FOREMOTION: LEARNING CAUSAL CO-SPEECH GESTURES THROUGH FUTURE-INFORMED SUPERVISION

Anonymous authors
Paper under double-blind review

ABSTRACT

Generating co-speech gestures as speech unfolds requires predicting motion from incomplete acoustic and linguistic context. Complete training recordings, however, contain later speech that is unavailable to a causal generator at deployment. The challenge is to use this additional context for learning while preserving the generator's causal speech access at inference. We present ForeMotion, a framework for learning causal co-speech gestures through future-informed gesture supervision. ForeMotion represents motion as discrete gesture tokens and trains a shared autoregressive generator using paired-view gesture distillation. Both views receive an identical rolling window of past and current speech. The future-conditioned gesture view additionally accesses later audio and text from the training clip. Both views predict the same gesture targets from the same permitted motion history. The future-conditioned view produces probability distributions over possible gesture tokens. A gesture-token distillation loss trains the causal view to match these distributions, with gradients stopped through the supervisory predictions. At inference, only the causal view is executed, generating motion autoregressively from available speech and previously generated motion without requiring future speech. Experiments on BEATX show that this training procedure improves selected gesture-quality measures on the evaluated backbone, with benefits varying across context scopes, speaker settings, and evaluation metrics.

1 INTRODUCTION

Co-speech gesture generation animates virtual characters with body movements that accompany speech (Cassell, 2000; Habibie et al., 2021; Ghorbani et al., 2023). When speech arrives incrementally, interactive avatars need to generate accompanying motion as the utterance unfolds (Cassell et al., 1998; Saleem et al., 2026). We focus on causal streaming, where predictions use only past and current speech. The generator must therefore choose movements without waiting for the complete utterance, while part of the context that could inform those movements is still unavailable.

Autoregressive models support this setting by generating successive motion tokens conditioned on preceding motion tokens and speech context. Recent causal gesture generators combine this sequential prediction with access restricted to available speech (Mughal et al., 2026; Saleem et al., 2026). Yet causal speech access can exclude cues relevant to the movement currently being generated. Studies of representational gestures find that movements often begin before their associated words are spoken (Morrel-Samuels & Krauss, 1992; ter Bekke et al., 2024; Donnellan et al., 2022). A causal generator may therefore need to initiate a movement before a relevant lexical cue becomes observable. Figure 1(A) illustrates the contrast with bidirectional speech access, which also exposes upcoming cues.

Completed training recordings make this missing speech context available for learning. Learning Using Privileged Information (LUPI) formalizes the use of richer inputs during training that are absent at inference (Vapnik & Vashist, 2009). Generalized distillation uses predictions informed by such inputs to supervise a model with restricted access (Lopez-Paz et al., 2016). Recent work applies future-conditioned supervision to language modeling and gaze estimation (Zhao et al., 2026; Li et al., 2026). For gestures, completed speech could similarly inform supervision for a causal view

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.

(Figure 1(B)). Speech admits multiple plausible gestures (Habibie et al., 2021; Li et al., 2021). We ask whether predictions conditioned on richer speech context from completed training recordings can provide useful supervision for a causal gesture generator without expanding its inference-time speech access.

We introduce **ForeMotion**, a framework for learning causal co-speech gestures through future-informed gesture supervision. Through paired-view gesture distillation, the same generator produces two predictive views during training. Both views receive an identical rolling window of past and current speech. The future-conditioned view additionally accesses later speech within the training clip. The gesture-token distillation loss uses a forward Kullback-Leibler divergence to align the causal predictions with detached future-conditioned token distributions across the regional motion-token streams. Both views retain the same permitted gesture history, and the privileged inputs contain no target or future gesture tokens. At inference, ForeMotion runs only the causal view. The additional future-conditioned computation is confined to training.

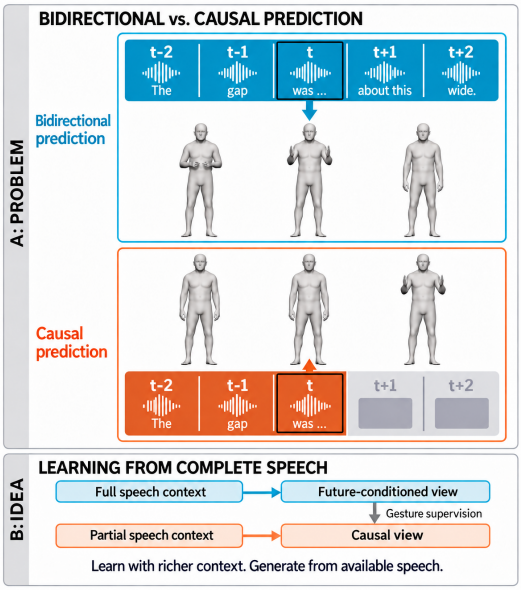


Figure 1: Motivation for future-informed gesture supervision. (A) Bidirectional speech access exposes upcoming cues unavailable to causal prediction at t. Poses illustrate possible timing differences, not measured outputs. (B) A future-conditioned view supervises the causal view of a shared generator during training; inference uses only available speech.

Our study treats future-context scope as an explicit experimental variable. We examine word-bounded, segment-bounded, and full-clip speech access, with their boundaries defined in Section 3.2. The motion representation and deployed generator are held fixed. The single-speaker study compares ForeMotion with a causal control using gesture token regularization, a supporting strategy for residual vector quantization (RVQ). Single- and multi-speaker comparisons with released baselines evaluate the complete training procedure.

Experiments on BEATX support the utility of this training procedure on the evaluated backbone. The observed benefits depend on context scope, speaker setting, and evaluation metric.

Our contributions are threefold:

- **A training formulation using expanded speech context as privileged supervision for causal gesture generation**, with explicit training and deployment information sets and three joint audio/text access scopes.
- **ForeMotion, implementing paired-view gesture distillation with shared parameters**, retaining the causal inference architecture and using gesture token regularization as a supporting component.
- **An empirical evaluation of the training procedure and its context scopes**, including a single-speaker regularized-control study, context-scope comparisons, and complete-system comparisons in single- and multi-speaker settings.

