SPEAKINGROBOT: HETEROGENEOUS DIFFUSION WITH SEMANTIC MOTION PRIORS FOR HUMANOID CO-SPEECH GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Humanoid robots are increasingly built for speaking roles, such as guides, receptionists, and tutors, where the body should convey what the speech means: semantic gestures must land on the right words, and the physical robot must be able to execute them. Existing approaches typically inject a gesture either as a soft condition, which lets it drift away from its word, or as a hard constraint, which copies a retrieved clip together with dynamics from its source context that the robot cannot reliably execute. We formulate this tension as a *preservation–adaptation* problem and address it with *editable semantic motion priors*: a retrieved gesture is treated as structured but uncertain, fixing which action occurs and at which word while leaving its dynamics and the whole-body support free to adapt. We present **SpeakingRobot**, which schedules sparse word-level semantic events with a language model and retrieves a robot-space prior for each event. Its heterogeneous whole-body diffusion initializes only the scheduled upper-body regions from lightly corrupted priors and everything else, including the torso, pelvis, and legs, from Gaussian noise; one denoiser refines them jointly, so the support can anticipate each gesture. Training exposes the denoiser to the same heterogeneous states and, through cross-instance priors, rewards adapting a prior rather than copying it. SpeakingRobot improves generation quality over the compared co-speech generators on ZeroEGGS and BEAT2 even without priors, and raises tracking success on BEAT2 under a shared whole-body tracker; with priors, it keeps gestures on their words nearly as precisely as hard constraints while tracking more reliably, and participants in a blind study on a physical humanoid rate the full system higher on every criterion.

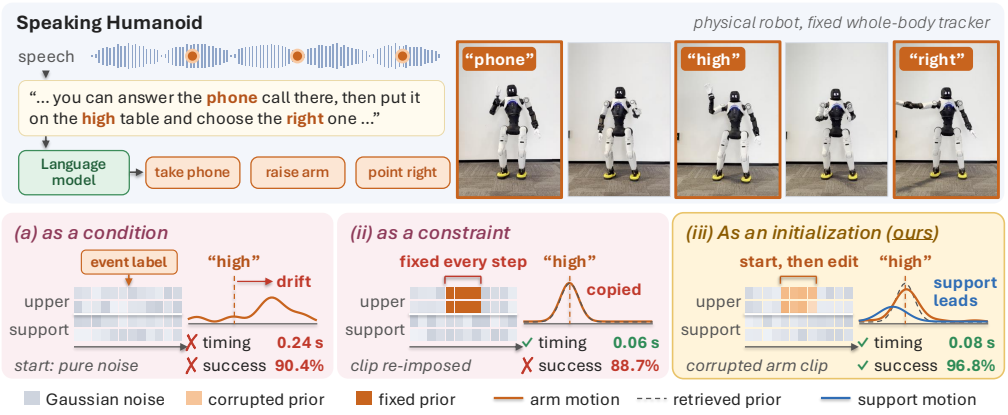

Figure 1: A speaking humanoid must place each semantic gesture on its word and stay balanced. Events scheduled on the transcript are performed on their trigger words (orange). **(a–c)** A scheduled gesture injected as a *condition* (event label only) drifts from its word; as a *constraint* (hard anchor), it is copied with foreign dynamics and tracking fails more often; as an *editable prior* initializing only the gesture region (ours), it keeps its timing while its dynamics and support adapt. Curves are schematic; values from Table 4.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Imagine that you would like a humanoid assistant to show a new colleague around the office. When it says "you can answer the phone call there, then put it on the high table and choose the right one" (Fig. 1a), you would expect it to gesture like a person: mime the call, raise a hand on *high*, and point at *the right one*. Such gestures carry part of the message: listeners integrate gesture and speech into one meaning (McNeill, 1992; Kendon, 2004) and understand better when gestures are present (Hostetter, 2011), and robots that gesture are humanlike and help listeners remember what they say (Salem et al., 2013; Huang & Mutlu, 2013; van Dijk et al., 2013). The roles of guide and tutor have long been explored with social robots (Burgard et al., 1999; Kanda et al., 2010; Belpaeme et al., 2018), and robots powered by language models now converse fluently, raising expectations for sophisticated nonverbal cues (Kim et al., 2024). Yet robots typically rely on pre-authored or rule-based speech–gesture association (Yoon et al., 2019), which cannot follow open-ended speech word by word. We study *humanoid co-speech generation*: producing, from speech, whole-body motion that conveys meaning at the right words and that the robot can execute.

Two properties define this problem, and the second is specific to robots. First, *meaning is sparse and time-locked*. Most co-speech motion is beat motion that follows prosody; the gestures that carry meaning are sparse events bound to particular words, in the long tail of the distribution (Liang et al., 2022; Zhang et al., 2024), and each stroke must coincide with the word it depicts (McNeill, 1992): a hand raised well after *high* no longer reads as *high*. Second, *on a humanoid, every gesture is also a balance event*. A fast and large reach shifts the center of mass, so the torso, pelvis, and feet must adjust *before* the stroke, not after it; a whole-body tracker (Chen et al., 2026b) only partially compensates for a reference lacking this support, and the rest appears as stumbling or falling (Sec. 4.5). A robot co-speech reference must therefore *preserve* each semantic gesture, its identity and its word-level timing—while *adapting* its dynamics and generating the whole-body support that makes it executable. We call this the *preservation–adaptation* problem.

Existing methods fail on at least one side of this problem. Co-speech generators inject meaning as a *condition*: text embeddings, semantic labels, retrieved exemplars, or coherence losses (Kucherenko et al., 2020; Liang et al., 2022; Liu et al., 2025a; Yang et al., 2025b). Conditions raise overall relevance, but do not bind a gesture to its word: under one shared peak detector, robot-space baselines place their peaks 0.20–0.31 s from the trigger word (Sec. 4.3). The opposite route treats a retrieved gesture as an *observation* to preserve, by inserting or blending it, inpainting around it, or re-imposing it at every reverse step (Lugmayr et al., 2022; Xie et al., 2024). In our controlled comparison, re-imposing the prior makes timing nearly exact (0.06 s), but the gesture is reproduced with the amplitude and velocity of its source clip, and the rest of the body must support a trajectory it cannot adapt: with the same fixed tracker, the success rate drops to 88.7% (ours: 96.8%; Sec. 4.5). Both failures arise from the *role* gesture plays in generation (Fig. 1b–d). Our key insight is that a retrieved gesture should act as an *editable semantic motion prior*: structured but uncertain, it should fix which action occurs and where its peak lies, while its dynamics and the support around it remain free to adapt. Diffusion makes this possible: sampling started from a partially corrupted signal keeps its coarse structure and re-synthesizes its details (Meng et al., 2022), so a gesture that *initializes* generation can be preserved and adapted at once. RAG-Gesture injects retrieved gestures in this spirit at inference time (Mughal et al., 2025); a humanoid additionally needs the prior confined to where it is informative and a denoiser that edits it rather than copies it.

Realizing this insight on a humanoid raises three research problems. **(i) What to initialize, and when.** Semantic events are sparse, open-vocabulary, and rare in co-speech corpora, and whether a word deserves a gesture depends on its meaning, so the events cannot be reliably inferred from prosody. **(ii) Where the prior may enter.** A retrieved clip shows how the arms moved in another context, so it should shape only the arms and only during the event, whereas the torso, pelvis, and legs must be generated for the current gesture and begin to respond before its stroke. **(iii) How to learn adaptation rather than reproduction.** A denoiser whose prior always equals its target can minimize the loss by copying the initialization and would then copy a foreign clip at test time. We present SpeakingRobot, which answers these problems in turn with language-model event scheduling, heterogeneous whole-body diffusion, and heterogeneous-noise training with cross-instance priors (Fig. 2). Our contributions are as follows.

