TalkFlow: Structured Latent Flow Matching for Speech Driven Holistic Motion Generation

XINMU WANG, Stony Brook University, USA
FUMI HONDA, Stony Brook University, USA
XIANG GAO, Stony Brook University, USA
XIANFENG GU, Stony Brook University, USA
XIYUN SONG, Futurewei Technologies Inc, USA
HEATHER YU, Futurewei Technologies Inc, USA

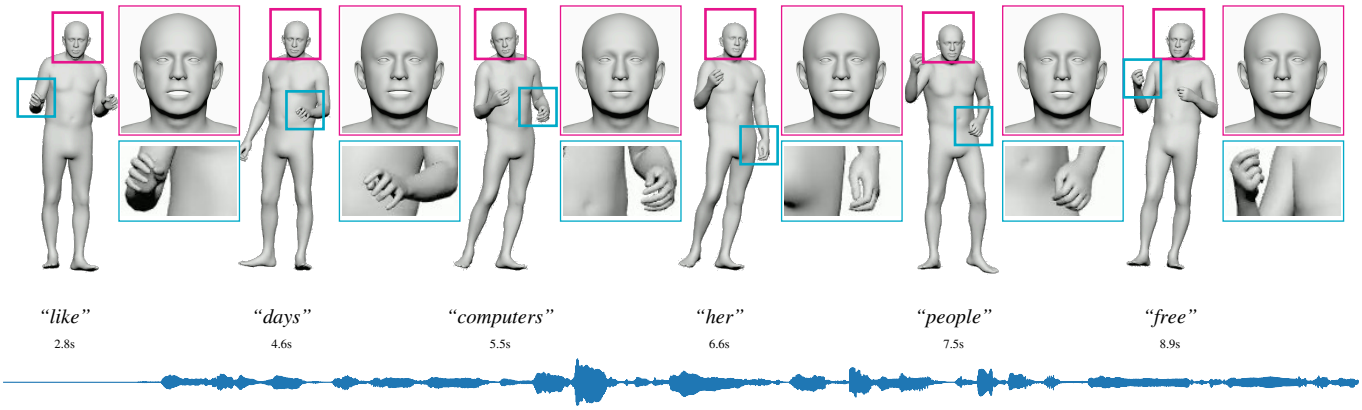


Fig. 1. **TalkFlow** generates coordinated body, hand, and facial motion from speech via structured latent flow matching. Conditioned on speech audio and aligned text features, our method synthesizes speech-aligned holistic motion: each frame shows the generated body pose with zoomed views of the face and hand. Frames are sampled at the onset of content words across a 9-second clip, with the corresponding waveform shown at the bottom.

Holistic co-speech motion generation aims to synthesize coordinated facial expressions, hand gestures, and body movements from speech, which is challenging due to heterogeneous part-specific dynamics, long-range temporal coherence, and cross-part coordination. Directly modeling high-dimensional motion can be unstable, while overly compact or unstructured latent spaces may lose fine-grained motion details. We present **TalkFlow**, a structured latent flow framework for speech-driven holistic motion generation. TalkFlow learns a continuous three-stream latent representation for upper-body motion, lower-body motion, and facial expressions, preserving part-specific dynamics without forcing heterogeneous motion components into a single latent space. To maintain holistic coordination, lightweight cross-stream interactions are incorporated in both latent representation learning and flow-based generation. Conditioned on multi-level speech features, TalkFlow performs speech-conditioned latent flow matching, learning a continuous transport process from source latents to realistic motion latents. Experiments on the BEAT2 dataset demonstrate strong performance across facial expression accuracy, body gesture quality, and speech-motion synchronization. Qualitative comparisons and user studies further show that TalkFlow generates more natural, expressive, and speech-aligned holistic motion than state-of-the-art methods. The project page is available at https://xinmuwang.github.io/TalkFlow/.

CCS Concepts: • **Computing methodologies** → **Computer vision**; **Animation**.

Additional Key Words and Phrases: Motion Generation, Flow Matching

ACM Reference Format:
Xinmu Wang, Fumi Honda, Xiang Gao, Xianfeng Gu, Xiyun Song, and Heather Yu. 2026. TalkFlow: Structured Latent Flow Matching for Speech Driven Holistic Motion Generation. In *SIGGRAPH Asia 2026 Conference Papers (SA Conference Papers '26), December 01–04, 2026, Kuala Lumpur, Malaysia.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3829340.3842366

Authors' Contact Information: Xinmu Wang, Department of Computer Science, Stony Brook University, Stony Brook, USA, xinmuwang@cs.stonybrook.edu; Fumi Honda, Department of Computer Science, Stony Brook University, Stony Brook, USA, fhonda@cs.stonybrook.edu; Xiang Gao, Department of Computer Science, Stony Brook University, Stony Brook, USA, gao2@cs.stonybrook.edu; Xianfeng Gu, Department of Computer Science, Stony Brook University, Stony Brook, USA, gu@cs.stonybrook.edu; Xiyun Song, Futurewei Technologies Inc, Santa Clara, USA, xsong@futurewei.com; Heather Yu, Futurewei Technologies Inc, Santa Clara, USA, hyu@futurewei.com.

This work is licensed under a Creative Commons Attribution 4.0 International License.
SA Conference Papers '26, Kuala Lumpur, Malaysia
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2842-6/2026/12
https://doi.org/10.1145/3829340.3842366

1 Introduction

Generating expressive human motion from speech is a central problem for digital avatars, virtual agents, augmented and virtual reality, and embodied communication. With parametric human models such as SMPL-X [Pavlakos et al. 2019], holistic human motion can be represented in a unified form that includes body pose, hand articulation, and facial expressions. Recent co-speech datasets and

benchmarks, such as BEAT2 [Liu et al. 2024b], further make it possible to study speech-driven 3D human motion at scale. However, generating realistic holistic motion remains challenging because speech-driven facial expressions, hand gestures, upper-body movements, and lower-body posture exhibit heterogeneous dynamics while requiring temporal coherence and cross-part coordination.