2 Related Work

2.1 Speech Conditioning, Gesture Variation, and Interaction

Gesture generators combine acoustic and textual cues, often with speaker information, to model speech–motion relationships (Kucherenko et al., 2020; Yoon et al., 2020; Liu et al., 2022a; 2024).

Conditional latent and diffusion models capture motion variation (Li et al., 2021; Mughal et al., 2024), while example- and prompt-based conditioning controls gesture style (Ghorbani et al., 2023; Ao et al., 2023). Holistic models jointly animate body, hands, and face (Habibie et al., 2021; Yi et al., 2023; Liu et al., 2024).

Semantic gesture generation uses alignment objectives and retrieval of speech-relevant motion (Liu et al., 2025b; Zhang et al., 2024; Mughal et al., 2025); TANGO combines audio-motion retrieval with diffusion interpolation for gesture-video reenactment (Liu et al., 2025a). Conversational models additionally address two-person interactions (Zhao et al., 2023; Ng et al., 2024; Qi et al., 2025). ForeMotion studies supervision for causal generation of individual speakers' gestures.

2.2 Streaming generation and speech access

Streaming output does not imply zero-lookahead speech access. STARGATE uses blocks containing future speech (Abel et al., 2024), and accelerated rolling diffusion denoises multiple frames jointly for continuous generation (Vu et al., 2026). MIBURI uses causal Temporal and Kinematic Transformers over regional RVQ streams, while LiveGesture combines streamable tokens, regional autoregressive experts, and causal fusion (Mughal et al., 2026; Saleem et al., 2026). ForeMotion retains MIBURI's deployed information boundary and studies richer supervision from completed training recordings.

2.3 Learning from privileged future observations

Knowledge distillation transfers softened teacher predictions (Hinton et al., 2015); Born Again Networks train separate students of the same architecture (Furlanello et al., 2018). LUPI introduces training-only information (Vapnik & Vashist, 2009), while generalized distillation transfers privileged-input guidance through teacher predictions (Lopez-Paz et al., 2016).

Regret Pre-training uses shared-parameter causal/future-aware passes and detached forward-KL supervision for language modeling (Zhao et al., 2026). ECOGaze applies this principle to gaze prediction with future video (Li et al., 2026). These are direct precedents for ForeMotion's paired-view gesture distillation. LocalRegret exposes the supervised language token, while GlobalRegret masks it; ForeMotion privileges future audio and text without exposing target or future gesture tokens.

Other preprints use different mechanisms. Privileged Foresight Distillation uses future video to train an action-prediction correction adapter retained at inference (Fang et al., 2026). Privileged Self-Distillation uses future interactions, momentum teacher targets, and a divergence gate in recommendation (Tang et al., 2026). ForeMotion matches regional gesture-token distributions using shared parameters, without a deployed correction adapter.

2.4 Context scope and limits of transfer

Future-context benefits depend on the task and model. ECOGaze reports its best results at intermediate tested horizons (Li et al., 2026); generalized distillation identifies limits from transferable information and student approximation (Lopez-Paz et al., 2016). Privileged Foresight Distillation tests temporal correspondence using matched and shuffled future observations (Fang et al., 2026). Variable gesture-speech timing likewise does not determine one useful lookahead boundary (Wagner et al., 2014; ter Bekke et al., 2024; Donnellan et al., 2022). ForeMotion compares word-bounded, segment-bounded, and full-clip access and examines supervision across speaker settings. These scopes vary in both duration and content.

3 Method

ForeMotion changes the supervision available during training while retaining the causal generator used at inference. During training, an additional pass of the same generator produces future-conditioned token distributions that supervise its causal predictions. Figure 2 summarizes the motion representation, shared generator, and paired-view gesture distillation. Gesture token regularization supports token learning and is also used by the regularized causal control.

