhang, Y., He, Z., Bao, L.: Audio2gestures: Generating diverse gestures from speech audio with conditional variational autoencoders. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11293–11302 (2021)
20. Li, R., Yang, S., Ross, D.A., Kanazawa, A.: Ai choreographer: Music conditioned 3d dance generation with aist++ (2021)
21. Lipman, Y., Chen, R.T.Q., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling (2023), https://arxiv.org/abs/2210.02747
22. Liu, H., Zhu, Z., Iwamoto, N., Peng, Y., Li, Z., Zhou, Y., Bozkurt, E., Zheng, B.: Beat: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis. In: European conference on computer vision. pp. 612–630. Springer (2022)
23. Liu, L., Ghaleb, E., Ozyurek, A., Yumak, Z.: Semges: Semantics-aware co-speech gesture generation using semantic coherence and relevance learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13963–13973 (2025)
24. Mascaro, E.V., Yan, Y., Lee, D.: Robot interaction behavior generation based on social motion forecasting for human-robot interaction. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 17264–17271. IEEE (2024)
25. Matheus, K., Ramnauth, R., Scassellati, B., Salomons, N.: Long-term interactions with social robots: Trends, insights, and recommendations. ACM Transactions on Human-Robot Interaction **14**(3), 1–42 (2025)
26. McNeill, D.: Hand and mind: What gestures reveal about thought. University of Chicago press (1992)
27. OpenAI: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
28. Parizet, E., Hamzaoui, N., Sabatie, G.: Comparison of some listening test methods: a case study. Acta Acustica united with Acustica **91**, 356–364 (2005)
29. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)
30. Perez, E., Strub, F., De Vries, H., Dumoulin, V., Courville, A.: Film: Visual reasoning with a general conditioning layer. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 32 (2018)
31. Qin, Y., Yang, W., Huang, B., Van Wyk, K., Su, H., Wang, X., Chao, Y.W., Fox, D.: Anyteleop: A general vision-based dexterous robot arm-hand teleoperation system. In: Robotics: Science and Systems (2023)
32. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
33. Scherf, L., Fröhlich, K., Koert, D.: Learning action conditions for automatic behavior tree generation from human demonstrations. In: Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction. pp. 950–954 (2024)

34. Tagliamonte, C., Maccaline, D., LeMasurier, G., Yanco, H.A.: A generalizable architecture for explaining robot failures using behavior trees and large language models. In: Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction. pp. 1038–1042 (2024)
35. Valls Mascaro, E., Lee, D.: Robot behavior generation for social human-robot interaction. International Journal of Social Robotics pp. 1–20 (2025)
36. Wang, Z., Chen, J., Chen, Z., Xie, P., Chen, R., Yi, L.: Genh2r: Learning generalizable human-to-robot handover via scalable simulation demonstration and imitation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16362–16372 (2024)
37. Wang, Z., Chen, Z., Chen, J., Wang, J., Yang, Y., Liu, Y., Liu, X., Wang, H., Yi, L.: Mobileh2r: Learning generalizable human to mobile robot handover exclusively from scalable and diverse synthetic data. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 17315–17325 (2025)
38. Yang, L., Huang, X., Wu, Z., Kanazawa, A., Abbeel, P., Sferrazza, C., Liu, C.K., Duan, R., Shi, G.: Omniretarget: Interaction-preserving data generation for humanoid whole-body loco-manipulation and scene interaction. arXiv preprint arXiv:2509.26633 (2025)
39. Yang, S., Wu, Z., Li, M., Zhang, Z., Hao, L., Bao, W., Cheng, M., Xiao, L.: 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. pp. 5860–5868. International Joint Conferences on Artificial Intelligence Organization (8 2023). https://doi.org/10.24963/ijcai.2023/650, https://doi.org/10.24963/ijcai.2023/650
40. Yoon, Y., Cha, B., Lee, J.H., Jang, M., Lee, J., Kim, J., Lee, G.: Speech gesture generation from the trimodal context of text, audio, and speaker identity. ACM Transactions on Graphics (TOG) **39**(6), 1–16 (2020)
41. Yoon, Y., Ko, W.R., Jang, M., Lee, J., Kim, J., Lee, G.: Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots. In: 2019 International Conference on Robotics and Automation (ICRA). pp. 4303–4309. IEEE (2019)
42. Ze, Y., Araújo, J.P., Wu, J., Liu, C.K.: Gmr: General motion retargeting (2025), https://github.com/YanjieZe/GMR, gitHub repository
43. Zhang, X., Cai, Y., Li, K., Yang, K., Zhou, Y., Li, Z., Chu, X., Zhang, J., Liu, H.: Personagesture: Single-reference co-speech gesture personalization for unseen speakers. arXiv preprint arXiv:2605.06064 (2026)
44. Zhang, X., Li, J., Ren, J., Zhang, J.: Mitigating error accumulation in co-speech motion generation via global rotation diffusion and multi-level constraints. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 12834–12842 (2026)
45. Zhang, X., Li, J., Zhang, J., Dang, Z., Ren, J., Bo, L., Tu, Z.: Semtalk: Holistic co-speech motion generation with frame-level semantic emphasis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13761–13771 (2025)
46. Zhang, X., Li, J., Zhang, J., Ren, J., Bo, L., Tu, Z.: Echomask: Speech-queried attention-based mask modeling for holistic co-speech motion generation. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 10827–10836 (2025)
47. Zhang, Y., Cao, C., Cheng, J., Lu, H.: Egogesture: A new dataset and benchmark for egocentric hand gesture recognition. IEEE Transactions on Multimedia **20**(5), 1038–1050 (2018)

48. Zhang, Z., Ao, T., Zhang, Y., Gao, Q., Lin, C., Chen, B., Liu, L.: Semantic gesticulator: Semantics-aware co-speech gesture synthesis. ACM Transactions on Graphics (TOG) **43**(4), 1–17 (2024)
49. Zhi, Y., Cun, X., Chen, X., Shen, X., Guo, W., Huang, S., Gao, S.: Livelyspeaker: Towards semantic-aware co-speech gesture generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 20807–20817 (October 2023)