Existing co-speech motion generation methods have made significant progress. Early learning-based approaches mainly relied on deterministic regression [Habibie et al. 2021; Ng et al. 2022], which often averages over multiple plausible gestures. More recent methods improve diversity and realism through discrete latent modeling, masked modeling, diffusion, or sequence modeling architectures [Chen et al. 2024; Liu et al. 2024b,a; Xu et al. 2024a; Yi et al. 2023; Zhang et al. 2025b]. Despite these advances, several limitations remain. Directly generating high-dimensional SMPL-X motion is difficult due to complex spatial–temporal dependencies. Latent generative models alleviate this issue, but overly compact or unstructured latent spaces may weaken fine-grained facial dynamics and body-part coordination. In particular, codebook-based motion latents introduce a discrete bottleneck that may cause quantization errors and limit the continuity required for subtle expression changes, smooth gesture transitions, and coordinated holistic motion. In addition, diffusion-based generation [Ho et al. 2020] often requires iterative denoising, which can be inefficient and may lead to inconsistent temporal evolution for long sequences.

To address these challenges, we propose **TalkFlow**, a structured latent flow framework for speech-driven holistic motion generation. TalkFlow first learns a compact continuous motion latent space using a structure-aware autoencoder, avoiding the discretization artifacts introduced by codebook-based motion tokens. The original SMPL-X axis-angle rotations are converted to the continuous 6D rotation representation [Zhou et al. 2019]. Rather than splitting the body before spatial encoding, TalkFlow first encodes the full SMPL-X pose graph with a holistic GCN to preserve global body topology. At the spatial encoding stage, full-body pose and facial expression are represented as two streams; the temporal encoder subsequently separates the pose representation into upper- and lower-body branches, yielding three bottleneck streams for upper-body motion, lower-body motion, and facial expressions.

The three-stream design preserves part-specific dynamics and avoids forcing heterogeneous motion components into a single latent space. Upper-body gestures are closely related to communicative intent and hand articulation, lower-body motion provides posture and global stability, and facial expressions capture fine-grained speech-related dynamics. However, these components are not independent in natural communication. We therefore introduce multi-stage cross-stream interactions in the autoencoder, including full pose–expression interaction before temporal compression, latent-level interaction among motion streams, and pose–expression interaction after temporal decoding. These interactions are implemented through lightweight FiLM-style [Perez et al. 2018] modulation and gated residual updates. This allows the model to exchange coordination cues across body and expression representations while preserving specialized stream dynamics.

Given speech input, TalkFlow performs generation through speech-conditioned latent flow matching [Lipman et al. 2022]. Multi-level acoustic and semantic features, including HuBERT [Hsu et al. 2021], Librosa [McFee et al. 2015], FastText [Bojanowski et al. 2017], and SBERT [Reimers and Gurevych 2019], are encoded into branch-specific conditions for the three motion streams. The flow model learns a continuous transport process from source latents to realistic motion latents using stream-specific Transformer branches, speech condition attention, and lightweight FiLM-gated cross-stream coordination. This design enables efficient latent-space generation while maintaining temporal coherence and holistic motion consistency.

By combining holistic graph-based motion encoding, continuous three-stream latent representation learning, and speech-conditioned latent flow matching, TalkFlow generates expressive and coordinated talking human motion. Experiments on the BEAT2 dataset [Liu et al. 2024b] demonstrate strong facial fidelity, competitive body-gesture quality, and effective speech-motion synchronization under the standard Speaker-2 evaluation protocol. Qualitative comparisons and user studies further show that TalkFlow produces more natural, speech-aligned, and perceptually preferred holistic motion compared with existing methods.

We summarize our main contributions as follows:

- We propose **TalkFlow**, a structured latent flow framework for speech-driven holistic motion generation that synthesizes coordinated facial expressions, upper-body gestures, and lower-body motion from speech.
- We introduce a **continuous structure-aware three-stream motion representation** that encodes the full-body SMPL-X pose graph and models upper-body motion, lower-body motion, and facial expressions with separate but interactive temporal streams.
- We formulate speech-driven holistic motion generation as **speech-conditioned latent flow matching**, using branch-specific speech conditions and a three-stream velocity backbone to achieve strong facial fidelity, competitive holistic motion performance, and improved perceptual quality on BEAT2 [Liu et al. 2024b].

2 Related Work

2.1 Co-speech and holistic motion generation.

Human motion generation can be conditioned on text, audio, scene context, or other control signals [Zhu et al. 2023]. Among them, speech-driven motion generation is most relevant to our work, as it requires motion to follow both acoustic dynamics and communicative semantics. Early co-speech systems relied on rule-based gesture mappings [Cassell et al. 2001; Kopp and Wachsmuth 2004; Levine et al. 2010; Poggi et al. 2005], while learning-based methods later modeled either facial animation [Cudeiro et al. 2019; Fan et al. 2022; Peng et al. 2023; Shen et al. 2023] or body gestures [Ao et al. 2023; Ginosar et al. 2019; Liu et al. 2022d; Yang et al. 2023; Yoon et al. 2020]. These part-specific approaches often struggle to maintain coordination across face, hands, and body. Recent holistic methods address this issue by jointly modeling multiple motion components. Habibie et al. [Habibie et al. 2021] used a shared audio encoder with separate body-part decoders, TalkSHOW [Yi et al. 2023] combined probabilistic gesture generation with facial animation, DiffSHEG [Chen et al. 2024] introduced directed face-to-body information flow, and EMAGE [Liu et al. 2024b] used masked gesture

reconstruction and audio-conditioned synthesis. MambaTalk [Xu et al. 2024a] and EchoMask [Zhang et al. 2025b] further explored efficient temporal modeling and speech-guided masked motion generation, while SemTalk [Zhang et al. 2025a] emphasizes frame-level semantic cues for holistic co-speech motion generation. More recently, EchoAvatar [Chen et al. 2026] targets real-time full-body avatar animation from streaming audio. Despite these advances, preserving part-specific dynamics and holistic coordination remains challenging.