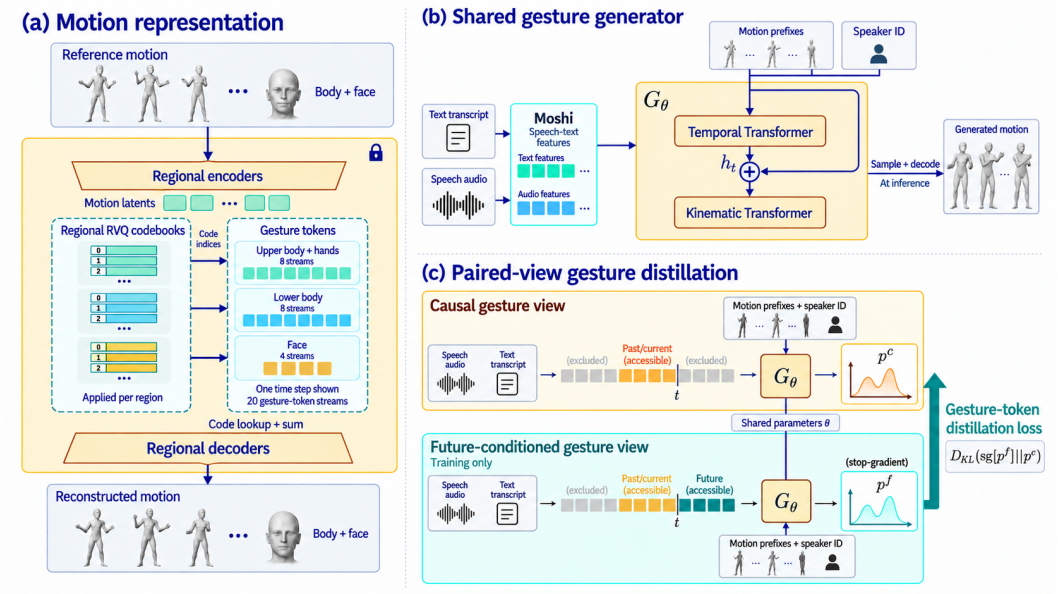


Figure 2: System overview of ForeMotion. (a) Frozen regional RVQ codecs represent upper-body and hand, lower-body, and facial motion using 8, 8, and 4 token streams. Code lookup and summation reconstruct regional latents for decoding. (b) G_θ denotes the shared temporal/kinematic generator, conditioned on speech, permitted motion prefixes and speaker identity; h_t abbreviates its view-dependent temporal state $\mathbf{h}_t^v$. Sampling and decoding are used at inference. (c) Paired-view gesture distillation matches causal predictions to detached future-conditioned distributions. Both views use identical past/current speech windows; the future-conditioned view additionally receives later speech up to the selected scope boundary (Equation 2). Here p^{c} and p^{f} abbreviate the real-code distillation distributions $\mathbf{p}_{b,t,k}^{v,\tau_d}$ of the two views, omitting sample, time, stream and temperature indices. Curves and the per-distribution KL term are schematic. Only the causal view executes at inference.

3.1 Preliminaries

We build on the motion representation and causal autoregressive generator of MIBURI (Mughal et al., 2026). These define the gesture targets, conditioning information, and prediction order used by ForeMotion.

Motion representation. Three frozen causal residual vector quantization (RVQ) codecs encode upper-body and hand motion, lower-body motion, and facial motion into eight, eight, and four token streams, respectively (Figure 2(a)). At step t, the gesture-token vector $\mathbf{G}_t = (g_{t,0}, \ldots, g_{t,K-1})$ contains $K = 20$ code indices. Each index selects one of V entries in its corresponding codebook, indexed by $\mathcal{V} = \{0, \ldots, V-1\}$. The streams concatenate three separate regional residual hierarchies. Code lookup, residual summation, and the regional decoders reconstruct motion from the token sequence. We call the reference tokens from deterministic codec encoding *canonical codec tokens*.

Causal prediction. Let A and X denote aligned audio and text token sequences, s the speaker identity, and $\mathcal{C}_t^{\mathrm{c}}$ the past and current speech permitted at step t. Let $\mathbf{G} = (\mathbf{G}_0, \ldots, \mathbf{G}_{T-1})$ be a sequence of T gesture-token steps. Writing $\mathbf{G}_{<t}$ for preceding vectors and $\mathbf{G}_{t,<k}$ for earlier streams within the current step, a generator with parameters θ models

$$p_\theta^{\mathrm{c}}(\mathbf{G} \mid A, X, s) = \prod_{t=0}^{T-1} \prod_{k=0}^{K-1} p_\theta(g_{t,k} \mid \mathbf{G}_{<t}, \mathbf{G}_{t,<k}, \mathcal{C}_t^{\mathrm{c}}, s)\,. \tag{1}$$

Temporal attention uses finite windows; the factorization specifies the prediction order. Sample indices are omitted except in batched losses.

Shared gesture generator. The Temporal Transformer summarizes the permitted motion history and speech into a state $\mathbf{h}_t^{\mathrm{c}}$ and predicts stream $k = 0$. The Kinematic Transformer uses this state and the preceding within-step tokens to predict streams $k = 1, \ldots, K-1$; its direct speech inputs are aligned to the current step (Figure 2(b)). ForeMotion retains these codecs, model parameterization, and generation order. Using this generator, we construct two training views with different speech access and shared parameters.

3.2 Paired-view gesture distillation

Following the shared-parameter formulation of Regret Pre-training (Zhao et al., 2026), we evaluate one generator under two speech-access policies. The *causal gesture view* uses $\mathcal{C}_t^{\mathrm{c}}$; the *future-conditioned gesture view* uses the expanded context $\mathcal{C}_t^{\mathrm{f}}$. Both views use identical past/current speech windows and receive the same preceding gesture-token vectors, preceding within-step codes, and speaker identity. Neither receives the token being predicted or later motion tokens. The future conditioning consists of audio and text admitted jointly.

Speech access. Let q index speech groups on the gesture-token clock, L_s denote the causal speech-window length, and $B_\sigma(t)$ be the final speech group admitted by scope σ. The direct speech cross-attention positions for the two views are

$$\begin{aligned} \mathcal{J}_t^{\mathrm{c}} &= \{q : \max(0, t - L_s + 1) \le q \le t\}, \\ \mathcal{J}_t^{\mathrm{f}} &= \{q : \max(0, t - L_s + 1) \le q \le B_\sigma(t)\}. \end{aligned} \tag{2}$$

For view $v \in \{\mathrm{c}, \mathrm{f}\}$, $\mathcal{C}_t^v$ contains the aligned audio/text groups indexed by $\mathcal{J}_t^v$. The masks assign zero attention bias to admitted positions and $-\infty$ elsewhere. They control direct audio/text cross-attention in the Temporal Transformer; gesture self-attention remains causal. Both views share the same lower boundary; only the future-conditioned view extends access beyond t. The scopes specify its upper boundary:

- **Word-bounded:** the end of the word containing the current aligned timestamp.
- **Segment-bounded:** the end of the transcript segment containing the current aligned timestamp.
- **Full-clip:** the last speech group of the training clip, $B_\sigma(t) = T - 1$.

Supplementary Section S1.1 specifies timestamp alignment and boundary cases.

From speech context to gesture distributions. The changed speech mask first produces a future-conditioned temporal state $\mathbf{h}_t^{\mathrm{f}}$. This state is also passed to the Kinematic Transformer, carrying the additional context into the remaining 19 predictions. The kinematic stage receives the same current-step speech slice and within-step prefix in both views; it has no separately extended speech window. This yields logits $\mathbf{z}_{b,t,k}^v$ for batch sample b and view v at every stream. The two views predict the same gesture token from these differently conditioned states.