- We formulate humanoid co-speech generation as a **preservation–adaptation problem** and show, under a fixed whole-body tracker, that soft conditioning loses word-level timing, whereas hard constraints keep the timing but lower tracking success.
- We introduce **editable semantic motion priors**, which fix a gesture's identity and word-level timing but leave its dynamics editable, and realize in SpeakingRobot with language-model event scheduling, **heterogeneous whole-body diffusion** that initializes only the scheduled upper-body regions from priors, and **heterogeneous noise training with cross-instance priors**, which teaches the denoiser to adapt priors rather than copy them.
- Experiments on two benchmarks, under a shared whole-body tracker, and in a blind study on a physical humanoid show that SpeakingRobot improves generation quality and tracking success over the evaluated co-speech generators, and that editable priors keep gestures on their words while executing more reliably than hard constraints.

2 Related Work

Co-speech gesture generation. Gesture synthesis (Nyatsanga et al., 2023) has progressed from rule-based templates (Cassell et al., 1994; Neff et al., 2008) through learned audio- and text-driven models (Ginosar et al., 2019; Yoon et al., 2020; Kucherenko et al., 2020; Ao et al., 2022) to diffusion, flow-matching, and token-based generators (Zhu et al., 2023; Yang et al., 2023; Ao et al., 2023; Liu et al., 2024; 2025b; Guichoux et al., 2025; Saleem et al., 2026). Semantics are strengthened by learned representations, language guidance, and motion priors (Liang et al., 2022; Zhi et al., 2023; Zhang et al., 2025; Liu et al., 2025a; Gao et al., 2025; Chen et al., 2025; Cheng & Huang, 2025), and retrieval improves long-tail coverage (Mughal et al., 2025; Yang et al., 2025b; Suresh et al., 2026). Retrieved or predicted gestures serve as conditions (Yang et al., 2025b) or key-pose anchors (Zhang et al., 2026a), are aligned into the output (Zhang et al., 2024), or are injected at inference time by DDIM inversion (Mughal et al., 2025). These methods target virtual characters, whose motion need not be physically executable.

Controllable and prior-initialized diffusion. Motion in-betweening (Zhou et al., 2020; Fan et al., 2026) and diffusion models support sparse temporal, spatial, joint-level, and part-level control (Tevet et al., 2023; Shafir et al., 2024; Karunratanakul et al., 2023; Xie et al., 2024; Cohan et al., 2024; Bae et al., 2025; Li et al., 2026a; Dai et al., 2026), and retrieval augments generation (Zhang et al., 2023). Keyframes, inpainted regions (Lugmayr et al., 2022), and masked observations (Liu et al., 2024) are treated as exact, which suits ground truth but not a gesture whose dynamics come from another context. SDEdit starts sampling from a corrupted input (Meng et al., 2022), retrieved motion has been noised and refined (Oba & Ukita, 2025; Liao et al., 2024), edit strength and noise level can vary per region or token (Levin & Fried, 2023; Chen et al., 2024), and optimized or retrieved noise can steer motion semantics (Karunratanakul et al., 2024; Liu et al., 2026; Ota et al., 2026; Chen et al., 2026a). We restrict the corrupted prior to the scheduled upper-body entries and train with cross-instance priors, so that one sampling pass adapts the gesture and generates its support.

Speaking robots and humanoid execution. Social robots have long gestured while speaking (Salem et al., 2013; Huang & Mutlu, 2013), using gestures learned from speech (Yoon et al., 2019) or generated and selected by language models (Huang et al., 2024; Mahadevan et al., 2024; Bao et al., 2025; Zhang, 2025; Montiel-Vazquez et al., 2026). Full-size humanoids now execute diverse human motion via retargeting and whole-body tracking (He et al., 2024; Cheng et al., 2024; Fu et al., 2024; Chen et al., 2026b; Ze et al., 2026; Luo et al., 2026), whose robustness depends on retargeting quality (Araujo et al., 2026; Zhao et al., 2026). Generated references are paired with tracking policies (Xie et al., 2026; Tevet et al., 2025), gated when the tracker cannot follow them (Cho et al., 2026), or generated directly for the robot from language or audio (Li et al., 2026c; Jiang et al., 2025; Li et al., 2026b), including streaming whole-body co-speech motion on a real humanoid (Zhang et al., 2026b) and concurrent robot co-speech work (Liang et al., 2026; Wang et al., 2026). Like them, we generate for the robot; unlike them, semantics enter as scheduled word-level events whose editable priors initialize whole-body diffusion.

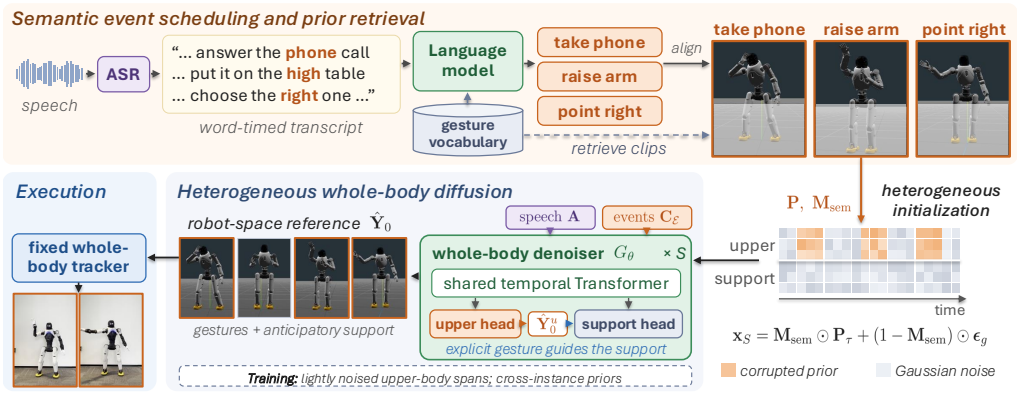

Figure 2: Overview of SpeakingRobot. A language model schedules sparse events on the transcript; each retrieves a robot-space upper-body prior whose peak is aligned to its trigger word. Only the scheduled upper-body regions start from the corrupted prior $\mathbf{P}_\tau$, everything else from Gaussian noise. One denoiser, conditioned on speech and events, predicts the upper body explicitly and the support with a head guided by it, so the support anticipates each gesture. A fixed whole-body tracker executes the reference.

3 METHOD

3.1 OVERVIEW AND FORMULATION

Given a speech waveform and its word-timed transcript, our goal is to generate a whole-body motion in robot-space $\mathbf{Y} = \{\mathbf{y}_t\}_{t=1}^{T} \in \mathbb{R}^{T \times D}$ that places each semantic gesture on its word, follows the speech rhythm and can be executed using a fixed whole-body tracking policy. Each frame $\mathbf{y}_t$ concatenates the angles of the robot joints $\mathbf{q}_t$ and their temporal differences $\Delta\mathbf{q}_t$, local root displacement $\Delta\mathbf{p}_t^{\text{loc}}$, yaw change $(\sin\Delta\psi_t, \cos\Delta\psi_t)$, local linear and angular velocities $\mathbf{v}_t^{\text{loc}}$ and $\boldsymbol{\omega}_t^{\text{loc}}$, and contact states $\mathbf{c}_t$. Generating in the robot's own joint space lets joint limits, available degrees of freedom, and support demand shape the motion while it is formed, whereas retargeting afterwards introduces artifacts that carry over into tracking (Araujo et al., 2026). We divide the D channels into semantic upper-body channels $\mathcal{D}_u$ (shoulders, elbows, wrists; $D_u = |\mathcal{D}_u|$) and support channels $\mathcal{D}_s$ (torso, pelvis, root, lower body, local trajectory and contacts), since meaning and balance live in different parts of the body: arms carry gestures, and the rest of the body keeps the robot upright.

The central idea of SpeakingRobot is to treat each retrieved gesture as an *editable semantic motion prior*. A retrieved clip, aligned to its word, reliably encodes *which* action to perform and *when* its expressive peak should occur, but not how this utterance and this robot should perform it: its dynamics and the support they require must be generated anew. SpeakingRobot thus trusts the prior for identity and timing and keeps everything else editable, in three stages (Fig. 2): event scheduling and prior retrieval (Sec. 3.2), heterogeneous whole-body diffusion (Sec. 3.3), and heterogeneous-noise training with cross-instance priors (Sec. 3.4).