2.2 Latent representations for holistic motion.

Latent motion representations are widely used to reduce the complexity of high-dimensional motion generation. VQ-VAE [Van Den Oord et al. 2017] and PQ-VAE [Jegou et al. 2010; Wu and Flierl 2019] provide compact discrete motion priors and have been adopted in co-speech generation. ProbTalk [Liu et al. 2024a] represents facial and upper-body motion in a shared discrete latent space, while Combo [Xu et al. 2024b] introduces part-aware latent decomposition and interaction. HoloDiff [Wang et al. 2026] performs speech-conditioned diffusion in a structured latent space for holistic talking human animation. Building on structured latent modeling, TalkFlow learns continuous upper-body, lower-body, and expression streams and coordinates them through lightweight FiLM-gated interactions, preserving part-specific dynamics without relying on discrete motion tokens.

2.3 Flow-based generative motion modeling.

Flow-based generative models have also been explored for motion synthesis. MoGlow [Henter et al. 2020] introduced normalizing flows for probabilistic and controllable motion generation. More recently, Flow Matching [Lipman et al. 2022] and Rectified Flow [Liu et al. 2022a] provide efficient continuous-time formulations for learning generative transport dynamics. In motion generation, Motion Flow Matching [Hu et al. 2023] applies flow matching to text- and action-conditioned human motion, and MotionLab [Guo et al. 2025] further studies flow-based human motion generation and editing in a unified framework. FLOAT [Ki et al. 2025] demonstrates motion latent flow matching for audio-driven talking portraits, while FlowerDance [Yang et al. 2025] explores efficient flow-based generation for music-driven dance synthesis. These works show the potential of flow-based generation, but they mainly focus on locomotion, text/action-conditioned motion, editing, dance, or talking portrait animation. TalkFlow instead applies latent flow matching to holistic 3D co-speech motion using branch-specific speech conditions and interacting upper-body, lower-body, and expression streams.

3 Method

3.1 Problem Definition

Given a speech clip of length N, our goal is to generate a sequence of holistic talking human motions that are temporally aligned with the speech signal. We adopt SMPL-X [Pavlakos et al. 2019] as the motion representation, which models facial expressions, hand gestures, and full-body poses in a unified parametric form.

Let $A = [a_1, \ldots, a_N]$ denote the input speech features. The target motion consists of articulated body poses and facial expressions. The pose sequence is denoted as $R = [r_1, \ldots, r_N]$, where each pose is represented using the continuous Rot6D representation [Zhou et al. 2019]. Facial expressions are represented by blendshape coefficients $E = [e_1, \ldots, e_N]$. Each frame contains 55×6 pose rotations and 100 expression coefficients. To model heterogeneous body dynamics, we decompose the body pose into upper-body and lower-body components:

$$R = \{R^u, R^l\}, \tag{1}$$

where R^u includes upper-body and hand joints, and R^l includes lower-body joints. Our objective is to learn a speech-conditioned generative model:

$$p(R^u, R^l, E \mid A), \tag{2}$$

which produces coordinated upper-body motion, lower-body motion, and facial expressions.

3.2 Data Preprocessing

We preprocess the BEAT2 motion, audio, and text data into temporally aligned training clips. For motion data, we use SMPL-X pose and expression parameters, where the original axis-angle pose rotations are converted into the continuous 6D rotation representation [Zhou et al. 2019]. We segment each sequence into 240-frame clips with a 60-frame overlap. Since BEAT2 is captured at 30 FPS, each clip corresponds to 8 seconds of motion and is encoded into a latent sequence of length 60 with a temporal downsampling factor of 4. The 60-frame overlap is used for training-clip construction and is distinct from the 8-frame overlap used for long-sequence inference. For audio conditioning, we extract HuBERT features [Hsu et al. 2021] and handcrafted Librosa features [McFee et al. 2015] offline and align them with the motion clips. For textual conditioning, following MambaTalk [Xu et al. 2024a], we pre-extract word-level FastText embeddings [Bojanowski et al. 2017] and sentence-level SBERT embeddings [Reimers and Gurevych 2019] from aligned transcripts, which are used as auxiliary conditioning signals together with audio features.

TalkFlow follows a two-stage design, as shown in Figure 2. The first stage learns a continuous structured motion space that reduces the complexity of high-dimensional SMPL-X motion while preserving part-specific dynamics and holistic coordination. The second stage performs speech-conditioned generation in this learned latent space using flow matching, allowing the flow model to operate on compact, interaction-aware motion latents.