For distillation, we remove the padding-class logit and normalize over the V motion-code entries: $\mathbf{p}_{b,t,k}^{v,\tau_d} = \mathrm{softmax}(\mathbf{z}_{b,t,k,\mathcal{V}}^v / \tau_d)$, where the indexed vector contains all real-code logits and $\tau_d > 0$ is the distillation temperature. Let $m_{b,t,k} \in \{0, 1\}$ indicate a valid target and $\mathcal{K}_+$ contain streams with at least one valid target in the batch. With sg denoting stop-gradient and D_{KL} the Kullback–Leibler (KL) divergence, the *gesture-token distillation loss* is

$$\mathcal{L}_{\mathrm{distill}} = \frac{\tau_d^2}{|\mathcal{K}_+|} \sum_{k \in \mathcal{K}_+} \frac{\sum_{b,t} m_{b,t,k} D_{\mathrm{KL}}\left(\mathrm{sg}[\mathbf{p}_{b,t,k}^{\mathrm{f},\tau_d}] \,\|\, \mathbf{p}_{b,t,k}^{\mathrm{c},\tau_d}\right)}{\sum_{b,t} m_{b,t,k}}. \tag{3}$$

We set $\mathcal{L}_{\mathrm{distill}} = 0$ when $\mathcal{K}_+ = \varnothing$. Each valid stream, including facial streams, contributes its mean forward KL with equal weight. Detached future-conditioned distributions supervise causal predictions of the same gesture token, and the shared parameters are updated through the causal pass. The supervisory role is defined by this objective; richer context does not guarantee higher held-out token accuracy.

3.3 GESTURE TOKEN REGULARIZATION

Under standard teacher forcing, the kinematic predictor uses reference token prefixes; at inference it uses its own predictions. Gesture token regularization exposes it to alternative prefixes through codebook-distance supervision and stochastic code sampling adapted from RQ-Transformer (Lee et al., 2022).

At zero-based regional depth ℓ, let $\mathbf{r}_\ell$ be the residual and $\boldsymbol{\pi}_\ell$ its codebook-distance target distribution, omitting sample, time and region indices. Select $\widetilde{g}_\ell$ by sampling from $\boldsymbol{\pi}_\ell$ with probability $p_s \in [0, 1]$, otherwise choosing the nearest code. Before constructing the next distribution, update

$$\mathbf{r}_{\ell+1} = \mathbf{r}_\ell - \mathbf{e}_{\ell,\widetilde{g}_\ell}. \tag{4}$$

where $\mathbf{e}_{\ell,\widetilde{g}_\ell}$ is the decoded code vector in residual space. Each region retains its fixed depth and restarts the residual traversal. Both views reuse the same sampled within-step prefix; temporal history remains canonical.

For $k > 0$, token supervision mixes selected-code and codebook-distance cross-entropy; $k = 0$ retains the canonical hard target. The regularized causal control uses the same procedure. Supplementary Section S1.2 defines the codebook-distance distribution, mixture and stream-specific prefixes.

3.4 TRAINING OBJECTIVE AND CAUSAL GENERATION

The complete objective combines the token loss $\mathcal{L}_{\text{token}}$, the distillation loss $\mathcal{L}_{\text{distill}}$, and the inherited voice-activity loss $\mathcal{L}_{\text{VAD}}$ and regional latent contrastive loss $\mathcal{L}_{\text{con}}$ (Mughal et al., 2026):

$$\mathcal{L} = \mathcal{L}_{\text{token}} + \lambda_d \mathcal{L}_{\text{distill}} + \lambda_v \mathcal{L}_{\text{VAD}} + \lambda_c \mathcal{L}_{\text{con}}. \tag{5}$$

The weights λ_d, λ_v and λ_c govern distillation, voice activity and contrastive learning, respectively, and may vary over training. Reference coefficient schedules are specified in Supplementary Section S2. Each training step constructs temporal history from canonical codec tokens and, when enabled, one stochastic residual path, then evaluates the causal pass. When $\lambda_d > 0$, it also evaluates the future-conditioned pass without gradient tracking. The combined objective updates the shared generator; the frozen codecs receive no updates. Setting $\lambda_d = 0$ removes future-speech supervision for the regularized causal control.

At inference, only the causal view executes. It samples the first token from the temporal state, generates the remaining streams in kinematic order, reconstructs motion with the regional codecs, and feeds the generated token vector into the next step. No future-conditioned pass or later speech is introduced by the training objective. Section 4 evaluates generated-motion quality across context scopes and speaker settings under this common deployment procedure.

The causal boundary is defined on supplied, aligned speech tokens. Evaluation uses cached audio codes and timestamp-aligned text tokens, embedded through Moshi-initialized lookup tables (Défossez et al., 2024), without running a full-sequence contextual language model. The generator attends only to current and preceding supplied speech positions. Each 80-ms step produces two motion frames. Supplementary Section S4 explains alignment and the separate requirement of token availability in an online frontend.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETTING

Datasets. We evaluate on BEATX, the mesh-based extension of BEAT introduced with EMAGE (Liu et al., 2022a; 2024), which pairs speech audio and transcripts with body, hand, and facial motion represented by SMPL-X (Pavlakos et al., 2019). We use two settings: a single-speaker study on Scott to compare training components and future-context scopes with speaker identity held fixed, and a multi-speaker study across 23 retained speakers to assess the intervention under broader speaker coverage. Training uses 10-second clips with motion sampled at 25 fps and audio at 24 kHz; evaluation uses complete held-out utterances. ForeMotion generates motion autoregressively using only its causal view at evaluation.

Evaluation metrics. We report Fréchet Gesture Distance (FGD) (Yoon et al., 2020) to compare generated and reference motion-feature distributions, and beat alignment (BeatAlign) to quantify correspondence between detected acoustic and kinematic beats (Mughal et al., 2026). We also report L1 diversity (L1-Div), computed as the mean L1 deviation of joint positions from their sequence mean, and facial mean-squared error (Facial-MSE) against reference facial motion. Lower FGD and Facial-MSE are preferred. For BeatAlign and L1-Div, we evaluate closeness to the corresponding ground-truth values, indicated by $\rightarrow$. For perceptual evaluation, we conduct a blind pairwise naturalness study with 25 participants using 15 held-out utterances from the single-speaker setting. Each trial pairs ForeMotion's output with a video of the same utterance from one of four baselines or ground-truth motion. System identities are concealed, and participants select the video whose motion appears more natural.