3.2 SEMANTIC MOTION PRIORS

Semantic event scheduling. Whether a word deserves a gesture, and which one, depends on what the word means rather than how it is spoken: *high* calls for a raised arm and *phone call* for a hand at the ear. Such gestures are sparse, open-vocabulary, and rare in co-speech corpora, so an audio-conditioned generator learns them poorly (Liang et al., 2022; Zhang et al., 2024), whereas a language model brings the lexical and world knowledge the decision requires. A speech recognizer produces a word-timed transcript $\mathcal{W} = \{(w_i, t_i^{\text{start}}, t_i^{\text{end}})\}_{i=1}^{N}$ with word midpoints $t_i^{\text{mid}} = \frac{1}{2}(t_i^{\text{start}} + t_i^{\text{end}})$. Given $\mathcal{W}$ and a robot-space gesture vocabulary $\mathcal{G}$, the language model predicts a sparse event set $\mathcal{E} = \{e_j\}_{j=1}^{M}$, $e_j = (k_j, i_j, \rho_j)$, with gesture category k_j, trigger-word index i_j, and salience ρ_j; a deterministic validator removes invalid categories and overlapping events and enforces an event-density limit. The scheduler decides only *what* occurs and *when*, never joint trajectories: motion remains the generator's job, and the scheduler needs no retraining when the vocabulary grows.

Prior retrieval and temporal alignment. A category label says what a gesture is, not how it unfolds. A retrieved exemplar supplies this missing structure—the preparation, stroke, and retraction of the gesture (Kendon, 2004)—which a generator would otherwise have to learn from few examples of a rare category, and it already respects the robot's joint limits. The vocabulary $\mathcal{G} = \{\mathcal{G}_k\}_{k=1}^{K}$ stores robot-space upper-body clips with descriptions, trigger phrases, and an annotated expressive peak. Since gesture–speech synchrony is defined by where the stroke falls relative to the word (McNeill, 1992), we align each exemplar's peak to its trigger word before generation starts. The target peak frame is $f_j = \mathrm{round}(F_{\mathrm{fps}}\, t_{i_j}^{\mathrm{mid}}) + \Delta_{k_j}$, where F_{fps} is the frame rate and Δ_{k_j} a category-specific trigger-to-peak offset; we retrieve an instance from $\mathcal{G}_{k_j}$, align its peak with f_j, and crop or resample it to $\mathbf{P}_j \in \mathbb{R}^{L_j \times D_u}$. Placing all aligned priors on the timeline gives $\mathbf{P} \in \mathbb{R}^{T \times D}$ and the region mask $\mathbf{M}_{\mathrm{sem}} = \mathbf{M}_{\mathrm{time}} \otimes \mathbf{M}_{\mathrm{upper}} \in \{0,1\}^{T \times D}$, where $\mathbf{M}_{\mathrm{time}}$ marks the scheduled intervals and $\mathbf{M}_{\mathrm{upper}}$ the channels in $\mathcal{D}_u$; where aligned clips overlap, the higher-salience prior fills the shared frames. The pair $(\mathbf{P}, \mathbf{M}_{\mathrm{sem}})$ marks where structured semantic information is *available* and imposes no constraint.

3.3 Heterogeneous whole-body diffusion

Diffusion lets a prior shape generation without fixing it: reverse diffusion from a partially corrupted signal keeps its coarse structure—for a gesture, the action and where its peak lies—and re-synthesizes the details (Meng et al., 2022). We therefore use the prior to *initialize* generation, and only where it is informative.

Prior corruption and heterogeneous initialization. We corrupt the prior to a level τ: a lower τ retains more of the retrieved dynamics, and a higher τ leaves more room for contextual adaptation. The prior is informative only where the semantics live, on the upper-body channels inside each scheduled interval; support motion from its source clip would belong to another context, and starting from it would import a foreign pattern that the denoiser must then undo. The initial state $\mathbf{x}_S$, from which the S reverse steps start, therefore draws from two sources:

$$\mathbf{P}_\tau = \sqrt{\bar{\alpha}_\tau}\,\mathbf{P} + \sqrt{1-\bar{\alpha}_\tau}\,\boldsymbol{\epsilon}_p, \qquad \mathbf{x}_S = \mathbf{M}_{\mathrm{sem}} \odot \mathbf{P}_\tau + (1 - \mathbf{M}_{\mathrm{sem}}) \odot \boldsymbol{\epsilon}_g, \tag{1}$$

with $\boldsymbol{\epsilon}_p, \boldsymbol{\epsilon}_g \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Unlike masked, inpainting, or hard-constrained diffusion, $\mathbf{M}_{\mathrm{sem}}$ marks no entries to preserve or reconstruct; it selects each entry's initialization source and tells the denoiser which entries came from a prior.

Unified whole-body denoising. Support must *anticipate* a gesture: the torso and feet have to adjust before the stroke, so the support at frame t depends on upper-body motion after t, and the gesture must in turn stay within what the support can deliver. We therefore generate the whole body in a single pass of one network, with the support following an explicit estimate of the gesture. At reverse step s, a temporal Transformer encodes the noisy window and two heads decode it:

$$\mathbf{H} = G_\theta(\mathbf{x}_s, s, \mathbf{A}, \mathbf{C}_{\mathcal{E}}, \mathbf{M}_{\mathrm{sem}}), \qquad \hat{\mathbf{Y}}_0^u = h_u(\mathbf{H}), \qquad \hat{\mathbf{Y}}_0^s = h_s(\mathbf{H}, \hat{\mathbf{Y}}_0^u), \tag{2}$$

where $\mathbf{A}$ are temporally aligned speech features, $\mathbf{C}_{\mathcal{E}}$ embeds each event's category, salience, and relative time, and $\hat{\mathbf{Y}}_0 = [\hat{\mathbf{Y}}_0^u, \hat{\mathbf{Y}}_0^s]$. The upper head predicts the gesture explicitly and is supervised directly; the support head takes this explicit upper-body motion as its guide. Because attention is bidirectional over the window, the support at each frame sees the gesture before it happens and can prepare the torso, legs, and contacts ahead of the peak.

Contextual adaptation and anticipatory support. The prior is never re-imposed during the reverse process, so the denoiser can change the gesture's amplitude, velocity, preparation, and recovery, while the identity and approximate word-level timing set in $\mathbf{x}_S$ persist. The support starts from noise *even inside* semantic intervals, so it is generated from the evolving gesture and speech rather than copied from the retrieved clip, with no stability rule or post-hoc correction. We sample with the standard conditional reverse process and classifier-free guidance (Ho & Salimans, 2022).

3.4 Heterogeneous-noise training with cross-instance priors

At inference, the denoiser starts from a state that standard diffusion training never produces: lightly corrupted regions that carry structure, surrounded by pure noise (Eq. 1). We train it on the same kind of state. In each training window, a mask $\mathbf{M}$ selects prior regions on the upper-body channels; these

entries are corrupted only to level τ instead of being noised like the rest of the window. Recovering the clean window teaches the denoiser to refine these regions and to generate everything around them: the transitions into and out of each region and the support beneath it. A denoiser whose prior always equals its target, however, can minimize the loss by copying its initialization, and would then copy a foreign clip at test time—exactly the failure of hard constraints. When a prior region holds a vocabulary gesture, we therefore draw its prior from another instance of the same category, $\mathbf{P}_j \sim \mathcal{G}_k$ with $\mathbf{P}_j \neq \mathbf{Y}^u_j$, the target's own upper-body segment. The prior then shares the gesture's identity but not its dynamics, as at inference, and copying is penalized by construction.