3.3 Stage I: Structured Three-Stream Motion Autoencoder

We learn a structured continuous latent space for upper-body motion, lower-body motion, and facial expressions. A key design choice is to preserve full-body spatial topology before temporal decomposition. Instead of encoding upper- and lower-body poses independently from the beginning, we first encode the complete SMPL-X pose graph with a holistic graph encoder, so that global body structure and cross-limb dependencies are retained before the motion is separated into temporal streams. The graph topology follows the anatomical design of Pose2Mesh [Choi et al. 2020]; detailed topology construction is provided in the supplementary material. We use Chebyshev graph convolutions [Defferrard et al. 2016] in

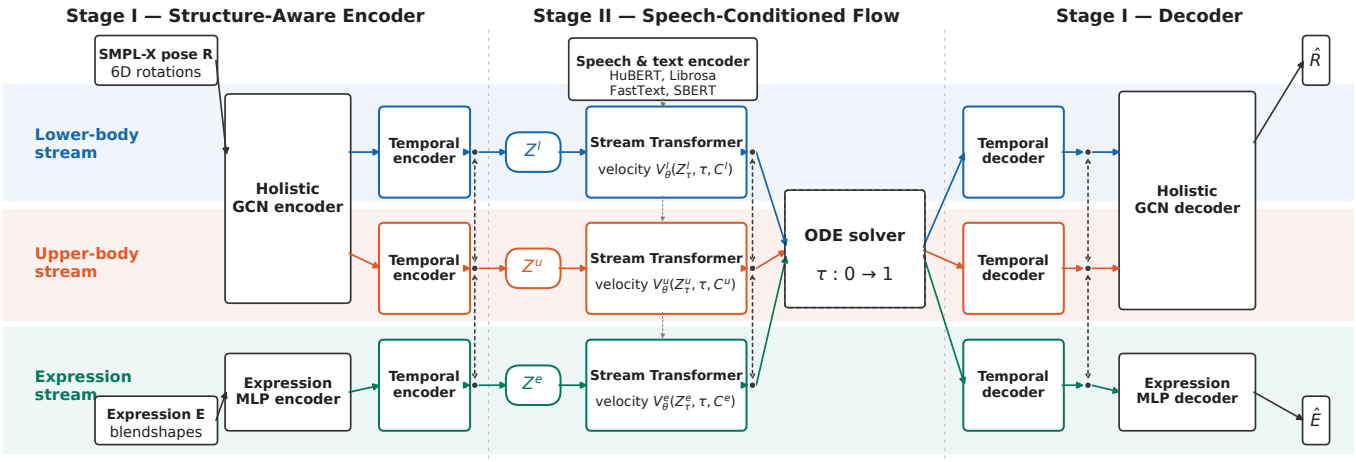


Fig. 2. **Overall architecture of TalkFlow.** Stage I first encodes full-body pose and expression as two spatial branches and then forms upper-body, lower-body, and expression latent streams through temporal encoding. Stage II encodes speech features into branch-specific conditions and generates the three motion latents using a structured flow backbone. Lightweight FiLM-gated interactions coordinate upper–lower and upper–expression streams without direct lower–expression interaction. The frozen Stage-I decoder reconstructs the final SMPL-X motion.

the pose encoder, while facial expressions are encoded by a lightweight frame-level MLP. Thus, the spatial encoding stage is initially pose–expression two-stream, whereas the compressed bottleneck is upper–lower–expression three-stream.

Given a motion sequence (R, E), the spatial encoders produce

$$(H^p, H^p_g) = \mathcal{G}_\psi(R), \quad H^e = \mathcal{M}_\psi(E), \tag{3}$$

where the graph encoder produces frame-level node features $h^p_t \in \mathbb{R}^{55\times 64}$ and global pose features $h^p_{g,t} \in \mathbb{R}^{128}$. Their temporal stacks are denoted by $H^p = [h^p_1, \dots, h^p_T]$ and $H^p_g = [h^p_{g,1}, \dots, h^p_{g,T}]$, respectively.

After spatial pose–expression interaction, the updated features $\widetilde{H}^p$ and $\widetilde{H}^e$ are routed to three temporal branches:

$$Z^u = \mathcal{T}^u_\psi(\widetilde{H}^p), \quad Z^l = \mathcal{T}^l_\psi(\widetilde{H}^p), \quad Z^e = \mathcal{T}^e_\psi(\widetilde{H}^e). \tag{4}$$

Here, Z^u, Z^l, and Z^e are continuous latents for upper-body motion, lower-body motion, and facial expressions, respectively. The temporal encoder downsamples the sequence by a factor of f, resulting in $T' = T/f$ latent tokens. This decomposition avoids forcing heterogeneous motion components into one latent space: upper-body gestures, lower-body posture, and facial expressions exhibit different temporal dynamics and relationships to speech. The shared spatial encoder and interaction modules retain holistic structure, while the temporal branches preserve part-specific dynamics.

The decoder reconstructs upper- and lower-body pose features, merges them into a full-body representation, and outputs the original SMPL-X pose and expression:

$$\hat{R}, \hat{E} = \mathcal{D}_\psi(Z^u, Z^l, Z^e). \tag{5}$$

Although the three-stream representation preserves part-specific dynamics, independently encoding these streams would weaken holistic coordination. All interaction modules use FiLM-style modulation [Perez et al. 2018] and gated residual updates. Given a target feature x and a temporally aligned conditioning feature c, we define

$$\begin{aligned} h &= \phi(c), \\ (\gamma, \beta, g) &= \big(W_\gamma h,\ W_\beta h,\ \sigma(W_g h)\big), \\ \mathcal{F}(x, c) &= x + \alpha\, g \odot \big(\gamma \odot x + \beta\big), \\ \mathcal{R}(x, c) &= x + \alpha\, g \odot W_\Delta h. \end{aligned} \tag{6}$$

where $\mathcal{F}$ and $\mathcal{R}$ denote the FiLM-gated and residual-gated updates, $\phi(\cdot)$ is a lightweight projection, and α is a learnable scalar initialized to a small value. For node-wise pose features, the modulation parameters are broadcast across the joint dimension.