Implementation details. The generator has four temporal layers (two heads) and two kinematic layers (one head), with 2,048 entries per motion codebook. The documented reference configuration uses $\tau_d = 1$, eight nearest RVQ codes and $\tau_r = 0.5$, with p_s ramping to 0.15 and β to 0.5. Regional codecs are frozen. Speech and motion-token embeddings are initialized from Moshi and the regional codecs, respectively, and remain trainable; the generator does not resume a gesture-generator checkpoint. Supplementary Section S2 provides schedules, optimization settings and the reference decoder; Supplementary Section S3.1 records speaker selection and splits.

4.2 Main results

Table 1: Reported BEATX results in the single- and multi-speaker settings. ForeMotion uses full-clip supervision. Streaming marks the causal online group (✓); × marks other methods. DiffSHEG is listed with the offline baselines but supports chunk-based streaming. Lower FGD and Facial-MSE are preferred. Bold and underlined values mark the best and second-best reported scores per column, excluding ground truth; BeatAlign and L1-Div are ranked by absolute distance to ground truth. Baselines are evaluated from released checkpoints; ForeMotion is trained in this study. Baseline checkpoint sources are summarized in Supplementary Section S3.2. Dashes denote unavailable values for the stated metric.

		Single speaker (Scott)				Multi-speaker (23 speakers)			
Method	Streaming	FGD ↓	BeatAlign →	L1-Div →	Facial-MSE ↓	FGD ↓	BeatAlign →	L1-Div →	Facial-MSE ↓
Ground truth	—	—	0.749	13.22	—	—	0.446	8.45	—
Offline methods									
HA2G	×	1.232	0.678	8.63	—	—	—	—	—
CaMN	×	0.969	0.698	10.61	—	0.736	0.176	<u>6.73</u>	—
TalkSHOW	×	0.615	0.686	**13.12**	7.79	—	—	—	—
DiffSHEG	×	0.899	0.714	11.91	7.67	—	—	—	—
EMAGE	×	0.552	0.795	<u>13.06</u>	7.68	0.859	0.236	6.58	<u>4.60</u>
MambaTalk	×	0.530	0.779	12.99	**6.25**	1.375	0.080	3.73	**4.12**
GestureLSM	×	**0.410**	0.729	13.76	—	0.537	0.481	**8.41**	—
PyraMotion	×	<u>0.461</u>	**0.742**	13.42	<u>7.18</u>	—	—	—	—
Online methods									
MIBURI	✓	0.753	0.790	15.85	8.85	<u>0.480</u>	<u>0.461</u>	10.44	7.77
ForeMotion	✓	0.583	<u>0.731</u>	13.77	7.82	**0.470**	**0.453**	10.42	5.21

Table 1 compares full-clip ForeMotion with offline and online methods: HA2G (Liu et al., 2022b), CaMN (Liu et al., 2022a), TalkSHOW (Yi et al., 2023), EMAGE (Liu et al., 2024), MambaTalk (Xu et al., 2024), GestureLSM (Liu et al., 2025c), PyraMotion (Yin et al., 2025), DiffSHEG (Chen et al., 2024), and MIBURI (Mughal et al., 2026). Baseline checkpoint sources are detailed in Supplementary Section S3.2. The MIBURI comparison evaluates the complete ForeMotion training procedure, including gesture token regularization and paired-view distillation.

Single-speaker results. On Scott, ForeMotion reduces FGD from MIBURI's 0.753 to 0.583, a 22.6% reduction, and lowers Facial-MSE from 8.85 to 7.82. BeatAlign moves from 0.790 to 0.731,

closer to the reference value of 0.749, while L1-Div moves from 15.85 to 13.77, closer to the reference value of 13.22. Four offline comparison methods attain lower single-speaker FGD. ForeMotion thus improves several reported measures over the online reference, while retaining a causal generation path, without leading every metric or comparison system. This complete-system comparison does not isolate the effect of adding paired-view distillation.

Multi-speaker results. The 23-speaker study compares the complete ForeMotion procedure with released baselines; the regularized-control comparison is single-speaker. ForeMotion obtains the lowest reported FGD in Table 1, at 0.470, close to MIBURI's 0.480. These are descriptive point estimates; uncertainty across independent training runs is not reported. Facial-MSE decreases from 7.77 to 5.21 relative to MIBURI, a 32.9% reduction, while BeatAlign moves closer to ground truth and L1-Div remains similar. EMAGE and MambaTalk retain lower Facial-MSE than ForeMotion. The two speaker settings therefore show different patterns of measured benefit; they do not establish a speaker-scaling trend or statistical equivalence in body-gesture quality.

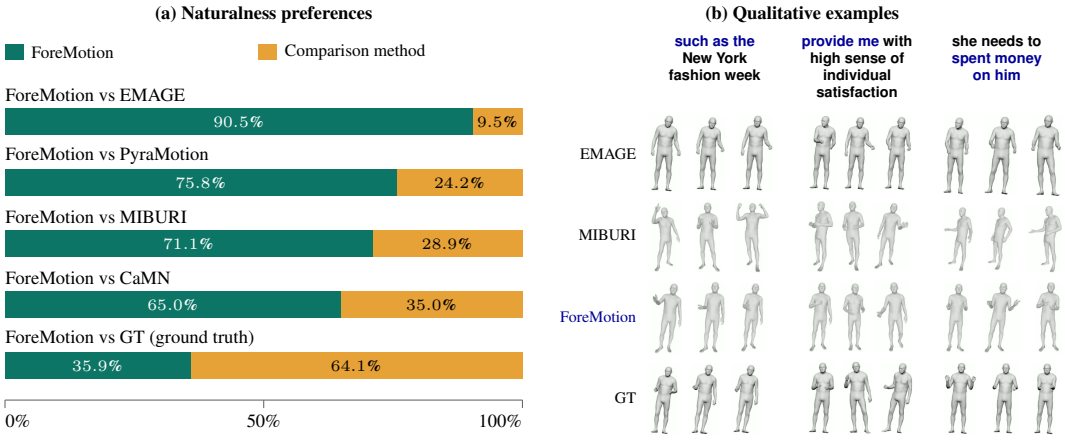


Figure 3: Perceptual evaluation and qualitative examples. (a) Pooled naturalness preferences from 25 participants over 15 single-speaker utterances; each bar shows the shares choosing ForeMotion and its named comparator. The rates are descriptive. (b) Selected motion frames from EMAGE, MIBURI, ForeMotion, and ground truth (GT) for three speech excerpts, shown above their corresponding example groups.