Learning objectives. We train with an x_0-prediction objective,

$$\mathcal{L} = \mathcal{L}^u_{\text{rec}} + \mathcal{L}^s_{\text{rec}} + \lambda_{\text{sem}}\mathcal{L}_{\text{sem}} + \lambda_{\text{vel}}\mathcal{L}_{\text{vel}} + \lambda_{\text{acc}}\mathcal{L}_{\text{acc}} + \lambda_{\text{con}}\mathcal{L}_{\text{con}}, \tag{3}$$

where $\mathcal{L}^u_{\text{rec}} = \|\mathbf{Y}^u - \hat{\mathbf{Y}}^u_0\|_1$ supervises the upper head's explicit output and $\mathcal{L}^s_{\text{rec}}$ the support head (Eq. 4), $\mathcal{L}_{\text{sem}} = \|\mathbf{M} \odot (\mathbf{Y} - \hat{\mathbf{Y}}_0)\|_1$ emphasizes the prior regions, $\mathcal{L}_{\text{vel}}$ and $\mathcal{L}_{\text{acc}}$ supervise first- and second-order differences, and $\mathcal{L}_{\text{con}}$ is a binary cross-entropy over contacts. Balance demand peaks on a gesture's highlight frames, where it is most expressive, so the support is weighted most there and less toward the surrounding rhythmic motion:

$$w^{\text{evt}}_t = 1 + \lambda_{\text{evt}} \sum_{j=1}^{M} \rho_j \exp\left(-\frac{(t - f_j)^2}{2\gamma_j^2}\right), \qquad \mathcal{L}^s_{\text{rec}} = \sum_{t=1}^{T} w^{\text{evt}}_t \left\|\mathbf{Y}^s_t - \hat{\mathbf{Y}}^s_t\right\|_1, \tag{4}$$

where $\mathbf{Y}^s_t$ are the support channels of frame t, f_j is the aligned highlight of event e_j, and γ_j sets the width of the emphasis; the weighting concentrates supervision where support matters most but prescribes no support pattern.

4 Experiments

4.1 Experimental setup

Datasets and representation. We use ZeroEGGS (Ghorbani et al., 2023) and BEAT2 (Liu et al., 2024); all robot-space methods use the common 29-DoF representation of Sec. 3.1. The gesture vocabulary is built from BABEL (Punnakkal et al., 2021) retargeted to the robot and holds 133 categories and 682 upper-body clips, each with an annotated expressive peak.

Implementation details. Word timestamps come from FunASR (Gao et al., 2023); Qwen3 (Yang et al., 2025a), prompted with the vocabulary, schedules the events (App. 2); and WavLM (Chen et al., 2022) provides speech features. The denoiser is an eight-layer temporal Transformer (hidden size 512, eight heads). Motions are sampled at 30 fps in 72-frame windows whose first 12 frames give motion history. We train with 50 diffusion levels using AdamW, batch size 64, learning rate 2×10^{-4}, and one million updates; inference uses 50 reverse steps. Prior regions are corrupted to $\tau = 10$ of the 50 diffusion levels.

Baselines and protocol. We compare with CaMN (Liu et al., 2022), DiffStyleGesture (Yang et al., 2023), EMAGE (Liu et al., 2024), Semantic Gesticulator (Zhang et al., 2024), SemTalk (Zhang et al., 2025), and HoloGest (Cheng & Huang, 2025), which condition on speech and, where available, on text, semantic labels, or retrieved gestures. We use two paired settings. In **Retarget**, a method generates in its native human or avatar representation and the result is retargeted to the humanoid; in **Robot** (suffix "-Robot"), the same architecture is retrained on the common robot representation, changing only motion-related input and output dimensions. The pair isolates whether dynamics are learned in the target space or converted after generation. All robot-space methods share data splits, frame rate, audio input, and comparable optimization budgets.

Physical execution. Every reference is executed in physics simulation by the same future-aware whole-body tracking policy, GMT (Chen et al., 2026b), with identical weights, observation and action interfaces, control frequency, reference horizon, and termination criteria. The tracker receives no labels, schedules, or priors, so physical differences reflect the reference, not the controller.

Priors in each evaluation. Real-robot experiments (Fig. 3, Table 5) use the complete system, with language-model scheduling and priors retrieved from the vocabulary. The quantitative benchmarks (Tables 1, 2, and the semantic columns of Table 4) initialize the upper-body regions of LLM-

Table 1: General co-speech generation in the common 29-DoF space. Baseline cells read retargeted → retrained in robot space. †LLM-scheduled semantic upper-body regions are initialized with forward-noised semantic gestures rather than pure Gaussian noise (Sec. 4.1); w/o priors uses Gaussian initialization (Table 4).

Method	ZeroEGGS			BEAT2		
	FGD↓	BC↑	DIV↑	FGD↓	BC↑	DIV↑
DiffStyleGesture	13.42→11.86	6.88→7.31	8.51→**8.88**	19.14→16.98	6.74→7.19	8.83→**9.11**
EMAGE	9.12→7.79	7.13→7.62	7.62→7.91	15.48→13.76	6.91→7.23	8.13→8.42
Sem. Gesticulator	8.84→7.53	7.29→7.81	7.48→7.79	13.37→11.59	7.21→7.70	7.82→8.09
SemTalk	8.13→6.93	7.52→7.99	7.55→7.84	12.79→11.17	7.38→7.91	8.17→8.50
HoloGest	8.29→7.02	7.47→7.93	7.79→8.14	12.63→11.06	7.35→7.89	7.87→8.13
Ours — robot space by construction, single value per cell						
SpeakingRobot w/o priors	6.75	8.01	8.31	10.18	7.98	8.76
SpeakingRobot†	**6.17**	**8.21**	8.54	**9.36**	**8.17**	9.03

scheduled semantic events with semantic gestures forward-noised to level τ, rather than pure Gaussian noise. All remaining regions use Gaussian initialization. Our w/o-priors variant uses Gaussian initialization in the semantic regions as well, while retaining the same conditioning.

Metrics. Generation is measured by Fréchet Gesture Distance (FGD↓), Beat Consistency (BC↑), and Diversity (DIV↑) in the common 29-DoF space. Executability under the fixed tracker is measured by Success Rate (SR↑, sequences finished without a fall or termination), Mean Per-Joint Angular Error (MPJAE↓, rad), and Mean Per-Keypoint Position Error (MPKPE↓, m), with all valid frames before termination included for failed sequences. Semantic Score (SC↑) averages $\cos(E_m(\hat{\mathbf{Y}}), E_t(\mathcal{W}))$ over test sequences, with frozen motion and text encoders shared by all methods; Peak Timing Error $\mathrm{PTE} = \frac{1}{MF_{\mathrm{fps}}} \sum_j |\hat{f}_j - f_j|$ (↓) asks whether each expressive peak lands on its scheduled word, with $\hat{f}_j$ detected identically for every method (App. 1); and Prior Deviation (PD, App. 1) measures the distance from the aligned prior; it has no optimal direction (near zero means copying, very large a lost identity) and is read with SC and PTE.

4.2 General robot-space co-speech generation

Retargeting versus robot-space retraining. Retraining the same model in robot space (Table 1) lowers FGD by 11.6–15.3% on ZeroEGGS and 11.1–13.3% on BEAT2 and raises BC in all five pairs. Since each pair shares architecture and objective, the gain comes from learning the dynamics where joint ranges and support demand are visible.

Comparison with robot-space baselines. Even without priors, SpeakingRobot attains the lowest FGD and the highest BC among the compared methods on both datasets, lowering FGD by 2.6% over SemTalk-Robot (ZeroEGGS) and 8.0% over HoloGest-Robot (BEAT2). With the LLM-scheduled semantic upper-body regions initialized with forward-noised semantic gestures rather than pure Gaussian noise, FGD falls by a further 8.6% and 8.1% while DIV rises (8.31 → 8.54, 8.76 → 9.03): the model absorbs a prior without collapsing onto an average motion.

4.3 Semantic identity and word-level timing

Soft conditions do not bind gestures to words (Table 2): the baselines place their peaks 0.20–0.31 s from the scheduled words, and so does our model given the schedule only as a condition (Gaussian initialization, 0.24 s; Table 4), since a condition, even one encoding event time, leaves the denoiser free to shift or weaken a gesture. With a prior initializing each event region, SC rises from 0.46 to 0.69 and PTE falls from 0.20 s to 0.08 s, as the peak is placed in $\mathbf{x}_S$ before denoising starts. Sec. 4.5 compares ways of using the same prior.