At the spatial stage, expression features modulate pose-node features through $\mathcal{F}$; the updated pose nodes are pooled and softly blended with the global pose feature, which then updates the expression stream through $\mathcal{R}$. At the bottleneck, interactions are applied bidirectionally between upper–lower and upper–expression streams, while direct lower–expression interaction is omitted. After temporal decoding, a final pose–expression interaction is applied before the output decoders. This design coordinates the streams without collapsing them into a single latent representation.

The autoencoder is optimized with reconstruction and temporal consistency losses:

$$\mathcal{L}_{\text{AE}} = \mathcal{L}_{\text{rec}} + \lambda_{\text{vel}}\mathcal{L}_{\text{vel}} + \lambda_{\text{acc}}\mathcal{L}_{\text{acc}}, \tag{7}$$

where $\mathcal{L}_{\text{rec}}$ supervises pose and expression reconstruction, and $\mathcal{L}_{\text{vel}}$ and $\mathcal{L}_{\text{acc}}$ enforce first- and second-order temporal consistency. After training, the autoencoder is frozen and used to provide structured target latents and the decoder for speech-conditioned flow generation.

3.4 Stage II: Speech-Conditioned Latent Flow Matching

Stage II generates motion by learning a speech-conditioned transport process in the structured latent space. Compared with direct

motion-space generation, latent flow matching reduces the dimensionality and improves stability. Compared with codebook-based generation, it preserves a continuous motion representation and avoids quantization artifacts. The learned velocity field is integrated by an ODE solver to transport source latents to realistic motion latents.

Given the pre-extracted HuBERT, Librosa, FastText, and SBERT features, we construct branch-specific speech conditions:

$$C^u, C^l, C^e = C_\omega(A^h, A^r, T^w, T^g), \tag{8}$$

where A^h and A^r denote HuBERT and Librosa audio features, and T^w and T^g denote word-level FastText and sentence-level SBERT features. The condition encoder produces separate conditions because facial expressions depend more strongly on phonetic and semantic content, upper-body gestures respond to prosody and communicative emphasis, and lower-body motion mainly reflects posture and stability. It uses branch-specific audio/text projections, semantic modulation, residual gating, and temporal downsampling to produce conditions aligned with the latent resolution.

Let

$$Z_1 = \{Z_1^u, Z_1^l, Z_1^e\} \tag{9}$$

denote target latents obtained from the frozen autoencoder. We independently sample source latents:

$$Z_0 = \{Z_0^u, Z_0^l, Z_0^e\}, \qquad Z_0^s \sim \mathcal{N}(0, I), \quad s \in \{u, l, e\}. \tag{10}$$

Although ODE integration is deterministic for a fixed Z_0, different Gaussian source latents produce different transport trajectories and generated motions. For $\tau \sim \mathcal{U}(0, 1)$, we construct the interpolated latent and target velocity:

$$Z_\tau = (1 - \tau)Z_0 + \tau Z_1, \qquad U_\tau = Z_1 - Z_0. \tag{11}$$

The flow model uses a three-stream Transformer backbone to predict velocities for the structured latent streams. Each stream performs speech-condition cross-attention with its corresponding branch-specific condition, while the motion streams exchange information through the FiLM-gated interaction defined above. We use the interaction neighborhoods $\mathcal{N}(u) = \{l, e\}$, $\mathcal{N}(l) = \{u\}$, and $\mathcal{N}(e) = \{u\}$. The velocity prediction for stream s is

$$\widehat{U}_\tau^s = V_\theta^s\left(Z_\tau^s, \tau, C^s; \{Z_\tau^r\}_{r \in \mathcal{N}(s)}\right), \quad s \in \{u, l, e\}. \tag{12}$$

Thus, $C = \{C^u, C^l, C^e\}$ contains three branch-specific conditions rather than one shared condition.

The latent flow matching objective is

$$\mathcal{L}_{\text{FM}} = \sum_{s \in \{u,l,e\}} \lambda_s \mathbb{E}_{\tau, Z_0, Z_1}\left[\left\|\widehat{U}_\tau^s - U_\tau^s\right\|_2^2\right]. \tag{13}$$

To improve compatibility between the learned latent flow and the frozen decoder, we integrate the learned field from Z_0 to obtain $\widetilde{Z}_1$, decode it through the frozen Stage-I decoder, and supervise the resulting pose and expression in the original motion space:

$$\mathcal{L}_{\text{gen}} = \mathcal{L}_{\text{FM}} + \lambda_{\text{roll}}\mathcal{L}_{\text{roll}}. \tag{14}$$

This rollout loss constrains the integrated trajectory after decoding and improves motion-space fidelity and temporal stability.

During inference, the condition encoder first produces $\{C^u, C^l, C^e\}$ from the input speech. Source latents are then integrated from $\tau = 0$ to $\tau = 1$ and decoded into SMPL-X motion. We use fixed-step Euler integration with 50 function evaluations. For arbitrary-length generation, 240-frame windows overlap by 8 original motion frames. The final overlap of the previous window is used as known context for the next window; overlapping predictions are linearly blended and smoothed with a Gaussian kernel of size 3 to reduce boundary discontinuities. This stitching is confined to the overlap region and is used only for long-sequence inference.

3.5 Training details.

The structured autoencoder is trained first and then frozen. The flow backbone is first optimized with $\mathcal{L}_{\text{FM}}$ and subsequently fine-tuned with $\mathcal{L}_{\text{gen}}$. We use temporal compression 4×, latent dimensions $(d_z^u, d_z^l, d_z^e) = (512, 128, 512)$, condition dimensions $(d_c^u, d_c^l, d_c^e) = (256, 128, 384)$, a 6-layer/8-head Transformer backbone with hidden dimension 384, and interaction hidden dimension 256. With pre-extracted speech features, Euler sampling with 50 NFEs runs at 0.006 s per frame on one RTX 4070 Ti SUPER. Remaining optimization and implementation details are provided in the supplementary material.