Perceptual results. Figure 3(a) reports the naturalness judgments from the single-speaker study. ForeMotion is selected more often than each of the four evaluated baselines, with pooled preference rates from 65.0% to 90.5%. In particular, participants favor ForeMotion over EMAGE and PyraMotion despite their lower FGD, illustrating why the distributional and perceptual measures are reported separately. Ground-truth motion remains preferred, with ForeMotion selected in 35.9% of those comparisons. These pooled rates are descriptive and do not account for repeated judgments by the same participant. The study assesses naturalness, and does not include GestureLSM or MambaTalk.

Qualitative examples. Figure 3(b) compares selected motion frames from ForeMotion, EMAGE, MIBURI, and ground truth for three speech excerpts. The examples show the generated arm and hand postures alongside reference motion, complementing the quantitative and perceptual evaluations.

4.3 Ablation Study

Table 2(a) compares the released MIBURI reference, the trained regularized causal control, and full-clip ForeMotion. Panel (b) varies future-context scope with both components retained.

Distillation comparison. Full-clip ForeMotion records lower FGD than the regularized causal control (0.583 versus 0.736, a 20.8% reduction) and lower Facial-MSE (7.82 versus 8.51). These

Table 2: Single-speaker component and context-scope comparisons (✓: enabled; ×: disabled). In (a), the first row reuses the released-checkpoint MIBURI evaluation from Table 1; the remaining rows are trained variants. Both final rows are full-clip ForeMotion; (b) retains both components. BeatAlign and L1-Div are read against Table 1's ground truth.

(a) Training components

Token reg.	Distill.	FGD ↓	BeatAlign →	L1-Div →	Facial-MSE ↓
×	×	0.753	0.790	15.85	8.85
✓	×	0.736	0.749	13.33	8.51
✓	✓	0.583	0.731	13.77	7.82

(b) Future-context scope

Training variant	FGD ↓	BeatAlign →	L1-Div →	Facial-MSE ↓
w/ word-bounded context	0.767	0.742	11.69	8.41
w/ segment-bounded context	0.703	0.736	12.12	7.71
ForeMotion	0.583	0.731	13.77	7.82

gains are selective: BeatAlign moves from 0.749 to 0.731 and L1-Div from 13.33 to 13.77, farther from their reference values of 0.749 and 13.22. The improvement therefore depends on the evaluated property of the generated motion.

Effect of context scope. Word-bounded supervision yields an FGD of 0.767, worse than the control, while segment-bounded supervision reaches 0.703 and full-clip supervision obtains the lowest FGD. Segment-bounded supervision instead gives the lowest Facial-MSE, at 7.71. The preferred scope thus depends on the metric. All scopes retain identical past/current speech windows across the two views (Section 3.2). These comparisons vary both the duration and content of future speech; they do not isolate its informative content from other effects of paired-view training.

Role of gesture token regularization. The regularized causal control has lower FGD than the released MIBURI reference (0.736 versus 0.753) and L1-Div closer to ground truth (13.33 versus 15.85). This comparison includes different training histories. Every distillation variant retains token regularization; their interaction is untested. The reported differences are point estimates without uncertainty estimates.

5 Conclusion and Future Work

We presented ForeMotion, a framework that uses expanded training speech context to supervise causal co-speech gesture generation through paired-view gesture distillation. A future-conditioned view of the shared generator provides detached gesture-token distributions during training; inference executes only the causal view with its original speech access. On the evaluated backbone, full-clip ForeMotion has lower single-speaker FGD and Facial-MSE than the regularized causal control, although beat alignment and L1-Div do not improve. In the single-speaker scope study, full-clip supervision yields the lowest FGD, whereas segment-bounded supervision yields the lowest Facial-MSE. The preferred context scope therefore depends on the metric. In the multi-speaker complete-system comparison, ForeMotion has lower Facial-MSE than MIBURI, while body FGD remains close to that reference. The naturalness study additionally records descriptive preferences for ForeMotion over the four evaluated baselines. These findings support the evaluated training procedure as a means of improving selected gesture-quality measures while preserving the generator's inference-time speech access. Its benefits depend on context scope, speaker setting, and metric.

Future work. Our current formulation models monologue: the generator conditions on the speaker's own audio, text, and identity, without an interlocutor input. Extending ForeMotion to dyadic conversation would test whether training-only speech context remains useful when gestures also respond to a partner. One hypothesis is that later speech provides clues to the speaker's evolving communicative plan. A partner's upcoming behavior reflects another participant's responses and intentions and is not directly observable in advance. A dyadic study could compare supervision informed by the speaker's future speech, the partner's future speech, or both, while restricting inference to past and current observations from both participants. Establishing when partner-informed supervision helps, fails to help, or interferes with gesture generation would clarify the limits of privileged supervision beyond monologue.

AI USE STATEMENT

Yes, to aid or polish writing. Details are described in the paper. AI tools assisted with manuscript preparation, including polishing wording, grammar, clarity, and readability. The authors are responsible for the final content.

REPRODUCIBILITY STATEMENT

Section 3 describes ForeMotion's architecture, speech-access constraints, and training objectives, while Section 4.1 summarizes the experimental setting. Supplementary Section S1 provides the token-target formulation and loss reductions; Supplementary Section S2 documents reference configurations, optimization schedules, initialization, and generation settings. Supplementary Section S3 describes the dataset and evaluation protocols, including baseline checkpoint provenance and perceptual evaluation. Supplementary Section S4 explains the speech-input interface and incremental generation procedure. Code and sample output videos will be included in the supplementary archive. We will publicly release the training code upon acceptance.