4.4 Physical executability

Under the fixed tracker (Table 3), robot-space retraining alone raises SR by 5.7 points for EMAGE and 4.6 for HoloGest and lowers both errors: a kinematically valid retargeted motion is not necessarily physically compatible. SpeakingRobot reaches 96.8% SR, 6.5 points above HoloGest-Robot,

Table 2: Semantic events on BEAT2. One peak detector and search window for every method. †LLM-scheduled semantic upper-body regions use semantic gestures forward-noised to τ rather than pure Gaussian noise.

Method	SC ↑	PTE (s) ↓
CaMN-Robot	0.25	0.31
EMAGE-Robot	0.32	0.27
DiffStyle-Robot	0.37	0.24
Sem. Gest.-Robot	0.46	0.20
SpeakingRobot†	**0.69**	**0.08**

Table 3: Physical executability on BEAT2 under one fixed tracker, so the spread reflects reference quality. RT: post-hoc retargeting.

Method	SR (%) ↑	MPJAE ↓	MPKPE ↓
EMAGE + RT	82.4	0.173	0.184
EMAGE-Robot	88.1	0.146	0.151
HoloGest + RT	85.7	0.158	0.162
HoloGest-Robot	90.3	0.131	0.139
SpeakingRobot	**96.8**	**0.097**	**0.106**

Table 4: Component ablation on BEAT2 (variants defined in App. 3). PD is two-sided: insertion and hard anchoring reproduce the prior, Gaussian initialization discards it. Direct insertion leaves discontinuities at clip boundaries (semantic metrics only); the support ablation changes no semantic pathway.

Variant	Semantic			Physical		
	SC ↑	PTE (s) ↓	PD	SR (%) ↑	MPJAE ↓	MPKPE ↓
Gaussian initialization	0.45	0.24	0.362	90.4	0.132	0.141
Direct insertion	0.64	0.08	0.000	–	–	–
Hard anchor	0.68	**0.06**	0.012	88.7	0.148	0.156
Homogeneous prior	0.63	0.11	0.074	91.5	0.126	0.134
w/o cross-instance priors	0.61	0.14	0.058	93.0	0.115	0.121
w/o anticipatory support	–	–	–	82.2	0.142	0.149
SpeakingRobot	**0.69**	0.08	0.137	**96.8**	**0.097**	**0.106**

and cuts MPJAE by 26.0% (0.131 → 0.097) and MPKPE by 23.7% (0.139 → 0.106). With the controller fixed, the reference itself carries the support response, so the tracker has less to correct (Sec. 4.5).

With the complete system on the physical robot (Fig. 3), EMAGE-Robot shows no recognizable semantic peak and sways laterally, and HoloGest-Robot keeps a coherent rhythm but weak semantic peak with unsteady lower-body adjustments; SpeakingRobot places clear gestures on *"this"* and *"upper"*.

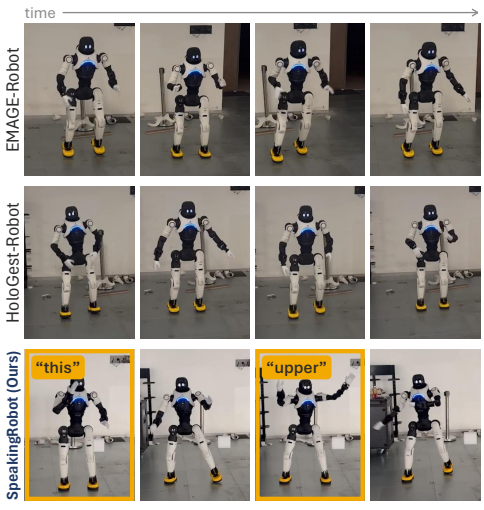

Figure 3: Real-robot comparison. Baselines show weak semantic peaks and unsteady support. SpeakingRobot places gestures on *"this"* and *"upper"* (highlighted), with anticipatory whole-body support. Full sequences appear in the supplementary video.

4.5 Ablation Study

The paired Retarget and Robot rows of Tables 1 and 3 already isolate the representation: robot-space training improves generation and execution in every evaluated pair. Table 4 tests the remaining designs of Sec. 3; all variants except Gaussian initialization receive the same prior regions, isolating how the prior is used.

Prior initialization. Replacing the prior with Gaussian noise while retaining all conditioning lowers SC from 0.69 to 0.45, increases PTE from 0.08 to 0.24 s, and reduces SR from 96.8% to 90.4%. Thus, a semantic schedule as a condition alone does not reliably preserve sparse actions; initializing the event regions provides structure that the denoiser can refine.

Editable prior versus rigid constraints. Direct insertion and hard anchoring achieve low PTE (0.08 and 0.06 s), but their near-zero PD (0.000 and 0.012) indicates copying. Hard anchoring also

Table 5: Real-robot user study. Mean ± s.e. across 20 participants, each averaging ten utterances. Bottom rows: two-sided Wilcoxon p-values, Holm-adjusted across six comparisons; * marks superiority to both baselines at adjusted $p < 0.05$.

Method	Naturalness ↑	Rhythm ↑	Semantic approp. ↑
EMAGE-Robot	3.45±0.14	3.53±0.09	2.98±0.12
HoloGest-Robot	3.76±0.10	3.81±0.09	3.64±0.09
SpeakingRobot	**4.14±0.07***	**4.20±0.08***	**4.24±0.08***
p_{Holm}: vs. EMAGE	0.00522	0.00103	1.14×10^{-5}
p_{Holm}: vs. HoloGest	0.01448	0.01160	0.00103

lowers SR to 88.7% and increases both tracking errors: preserving event timing is insufficient when gesture dynamics cannot adapt to whole-body support. Splicing and boundary blending likewise do not jointly generate the surrounding support motion. SpeakingRobot combines the highest SC with low PTE at non-zero PD (0.137), supporting an editable prior rather than a rigid trajectory constraint.

Heterogeneous initialization. Initializing all channels from the prior lowers SC to 0.63, increases PTE to 0.11 s, and reduces SR to 91.5%. Its smaller PD (0.074) indicates stronger dependence on the prior trajectory. Structured initialization therefore benefits semantic upper-body regions, while non-event and support channels benefit from the freedom to adapt through Gaussian initialization.

Cross-instance priors. Using each training gesture's own motion as its prior, rather than another instance of the same category, reduces PD to 0.058 but lowers SC to 0.61, increases PTE to 0.14 s, and reduces SR to 93.0%. Closer reproduction of the prior does not yield better semantics or timing. This supports cross-instance training as a means of learning to adapt a gesture to its context instead of copying its trajectory (Sec. 3.4).

Anticipatory whole-body support. Removing jointly generated support causes the largest SR drop, from 96.8% to 82.2%, and increases MPJAE and MPKPE by 46.4% and 40.6%. The semantic pathway and tracker remain unchanged, isolating the contribution of anticipatory torso, pelvis, root, lower-body, and contact responses to executable gestures.

4.6 Real-robot user study

Twenty participants rate 30 robot videos (ten utterances, three methods) on five-point Naturalness, Rhythm, and Semantic Appropriateness scales. Robot, camera, scene, audio level, and playback speed are fixed; method identities are hidden and presentation order is randomized per participant. We average ratings over utterances within each participant and method, giving 20 paired observations.

Significance. Tie-corrected Friedman tests with 99,999 within-participant permutations give $Q = 8.949, 9.564, 28.900$ and $p = 0.01004, 0.00641, 0.00001$ for the three criteria, respectively. Two-sided Wilcoxon tests use exact sign-permutation distributions, retaining tied ranks and excluding zeros. All six baseline comparisons remain significant after Holm correction (Table 5). Matched-pairs rank-biserial effects range from 0.626 to 1.000; these results concern the ten evaluated utterances.

5 Conclusion

We presented SpeakingRobot, which treats each retrieved gesture as an editable semantic motion prior which is a region-restricted initialization of whole-body diffusion rather than a condition or a constraint and trains its denoiser to adapt such priors rather than copy them, so that a humanoid places semantic gestures on their words while generating the support needed to execute them. The scheduler, retriever, and tracker interact with the generator only through the schedule, priors, and reference, so each can be improved independently, e.g., with learned retargeting (Zhao et al., 2026) for new embodiments.