4 Experiments

4.1 Experimental Settings

4.1.1 Dataset. For training and evaluation, we use the BEAT2 dataset [Liu et al. 2024b], which contains 60 hours of high-quality motion data from 25 speakers, including 12 female and 13 male speakers. The dataset consists of 1,762 sequences, with an average duration of 65.66 seconds, where speakers respond to daily inquiries. We divide the dataset into training, validation, and test sets with ratios of 85%, 7.5%, and 7.5%, respectively. To ensure a fair comparison with prior work, we follow EMAGE [Liu et al. 2024b] and use the Speaker-2 subset for training, validation, and testing. This single-speaker protocol provides a controlled evaluation setting, but it also limits the range of speaker-specific gesture styles observed during training.

4.1.2 Metrics. We evaluate generated body gestures using Fréchet Gesture Distance (FGD) [Yoon et al. 2020], which measures the distributional distance between generated and ground-truth gestures. We compute FGD using the feature extractor provided by the official EMAGE codebase. Following EMAGE [Liu et al. 2024b], DIV [Li et al. 2021a] is computed as the average L1 distance between generated gesture clips, and therefore mainly reflects coordinate-level motion variation. To evaluate speech-motion synchrony, we use Beat Consistency Score (BCS) [Li et al. 2021b], which measures the temporal alignment between gesture rhythm and low-level audio beats. For facial motion evaluation, we report vertex mean squared error (MSE) and vertex L1 difference (LVD) [Yi et al. 2023], which measure the discrepancy between generated and ground-truth facial vertices. Since quantitative metrics may not fully capture semantic appropriateness and perceptual naturalness, we further conduct qualitative comparisons and a user perception study.

4.2 Quantitative Results

As shown in Table 1, TalkFlow achieves the best facial MSE and LVD, together with the second-best FGD and BCS. The improvements

Table 1. **Quantitative comparison on the BEAT2 dataset.** Following prior work, methods are grouped into facial expression generation, body gesture generation, and holistic motion generation. FGD is reported in 10^{-1} scale, BCS in 10^{-1} scale, MSE in 10^{-8} scale, and LVD in 10^{-5} scale. Bold indicates the best result, and underline indicates the second-best result.

Methods	FGD↓	BCS↑	DIV↑	MSE↓	LVD↓
Ground Truth	0	6.86	13.05	0	0
Facial Expression Generation					
FaceFormer [Fan et al. 2022]	-	-	-	7.79	7.59
CodeTalker [Xing et al. 2023]	-	-	-	8.03	7.77
Body Gesture Generation					
S2G [Ginosar et al. 2019]	28.15	4.68	5.97	-	-
Trimodal [Yoon et al. 2020]	12.41	5.93	7.72	-	-
HA2G [Liu et al. 2022c]	12.32	6.78	8.63	-	-
DisCo [Liu et al. 2022b]	9.42	6.44	9.91	-	-
CaMN [Liu et al. 2022d]	6.64	6.77	10.86	-	-
DiffStyleGesture [Yang et al. 2023]	8.81	7.24	11.49	-	-
Holistic Motion Generation					
Habibie et al. [Habibie et al. 2021]	9.04	7.72	8.21	8.61	8.04
TalkSHOW [Yi et al. 2023]	6.21	6.95	**13.47**	7.79	7.77
EMAGE [Liu et al. 2024b]	5.51	7.72	13.06	7.68	7.56
DiffSHEG [Chen et al. 2024]	8.99	7.14	11.91	7.67	8.67
ProbTalk [Liu et al. 2024a]	6.17	8.10	10.43	8.99	8.39
MambaTalk [Xu et al. 2024a]	5.37	7.81	13.05	<u>6.29</u>	<u>6.90</u>
EchoMask [Zhang et al. 2025b]	**4.62**	7.74	<u>13.37</u>	6.76	7.29
TalkFlow (Ours)	<u>5.24</u>	<u>7.84</u>	11.48	**5.77**	**6.65**

in MSE and LVD indicate that the structured latent representation preserves fine-grained facial dynamics. ProbTalk obtains a slightly higher BCS, whereas TalkFlow performs better on FGD, DIV, MSE, and LVD. These results also indicate that no single metric provides a complete measure of holistic generation quality. Combo [Xu et al. 2024b] is not included because its results use a different four-subject train/test protocol, and its code and pretrained models are unavailable for retraining under the common Speaker-2 setting.

For body gesture evaluation, TalkFlow achieves the second-best FGD. FGD evaluates distributional similarity in a learned gesture feature space, but does not directly measure perceptual realism, semantic appropriateness, or face–body coordination. EchoMask obtains the lowest FGD, whereas TalkFlow obtains the lowest mean realism rank in the user study. This difference suggests that feature-distribution distance and perceived motion quality capture complementary aspects of generation performance.

BCS measures low-level coincidence between audio and motion beats. Notably, the ground-truth BCS of 6.86 is lower than that of every holistic generation method in Table 1. Therefore, a higher BCS does not necessarily imply more natural or semantically appropriate motion and may favor frequent or regular beat responses. In contrast, the user study evaluates perceived speech-motion synchronization at a higher perceptual level, where TalkFlow obtains the lowest mean rank among the evaluated methods. BCS should thus be interpreted as a low-level rhythmic alignment metric rather than a standalone measure of overall synchronization quality.