REFERENCES

Louis Abel, Vincent Colotte, and Slim Ouni. Towards realtime co-speech gestures synthesis using STARGATE. In *Proceedings of Interspeech 2024*, pp. 4893–4897, 2024. doi: 10.21437/Interspeech.2024-302. URL https://www.isca-archive.org/interspeech_2024/abel24_interspeech.html.

Tenglong Ao, Zeyi Zhang, and Libin Liu. GestureDiffuCLIP: Gesture diffusion model with CLIP latents. *ACM Transactions on Graphics*, 42(4):42, 2023. doi: 10.1145/3592097. URL https://pku-mocca.github.io/GestureDiffuCLIP-Page/.

Justine Cassell. Embodied conversational interface agents. *Communications of the ACM*, 43(4): 70–78, 2000. doi: 10.1145/332051.332075. URL https://doi.org/10.1145/332051.332075.

Justine Cassell, Timothy Bickmore, Mark Billinghurst, Lee Campbell, Kenny Chang, Hannes Vilhjálmsson, and Hao Yan. An architecture for embodied conversational characters. In *Proceedings of the First Workshop on Embodied Conversational Characters*, 1998. URL https://www.justinecassell.com/publications/wecc_final.pdf.

Junming Chen, Yunfei Liu, Jianan Wang, Ailing Zeng, Yu Li, and Qifeng Chen. DiffSHEG: A diffusion-based approach for real-time speech-driven holistic 3D expression and gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7352–7361, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Chen_DiffSHEG_A_Diffusion-Based_Approach_for_Real-Time_Speech-driven_Holistic_3D_Expression_CVPR_2024_paper.html.

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. Moshi: A speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00037*, 2024. URL https://arxiv.org/abs/2410.00037.

Ed Donnellan, Levent Emir Özder, Hillarie Man, Beata Grzyb, Yan Gu, and Gabriella Vigliocco. Timing relationships between representational gestures and speech: A corpus based investigation. In Jennifer Culbertson, Amy Perfors, Hugh Rabagliati, and Verónica Ramenzoni (eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society*, pp. 2052–2058, 2022. URL https://escholarship.org/uc/item/7w349725.

Pengcheng Fang, Hongli Chen, and Xiaohao Cai. Privileged foresight distillation: Zero-cost future correction for world action models. *arXiv preprint arXiv:2604.25859*, 2026. URL https://arxiv.org/abs/2604.25859.

Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. In *Proceedings of the 35th International Conference on Machine*

Learning, volume 80 of *Proceedings of Machine Learning Research*, pp. 1607–1616. PMLR, 2018. URL https://proceedings.mlr.press/v80/furlanello18a.html.

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. doi: 10.1111/cgf.14734. URL https://doi.org/10.1111/cgf.14734.

Ikhsanul Habibie, Weipeng Xu, Dushyant Mehta, Lingjie Liu, Hans-Peter Seidel, Gerard Pons-Moll, Mohamed Elgharib, and Christian Theobalt. Learning speech-driven 3D conversational gestures from video. In *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents*, pp. 101–108, 2021. doi: 10.1145/3472306.3478335. URL https://vcai.mpi-inf.mpg.de/projects/3d_speech_driven_gesture/.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. URL https://arxiv.org/abs/1503.02531.

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*, 2020. doi: 10.1145/3382507.3418815. URL https://svito-zar.github.io/gesticulator/.

Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11523–11532, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Lee_Autoregressive_Image_Generation_Using_Residual_Quantization_CVPR_2022_paper.html.

Jia Li, Wenjie Zhao, Fnu Atisri, Sanskriti Aripineni, Shijian Deng, Jon E. Froehlich, Yuhang Zhao, and Yapeng Tian. How much future helps? a controlled study of future-privileged supervision for causal egocentric gaze estimation. *arXiv preprint arXiv:2607.01437*, 2026. URL https://arxiv.org/abs/2607.01437. Accepted to the 7th International Workshop on Eye and Gaze in Computer Vision (GAZE 2026).

Jing Li, Di Kang, Wenjie Pei, Xuefei Zhe, Ying Zhang, Zhenyu He, and Linchao Bao. 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. URL https://openaccess.thecvf.com/content/ICCV2021/html/Li_Audio2Gestures_Generating_Diverse_Gestures_From_Speech_Audio_With_Conditional_Variational_ICCV_2021_paper.html.

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 *Computer Vision – ECCV 2022*, pp. 612–630, 2022a. doi: 10.1007/978-3-031-20071-7_36. URL https://www.ecva.net/papers/eccv_2022/papers_ECCV/html/296_ECCV_2022_paper.php.

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*, pp. 1144–1154, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Liu_EMAGE_Towards_Unified_Holistic_Co-Speech_Gesture_Generation_via_Expressive_Masked_CVPR_2024_paper.html.

Haiyang Liu, Xingchao Yang, Tomoya Akiyama, Yuantian Huang, Qiaoge Li, Shigeru Kuriyama, and Takafumi Taketomi. TANGO: Co-speech gesture video reenactment with hierarchical audio motion embedding and diffusion interpolation. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=LbEWwJOufy.

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*, pp. 13963–13973, 2025b. URL https://openaccess.thecvf.com/content/ICCV2025/html/Liu_SemGes_Semantics-aware_Co-Speech_Gesture_Generation_using_Semantic_Coherence_and_Relevance_ICCV_2025_paper.html.

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*, pp. 10929–10939, 2025c. URL https://openaccess.thecvf.com/content/ICCV2025/html/Liu_GestureLSM_Latent_Shortcut_based_Co-Speech_Gesture_Generation_with_Spatial-Temporal_Modeling_ICCV_2025_paper.html.