AI USE STATEMENT

In this work, a large language model is a component of the proposed method: it schedules semantic events (Sec. 3.2), and its role is limited to selecting gesture categories and trigger words; it generates no motion. Beyond this, we used generative AI tools to edit the manuscript for readability, to help restructure and draft parts of the text (including the discussion of results), to help create and revise figures, to search for and summarize related literature, and to format references. We have not used generative AI tools to generate data or to design or run experiments. We have reviewed all AI-assisted work, checked every reported number against our experiments, and verified every reference, and we take responsibility for the final content of this work, including text, claims, and figures produced with the aid of generative AI.

ETHICS STATEMENT

We have not identified any specific ethical concerns arising from this work.

REPRODUCIBILITY STATEMENT

Sec. 4.1 gives the datasets, motion representation, vocabulary size, network configuration, optimization settings, diffusion schedule, and the prior used in each evaluation; App. 1 specifies the shared peak detector used for PTE, and App. 2 describes the scheduler interface and validator. Every physical result uses one publicly described tracking policy (Chen et al., 2026b) with fixed weights and identical termination criteria across all methods. Code, the robot-space gesture vocabulary, and the evaluation scripts will be released.

REFERENCES

Tenglong Ao, Qingzhe Gao, Yuke Lou, Baoquan Chen, and Libin Liu. Rhythmic Gesticulator: Rhythm-aware co-speech gesture synthesis with hierarchical neural embeddings. *ACM Transactions on Graphics*, 41(6):1–19, 2022.

Tenglong Ao, Zeyi Zhang, and Libin Liu. GestureDiffuCLIP: Gesture diffusion model with CLIP latents. *ACM Transactions on Graphics*, 42(4):1–18, 2023.

Joao Pedro Araujo, Yanjie Ze, Pei Xu, Jiajun Wu, and C. Karen Liu. Retargeting matters: General motion retargeting for humanoid motion tracking. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2026.

Jinseok Bae, Inwoo Hwang, Young-Yoon Lee, Ziyu Guo, Joseph Liu, Yizhak Ben-Shabat, Young Min Kim, and Mubbasir Kapadia. Less is more: Improving motion diffusion models with sparse keyframes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11069–11078, 2025.

Lingfan Bao, Yan Pan, Tianhu Peng, Dimitrios Kanoulas, and Chengxu Zhou. Hierarchical intention-aware expressive motion generation for humanoid robots. *arXiv preprint arXiv:2506.01563*, 2025.

Tony Belpaeme, James Kennedy, Aditi Ramachandran, Brian Scassellati, and Fumihide Tanaka. Social robots for education: A review. *Science Robotics*, 3(21):eaat5954, 2018.

Wolfram Burgard, Armin B. Cremers, Dieter Fox, Dirk Hähnel, Gerhard Lakemeyer, Dirk Schulz, Walter Steiner, and Sebastian Thrun. Experiences with an interactive museum tour-guide robot. *Artificial Intelligence*, 114(1–2):3–55, 1999.

Justine Cassell, Catherine Pelachaud, Norman Badler, Mark Steedman, Brett Achorn, Tripp Becket, Brett Douville, Scott Prevost, and Matthew 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 (SIGGRAPH)*, pp. 413–420, 1994.

Bohong Chen, Yumeng Li, Youyi Zheng, Yao-Xiang Ding, and Kun Zhou. Motion-example-controlled co-speech gesture generation leveraging large language models. In *ACM SIGGRAPH 2025 Conference Papers*, 2025.

Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 24081–24125, 2024.

Chia-Wen Chen, Yan Wu, Korrawe Karunratanakul, and Siyu Tang. NaP-Control: Navigating diffusion prior for versatile and fast character control. In *European Conference on Computer Vision (ECCV)*, 2026a.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022.

Zixuan Chen, Mazeyu Ji, Xuxin Cheng, Xuanbin Peng, Xue Bin Peng, and Xiaolong Wang. GMT: General motion tracking for humanoid whole-body control. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2026b.

Xuxin Cheng, Yandong Ji, Junming Chen, Ruihan Yang, Ge Yang, and Xiaolong Wang. Expressive whole-body control for humanoid robots. In *Robotics: Science and Systems (RSS)*, 2024.

Yongkang Cheng and Shaoli Huang. HoloGest: Decoupled diffusion and motion priors for generating holisticly expressive co-speech gestures. In *Proceedings of the International Conference on 3D Vision (3DV)*, 2025.

Hanbyel Cho, Sang-Hun Kim, Jeonguk Kang, and Donghan Koo. SafeFlow: Real-time text-driven humanoid whole-body control via physics-guided rectified flow and selective safety gating. *arXiv preprint arXiv:2603.23983*, 2026.

Setareh Cohan, Guy Tevet, Daniele Reda, Xue Bin Peng, and Michiel van de Panne. Flexible motion in-betweening with diffusion models. In *ACM SIGGRAPH 2024 Conference Papers*, 2024.

Minyue Dai, Ke Fan, Anyi Rao, Jingbo Wang, and Bo Dai. Controllable text-to-motion generation via modular body-part phase control. *arXiv preprint arXiv:2603.19795*, 2026.

Shiyu Fan, Paul Henderson, and Edmond S. L. Ho. Generative motion in-betweening by diffusion over continuous implicit representations. *IEEE Transactions on Visualization and Computer Graphics*, 32(7):7107–7117, 2026.

Zipeng Fu, Qingqing Zhao, Qi Wu, Gordon Wetzstein, and Chelsea Finn. HumanPlus: Humanoid shadowing and imitation from humans. In *Conference on Robot Learning (CoRL)*, volume 270 of *Proceedings of Machine Learning Research*, pp. 2828–2844, 2024.

Nan Gao, Yihua Bao, Dongdong Weng, Jiayi Zhao, Jia Li, Yan Zhou, and Pengfei Wan. SARGes: Semantically aligned reliable gesture generation via intent chain. In *Proceedings of the International Workshop on Generation and Evaluation of Non-verbal Behaviour for Embodied Agents (GENEA)*, pp. 13–21, 2025.

Zhifu Gao, Zerui Li, Jiaming Wang, Haoneng Luo, Xian Shi, Mengzhe Chen, Yabin Li, Lingyun Zuo, Zhihao Du, and Shiliang Zhang. FunASR: A fundamental end-to-end speech recognition toolkit. In *Interspeech*, pp. 1593–1597, 2023.

Saeed Ghorbani, Ylva Ferstl, Daniel Holden, Nikolaus F. Troje, and Marc-André Carbonneau. ZeroEGGS: Zero-shot example-based gesture generation from speech. *Computer Graphics Forum*, 42(1):206–216, 2023.

Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, and Jitendra Malik. Learning individual styles of conversational gesture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3497–3506, 2019.

Téo Guichoux, Théodor Lemerle, Shivam Mehta, Jonas Beskow, Gustav Eje Henter, Laure Soulier, Catherine Pelachaud, and Nicolas Obin. Gelina: Unified speech and gesture synthesis via interleaved token prediction. *arXiv preprint arXiv:2510.12834*, 2025.

Tairan He, Zhengyi Luo, Wenli Xiao, Chong Zhang, Kris Kitani, Changliu Liu, and Guanya Shi. Learning human-to-humanoid real-time whole-body teleoperation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 8944–8951, 2024.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Autumn B. Hostetter. When do gestures communicate? a meta-analysis. *Psychological Bulletin*, 137(2):297–315, 2011.

Chien-Ming Huang and Bilge Mutlu. Modeling and evaluating narrative gestures for humanlike robots. In *Robotics: Science and Systems (RSS)*, 2013.

Peide Huang, Yuhan Hu, Nataliya Nechyporenko, Daehwa Kim, Walter Talbott, and Jian Zhang. EMOTION: Expressive motion sequence generation for humanoid robots with in-context learning. *arXiv preprint arXiv:2410.23234*, 2024.