TalkFlow obtains a lower quantitative DIV than several baselines. Since DIV is computed from L1 distances between generated gesture clips, it mainly reflects coordinate-level motion variation rather than high-level diversity of gesture patterns, expressive modes, or talking styles. This distinction is reflected in the user study: MambaTalk obtains the lowest mean diversity rank, while TalkFlow ranks second, despite their lower quantitative DIV than TalkSHOW and EchoMask. Flow matching is not inherently deterministic, because different Gaussian source latents produce different generation trajectories. However, the current model uses a simple white-noise prior without explicit talking-style conditioning, and the Speaker-2 protocol contains limited style variation. Style-conditioned priors and multi-speaker training may therefore further improve controllable diversity.

4.3 Ablation Study

Table 2 validates the contribution of the main components in TalkFlow. The single-latent variant compresses all motion components into one shared latent space. Separating pose and expression in the dual-latent variant improves facial MSE/LVD from 7.20/7.76 to 6.66/7.16, while also improving FGD, BCS, and DIV. The triple-latent variants further decompose full-body pose into upper- and lower-body streams, producing a substantial reduction in FGD and confirming the benefit of preserving heterogeneous part-specific dynamics.

Introducing autoencoder-stage interaction further improves both body and facial metrics, indicating that coordination during representation learning benefits holistic motion generation. The full model achieves the best results across all ablation metrics, showing that graph-based pose encoding, autoencoder and flow-stage interactions, and ODE rollout reconstruction are complementary. In

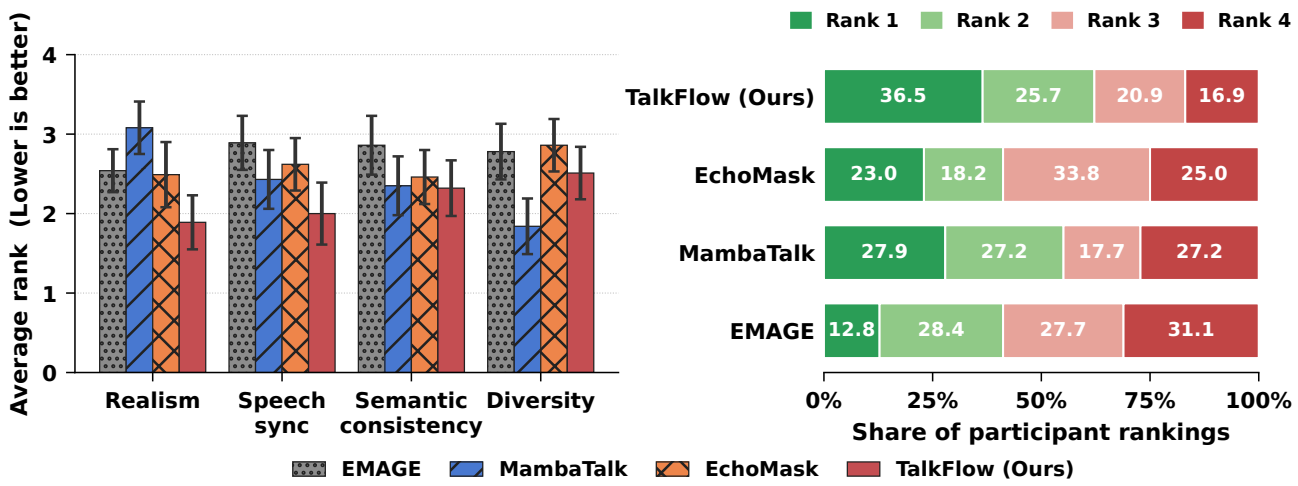


Fig. 3. **Perceptual user study.** N=37 participants ranked TalkFlow against three baselines on four explicitly labeled criteria. **Left:** mean rank per criterion, with error bars denoting one standard deviation across participants; lower is better. **Right:** percentage shares of Rank-1 to Rank-4 votes for each method. TalkFlow obtains the lowest mean rank in realism, semantic consistency, and speech-motion synchronization, and the largest overall share of first-place votes.

Table 2. **Ablation study on the BEAT2 dataset.** "Single latent", "Dual latent", and "Triple latent" denote the number of latent streams. "Int." denotes cross-stream interaction, and "Rollout Rec." denotes training with ODE rollout reconstruction loss.

Variant	FGD↓	BCS↑	DIV↑	MSE↓	LVD↓
Single latent	13.21	7.45	8.67	7.20	7.76
Dual latent	11.48	7.72	9.07	6.66	7.16
Triple latent					
w/o Int.	6.40	7.80	10.94	6.11	6.79
w/o Flow Int.	5.30	7.73	11.20	6.09	6.75
w/o AE Int.	5.40	7.72	11.01	6.30	7.09
w/o GCN	5.49	7.76	10.82	6.16	6.94
w/o Rollout Rec.	9.46	7.75	10.84	6.44	6.97
Full	**5.24**	**7.84**	**11.48**	**5.77**	**6.65**

particular, removing rollout reconstruction substantially degrades FGD, confirming that motion-space rollout supervision improves compatibility between the learned latent trajectory and the frozen decoder.

4.4 User Perception Study

Following the user study protocol of EchoMask [Zhang et al. 2025b], we conduct a perceptual evaluation to compare the visual quality and speech-motion consistency of different methods. We recruit 37 participants with normal or corrected-to-normal vision, normal hearing, and sufficient English proficiency to understand the evaluated speech clips. Participants are asked to rank generated videos from different methods according to four perceptual criteria: realism, semantic consistency, speech-motion synchronization, and motion diversity. We report the average rank for each method, where a lower rank indicates stronger user preference.