Xian Liu, Qianyi Wu, Hang Zhou, Yinghao Xu, Rui Qian, Xinyi Lin, Xiaowei Zhou, Wayne Wu, Bo Dai, and Bolei Zhou. Learning hierarchical cross-modal association for co-speech gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10462–10472, 2022b. URL https://openaccess.thecvf.com/content/CVPR2022/html/Liu_Learning_Hierarchical_Cross-Modal_Association_for_Co-Speech_Gesture_Generation_CVPR_2022_paper.html.

David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. In *4th International Conference on Learning Representations*, 2016. URL https://arxiv.org/abs/1511.03643.

Palmer Morrel-Samuels and Robert M. Krauss. Word familiarity predicts temporal asynchrony of hand gestures and speech. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18(3):615–622, 1992. doi: 10.1037/0278-7393.18.3.615. URL https://doi.org/10.1037/0278-7393.18.3.615.

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*, pp. 16578–16588, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Mughal_Retrieving_Semantics_from_the_Deep_an_RAG_Solution_for_Gesture_CVPR_2025_paper.html.

M. Hamza Mughal, Rishabh Dabral, Vera Demberg, and Christian Theobalt. MIBURI: Towards expressive interactive gesture synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/papers/Mughal_MIBURI_Towards_Expressive_Interactive_Gesture_Synthesis_CVPR_2026_paper.pdf.

Muhammad Hamza Mughal, Rishabh Dabral, Ikhsanul Habibie, Lucia Donatelli, Marc Habermann, and Christian Theobalt. ConvoFusion: Multi-modal conversational diffusion for co-speech gesture synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1388–1398, 2024. doi: 10.1109/CVPR52733.2024.00138. URL https://openaccess.thecvf.com/content/CVPR2024/html/Mughal_ConvoFusion_Multi-Modal_Conversational_Diffusion_for_Co-Speech_Gesture_Synthesis_CVPR_2024_paper.html.

Evonne Ng, Javier Romero, Timur Bagautdinov, Shaojie Bai, Trevor Darrell, Angjoo Kanazawa, and Alexander Richard. From audio to photoreal embodiment: Synthesizing humans in conversations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1001–1010, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Ng_From_Audio_to_Photoreal_Embodiment_Synthesizing_Humans_in_Conversations_CVPR_2024_paper.html.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10975–10985, 2019. URL https://smpl-x.is.tue.mpg.de/.

Xingqun Qi, Yatian Wang, Hengyuan Zhang, Jiahao Pan, Wei Xue, Shanghang Zhang, Wenhan Luo, Qifeng Liu, and Yike Guo. Co^{3}Gesture: Towards coherent concurrent co-speech 3d gesture generation with interactive diffusion. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=VaowElpVzd.

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*, pp. 2264–2273, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Saleem_LiveGesture_Streamable_Co-Speech_Gesture_Generation_Model_CVPR_2026_paper.html.

Jiakai Tang, Yang Zhang, See-Kiong Ng, Xu Chen, Wen Chen, Jian Wu, and Han Zhu. Learning from the future: Privileged self-distillation for sequential recommendation. *arXiv preprint arXiv:2607.27055*, 2026. URL https://arxiv.org/abs/2607.27055.

Marlijn ter Bekke, Linda Drijvers, and Judith Holler. Hand gestures have predictive potential during conversation: An investigation of the timing of gestures in relation to speech. *Cognitive Science*, 48(1):e13407, 2024. doi: 10.1111/cogs.13407. URL https://doi.org/10.1111/cogs.13407.

Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural Networks*, 22(5–6):544–557, 2009. doi: 10.1016/j.neunet.2009.06.042. URL https://doi.org/10.1016/j.neunet.2009.06.042.

Evgeniia Vu, Andrei Boiarov, and Dmitry Vetrov. Streaming generation of co-speech gestures via accelerated rolling diffusion. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40 (31):26054–26061, 2026. doi: 10.1609/aaai.v40i31.39807. URL https://ojs.aaai.org/index.php/AAAI/article/view/39807.

Petra Wagner, Zofia Malisz, and Stefan Kopp. Gesture and speech in interaction: An overview. *Speech Communication*, 57:209–232, 2014. doi: 10.1016/j.specom.2013.09.008. URL https://doi.org/10.1016/j.specom.2013.09.008.

Zunnan Xu, Yukang Lin, Haonan Han, Sicheng Yang, Ronghui Li, Yachao Zhang, and Xiu Li. MambaTalk: Efficient holistic gesture synthesis with selective state space models. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-0633. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/23c9c94227f937cfb50592a15e7fbb63-Abstract-Conference.html.

Hongwei Yi, Hualin Liang, Yifei Liu, Qiong Cao, Yandong Wen, Timo Bolkart, Dacheng Tao, and Michael J. Black. Generating holistic 3D human motion from speech. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 469–480, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/papers/Yi_Generating_Holistic_3D_Human_Motion_From_Speech_CVPR_2023_paper.pdf.

Zhizhuo Yin, Yuk Hang Tsui, and Pan Hui. PyraMotion: Attentional pyramid-structured motion integration for co-speech 3D gesture synthesis. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-4634. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/ca6403397d60b7dbf9c7158a1c19094a-Abstract-Conference.html.

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):222, 2020. doi: 10.1145/3414685.3417838. URL https://doi.org/10.1145/3414685.3417838.

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), 2024. doi: 10.1145/3658134. URL https://pku-mocca.github.io/Semantic-Gesticulator-Page/.

Mingkuan Zhao, Xiayu Sun, Wentao Hu, Suquan Chen, Jiaxuan Li, Xiaoyan Zhu, Xin Lai, and Jiayin Wang. Regret pre-training: Bridging prior and posterior views for enhanced knowledge grounding. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026. URL https://arxiv.org/abs/2606.03080.

Weiyu Zhao, Liangxiao Hu, and Shengping Zhang. DiffuGesture: Generating human gesture from two-person dialogue with diffusion models. In *Companion Publication of the 25th International Conference on Multimodal Interaction*, 2023. doi: 10.1145/3610661.3616552. URL https://doi.org/10.1145/3610661.3616552.