Nan Jiang, Zimo He, Wanhe Yu, Lexi Pang, Yunhao Li, Hongjie Li, Jieming Cui, Yuhan Li, Yizhou Wang, Yixin Zhu, and Siyuan Huang. UniAct: Unified motion generation and action streaming for humanoid robots. *arXiv preprint arXiv:2512.24321*, 2025.

Takayuki Kanda, Masahiro Shiomi, Zenta Miyashita, Hiroshi Ishiguro, and Norihiro Hagita. A communication robot in a shopping mall. *IEEE Transactions on Robotics*, 26(5):897–913, 2010.

Korrawe Karunratanakul, Konpat Preechakul, Supasorn Suwajanakorn, and Siyu Tang. Guided motion diffusion for controllable human motion synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2151–2162, 2023.

Korrawe Karunratanakul, Konpat Preechakul, Emre Aksan, Thabo Beeler, Supasorn Suwajanakorn, and Siyu Tang. Optimizing diffusion noise can serve as universal motion priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1334–1345, 2024.

Adam Kendon. *Gesture: Visible Action as Utterance*. Cambridge University Press, 2004.

Callie Y. Kim, Christine P. Lee, and Bilge Mutlu. Understanding large-language model (LLM)-powered human-robot interaction. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 371–380, 2024.

Taras Kucherenko, Patrik Jonell, Sanne van Waveren, Gustav Eje Henter, Simon Alexanderson, Iolanda Leite, and Hedvig Kjellström. Gesticulator: A framework for semantically-aware speech-driven gesture generation. In *Proceedings of the 2020 International Conference on Multimodal Interaction (ICMI)*, pp. 242–250, 2020.

Eran Levin and Ohad Fried. Differential diffusion: Giving each pixel its strength. *arXiv preprint arXiv:2306.00950*, 2023.

Chuqiao Li, Xianghui Xie, Yong Cao, Andreas Geiger, and Gerard Pons-Moll. FrankenMotion: Part-level human motion generation and composition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16420–16431, 2026a.

Zhe Li, Cheng Chi, Yangyang Wei, Boan Zhu, Tao Huang, Zhenguo Sun, Yibo Peng, Pengwei Wang, Zhongyuan Wang, Fangzhou Liu, Chang Xu, and Shanghang Zhang. Do you have freestyle? expressive humanoid locomotion via audio control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026b.

Zhe Li, Yangyang Wei, Boan Zhu, Yibo Peng, Tao Huang, Pengwei Wang, Zhongyuan Wang, Cheng Chi, Chang Xu, and Shanghang Zhang. From language to locomotion: Retargeting-free humanoid control via motion latent guidance. In *International Conference on Learning Representations (ICLR)*, 2026c.

Yuanzhi Liang, Qianyu Feng, Linchao Zhu, Li Hu, Pan Pan, and Yi Yang. SEEG: Semantic energized co-speech gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10473–10482, 2022.

Zhangzhao Liang, Xiaofen Xing, Mingyue Yang, Wenlve Zhou, and Xiangmin Xu. PhysDrift: Bridging the embodiment gap in humanoid co-speech motion generation. *arXiv preprint arXiv:2606.19935*, 2026.

Zhouyingcheng Liao, Mingyuan Zhang, Wenjia Wang, Lei Yang, and Taku Komura. RMD: A simple baseline for more general human motion generation via training-free retrieval-augmented motion diffuse. *arXiv preprint arXiv:2412.04343*, 2024.

Haiyang Liu, Zihao Zhu, Naoya Iwamoto, Yichen Peng, Zhengqing Li, You Zhou, Elif Bozkurt, and Bo Zheng. BEAT: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis. In *European Conference on Computer Vision (ECCV)*, pp. 612–630. Springer, 2022.

Haiyang Liu, Zihao Zhu, Giorgio Becherini, Yichen Peng, Mingyang Su, You Zhou, Xuefei Zhe, Naoya Iwamoto, Bo Zheng, and Michael J. Black. EMAGE: Towards unified holistic co-speech gesture generation via expressive masked audio gesture modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1144–1154, 2024.

Hanchao Liu, Fang-Lue Zhang, Shining Zhang, Tai-Jiang Mu, and Shi-Min Hu. Towards highly-constrained human motion generation with retrieval-guided diffusion noise optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. arXiv:2605.08054.

Lanmiao Liu, Esam Ghaleb, Aslı Özyürek, and Zerrin Yumak. SemGes: Semantics-aware co-speech gesture generation using semantic coherence and relevance learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13963–13973, 2025a.

Pinxin Liu, Luchuan Song, Junhua Huang, Haiyang Liu, and Chenliang Xu. GestureLSM: Latent shortcut based co-speech gesture generation with spatial-temporal modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10929–10939, 2025b.

Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. RePaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11461–11471, 2022.

Zhengyi Luo, Ye Yuan, Tingwu Wang, Chenran Li, Fernando Castañeda, Sirui Chen, Zi-Ang Cao, Jiefeng Li, David Minor, Qingwei Ben, Jinhyung Park, David Sami, Zi Wang, Xingye Da, Runyu Ding, Cyrus Hogg, Lina Song, Edy Lim, Eugene Jeong, Tairan He, Haoru Xue, Wenli Xiao, Simon Yuen, Jan Kautz, Yan Chang, Umar Iqbal, Linxi Fan, and Yuke Zhu. SONIC: Supersizing motion tracking for natural humanoid whole-body control. *Science Robotics*, 11(117):eaed4592, 2026.

Karthik Mahadevan, Jonathan Chien, Noah Brown, Zhuo Xu, Carolina Parada, Fei Xia, Andy Zeng, Leila Takayama, and Dorsa Sadigh. Generative expressive robot behaviors using large language models. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 482–491, 2024.

David McNeill. *Hand and Mind: What Gestures Reveal about Thought*. University of Chicago Press, 1992.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2022.

Edwin C. Montiel-Vazquez, Christian Arzate Cruz, Stefanos Gkikas, Thomas Kassiotis, Giorgos Giannakakis, and Randy Gomez. Efficient emotion-aware iconic gesture prediction for robot co-speech. *arXiv preprint arXiv:2604.11417*, 2026.

M. Hamza Mughal, Rishabh Dabral, Merel C. J. Scholman, Vera Demberg, and Christian Theobalt. Retrieving semantics from the deep: An RAG solution for gesture synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16578–16588, 2025.

Michael Neff, Michael Kipp, Irene Albrecht, and Hans-Peter Seidel. Gesture modeling and animation based on a probabilistic re-creation of speaker style. *ACM Transactions on Graphics*, 27(1): 1–24, 2008.

Simbarashe Nyatsanga, Taras Kucherenko, Chaitanya Ahuja, Gustav Eje Henter, and Michael Neff. A comprehensive review of data-driven co-speech gesture generation. *Computer Graphics Forum*, 42(2):569–596, 2023.

Takeru Oba and Norimichi Ukita. R2-Diff: Denoising by diffusion as a refinement of retrieved motion for image-based motion prediction. *Neurocomputing*, 647:130489, 2025.

Sakuya Ota, Qing Yu, Kent Fujiwara, Satoshi Ikehata, and Ikuro Sato. Retrieving and refining winning noise tickets for diffusion-based motion generation. In *European Conference on Computer Vision (ECCV)*, 2026.

Abhinanda R. Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J. Black. BABEL: Bodies, action and behavior with English labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 722–731, 2021.

Muhammad Usama Saleem, Mayur Jagdishbhai Patel, Ekkasit Pinyoanuntapong, Zhongxing Qin, Li Yang, Hongfei Xue, Ahmed Helmy, Chen Chen, and Pu Wang. LiveGesture: Streamable co-speech gesture generation model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Maha Salem, Friederike Eyssel, Katharina Rohlfing, Stefan Kopp, and Frank Joublin. To err is human(-like): Effects of robot gesture on perceived anthropomorphism and likability. *International Journal of Social Robotics*, 5(3):313–323, 2013.

Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit H. Bermano. Human motion diffusion as a generative prior. In *International Conference on Learning Representations (ICLR)*, 2024.

Varsha Suresh, Mohammad Mahdi Abootorabi, Mohamed Salman, M. Hamza Mughal, Christian Theobalt, Ashwin Ram, Jürgen Steimle, and Vera Demberg. Semantic motion anchors: Bridging motion and meaning in co-speech gestures. *arXiv preprint arXiv:2605.30608*, 2026.

Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H. Bermano. Human motion diffusion model. In *International Conference on Learning Representations (ICLR)*, 2023.

Guy Tevet, Sigal Raab, Setareh Cohan, Daniele Reda, Zhengyi Luo, Xue Bin Peng, Amit Haim Bermano, and Michiel van de Panne. CLoSD: Closing the loop between simulation and diffusion for multi-task character control. In *International Conference on Learning Representations (ICLR)*, 2025.

Elisabeth T. van Dijk, Elena Torta, and Raymond H. Cuijpers. Effects of eye contact and iconic gestures on message retention in human-robot interaction. *International Journal of Social Robotics*, 5(4):491–501, 2013.

Zifan Wang, Ziang Ren, Pengyang Shi, Zirui Wang, Chenghuai Lin, Tianze Wang, Zekun Qi, Liangliang Zhao, He Wang, and Li Yi. RoboGesture: Real-time semantic-aligned co-speech gestures generation for humanoid interaction. In *European Conference on Computer Vision (ECCV)*, 2026.

Weiji Xie, Jiakun Zheng, Jinrui Han, Jiyuan Shi, Weinan Zhang, Chenjia Bai, and Xuelong Li. TextOp: Real-time interactive text-driven humanoid robot motion generation and control. In *Conference on Robot Learning (CoRL)*, 2026.

Yiming Xie, Varun Jampani, Lei Zhong, Deqing Sun, and Huaizu Jiang. OmniControl: Control any joint at any time for human motion generation. In *International Conference on Learning Representations (ICLR)*, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Quanwei Yang, Luying Huang, Kaisiyuan Wang, Jiazhi Guan, Shengyi He, Fengguo Li, Hang Zhou, Lingyun Yu, Yingying Li, Haocheng Feng, and Hongtao Xie. GestureHYDRA: Semantic co-speech gesture synthesis via hybrid modality diffusion transformer and cascaded-synchronized retrieval-augmented generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12615–12625, 2025b.

Sicheng Yang, Zhiyong Wu, Minglei Li, Zhensong Zhang, Lei Hao, Weihong Bao, Ming Cheng, and Long Xiao. DiffuseStyleGesture: Stylized audio-driven co-speech gesture generation with diffusion models. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 5860–5868, 2023.

Youngwoo Yoon, Woo-Ri Ko, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4303–4309, 2019.

Youngwoo Yoon, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. Speech gesture generation from the trimodal context of text, audio, and speaker identity. *ACM Transactions on Graphics*, 39(6):1–16, 2020.

Yanjie Ze, Siheng Zhao, Weizhuo Wang, Angjoo Kanazawa, Rocky Duan, Pieter Abbeel, Guanya Shi, Jiajun Wu, and C. Karen Liu. TWIST2: Scalable, portable, and holistic humanoid data collection system. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2026.

Gang Zhang. Semantic co-speech gesture synthesis and real-time control for humanoid robots. *arXiv preprint arXiv:2512.17183*, 2025.

Mingyuan Zhang, Xinying Guo, Liang Pan, Zhongang Cai, Fangzhou Hong, Huirong Li, Lei Yang, and Ziwei Liu. ReMoDiffuse: Retrieval-augmented motion diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 364–373, 2023.

Xiangyue Zhang, Jianfang Li, Jiaxu Zhang, Ziqiang Dang, Jianqiang Ren, Liefeng Bo, and Zhigang Tu. SemTalk: Holistic co-speech motion generation with frame-level semantic emphasis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13761–13771, 2025.

Xiangyue Zhang, Jianfang Li, Jiaxu Zhang, Kaixing Yang, and Steven Hoi. StreamTalk: Streaming co-speech gesture generation with key-pose anchoring. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2026a.

Zeyi Zhang, Tenglong Ao, Yuyao Zhang, Qingzhe Gao, Chuan Lin, Baoquan Chen, and Libin Liu. Semantic Gesticulator: Semantics-aware co-speech gesture synthesis. *ACM Transactions on Graphics*, 43(4):1–17, 2024.

Zeyi Zhang, Zixi Kang, Ruijie Zhao, Yusen Feng, Biao Jiang, Hanyu Ji, and Libin Liu. ProAct: Harnessing streaming motion generation and agentic reasoning for realtime embodied social interaction. *ACM Transactions on Graphics (TOG)*, 45(6), 2026b. SIGGRAPH Asia 2026, to appear. arXiv:2602.14048.

Qingrui Zhao, Kaiyue Yang, Xiyu Wang, Shiqi Zhao, Yi Lu, Xinfang Zhang, Qiu Shen, Xiao-Xiao Long, and Xun Cao. Make tracking easy: Neural motion retargeting for humanoid whole-body control. *arXiv preprint arXiv:2603.22201*, 2026.

Yihao Zhi, Xiaodong Cun, Xuelin Chen, Xi Shen, Wen Guo, Shaoli Huang, and Shenghua Gao. LivelySpeaker: Towards semantic-aware co-speech gesture generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20807–20817, 2023.

Yi Zhou, Jingwan Lu, Connelly Barnes, Jimei Yang, Sitao Xiang, and Hao Li. Generative tweening: Long-term inbetweening of 3D human motions. *arXiv preprint arXiv:2005.08891*, 2020.

Lingting Zhu, Xian Liu, Xuanyu Liu, Rui Qian, Ziwei Liu, and Lequan Yu. Taming diffusion models for audio-driven co-speech gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10544–10553, 2023.

A Peak Detection and Metric Details

For event e_j we compute the normalized upper-body motion energy $r_j(t) = \sum_{d\in\mathcal{D}_u} \left|(\hat{Y}_{t,d} - \hat{Y}_{t-1,d})/\sigma_d\right|$, where $\mathcal{D}_u$ holds the bilateral shoulder, elbow, and wrist DoFs and σ_d is the training-set standard deviation of channel d. We smooth r_j with a one-dimensional Gaussian kernel ($\sigma_{\text{peak}} = 2$ frames) and detect $\hat{f}_j = \arg\max_{t\in\widetilde{\mathcal{I}}_j} \widetilde{r}_j(t)$, where $\widetilde{\mathcal{I}}_j$ is the scheduled interval extended by 0.25 s on both sides. Every method uses the same joints, normalization, search range, smoothing, and peak-selection rule, so PTE compares placement rather than energy profile. Prior Deviation is $\text{PD} = \frac{1}{M}\sum_j \|\hat{\mathbf{Y}}^u_j - \mathbf{P}_j\|_1/(L_j D_u)$, where $\hat{\mathbf{Y}}^u_j$ is the generated upper-body segment of event e_j.

B Semantic Event Scheduler

Interface. The scheduler receives the word-timed transcript $\mathcal{W}$ with word indices and, for every vocabulary category k, its name, a short description, and its trigger phrases. It returns a list of events $e_j = (k_j, i_j, \rho_j)$, i.e., a category, the index of the trigger word, and a salience score. It never outputs joint values, durations, or amplitudes; these come from the retrieved prior and the denoiser.
Validation. A deterministic validator (i) removes events whose category is not in the vocabulary or whose word index is out of range, (ii) removes temporally overlapping events, keeping the one with higher salience, and (iii) caps the number of events per window. Only validated events define $\mathbf{P}$ and $\mathbf{M}_{\text{sem}}$.

C Ablation Variants

Gaussian initialization starts the whole window from Gaussian noise and keeps all conditioning; it is the w/o-priors row of Table 1. **Direct insertion** pastes the aligned prior into the output, and **hard anchor** re-imposes it after every reverse step. **Homogeneous prior** initializes all channels of the scheduled intervals from the prior. **W/o cross-instance priors** takes each training gesture's own motion as its prior instead of another instance of the same category. **W/o anticipatory support** leaves the semantic pathway and the tracker unchanged but does not generate the support jointly with the gesture.