As shown in Fig. 3, TalkFlow obtains the lowest mean rank in realism, semantic consistency, and speech-motion synchronization. These perceptual results complement rather than replace the quantitative metrics. Although EchoMask obtains a lower FGD in Table 1, TalkFlow receives a lower mean realism rank, illustrating that FGD does not fully capture motion naturalness, temporal stability, and holistic coordination.

For speech-motion synchronization, TalkFlow obtains the lowest mean user-study rank among the evaluated methods, while its quantitative BCS is second-best. This difference is consistent with the distinct aspects captured by the two evaluations: BCS measures low-level audio-beat coincidence, whereas participants evaluate whether generated gestures and facial motion appear appropriately synchronized with the speech as a whole. The comparatively low ground-truth BCS further shows that maximizing beat coincidence is not equivalent to maximizing perceived synchronization.

For motion diversity, MambaTalk obtains the lowest mean rank and TalkFlow ranks second. This differs from quantitative DIV, where both methods score below TalkSHOW and EchoMask. DIV reflects coordinate-level variation and therefore does not necessarily capture high-level variation in gesture patterns, expressive modes, or talking styles. The user-study results suggest that TalkFlow retains perceptually meaningful motion variation despite its lower coordinate-level DIV.

Participant comments further support these observations. EMAGE often produces body movements that are too calm and do not fully match the agitated tone of the speech; however, its conservative motion also avoids severe torso deformation. MambaTalk exhibits the largest range of body movement and often matches the angry tone well, but participants also observe occasional unnatural jitters and sudden full-body twists. EchoMask sometimes produces repetitive arm postures, and its wrists may appear unnatural. In contrast, TalkFlow is generally perceived as the most balanced method, with

smoother pose transitions, more stable body motion, and better synchronization with the speech tempo.

4.5 Qualitative Results

We provide qualitative comparisons with representative holistic co-speech motion generation methods in the final two figure-only pages to allow larger rendered frames and clearer visual inspection. We first compare facial motion at selected spoken words in Fig. 4. Since BEAT2 facial motion is generally close to neutral and SMPL-X provides a relatively sparse facial representation, differences in static renderings are mainly visible through lip opening/closing and mouth-motion timing. TalkFlow generally produces mouth shapes closest to the ground truth, while several baseline results contain clear lip-shape errors, such as an open mouth when the ground-truth mouth is closed. Red boxes mark only these evident errors; unboxed results are not assigned a binary correct/incorrect label.

We further compare body motion on 14 word-level cases in Fig. 5 and 6. Each case corresponds to the body pose generated at a specific spoken word, allowing direct comparison of gesture timing, body posture, and face–body coordination under the same speech context. Orange boxes highlight representative ground-truth motion regions used as visual references, green boxes indicate generated motions closest to the corresponding references, and red boxes mark visible artifacts or semantically inconsistent motions. In most selected cases, TalkFlow remains close to the ground truth while avoiding obvious artifacts such as implausible body deformation, abrupt large movements, or unstable poses.

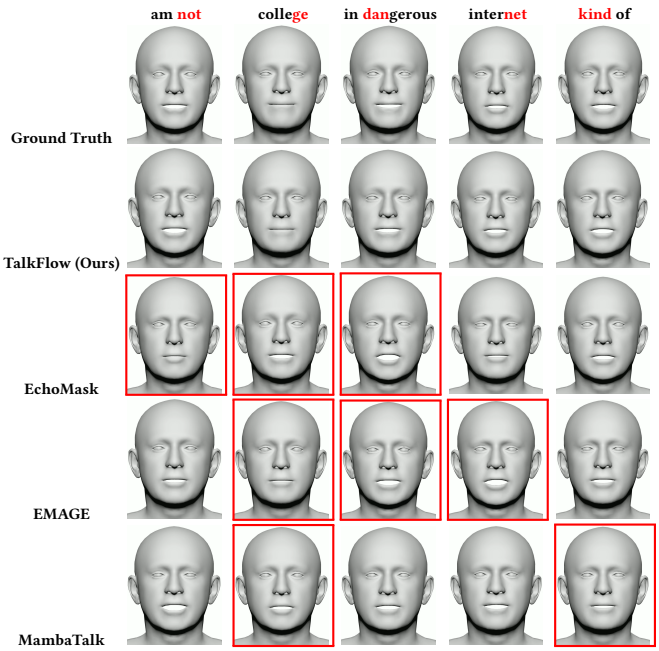


Fig. 4. Facial qualitative comparison on selected spoken words. Each column corresponds to a spoken word, with the highlighted syllable shown in red. Red boxes indicate evident lip-shape errors relative to the ground truth, such as an open mouth versus a closed mouth. TalkFlow generally remains closest to the ground-truth mouth shape across the selected comparisons.

5 Conclusion

We presented **TalkFlow**, a structured latent flow framework for speech-driven holistic motion generation. TalkFlow learns a continuous three-stream latent representation for upper-body motion, lower-body motion, and facial expressions, avoiding both direct high-dimensional motion generation and discrete tokenization. By encoding the full-body pose graph before temporal stream separation and coordinating the streams through lightweight FiLM-gated interactions, TalkFlow preserves part-specific dynamics while maintaining holistic face–body consistency.

We further introduced speech-conditioned latent flow matching to generate coordinated motion latents from acoustic and textual speech features. Experiments on BEAT2 show that TalkFlow achieves the best facial fidelity and competitive body-motion performance. Qualitative comparisons demonstrate plausible word-level lip dynamics and natural body gestures with fewer visible artifacts. The user study further shows that TalkFlow obtains the lowest mean rank in realism, semantic consistency, and speech-motion synchronization, together with the largest overall share of first-place rankings.

Several limitations remain. The single-speaker Speaker-2 protocol limits evidence for cross-speaker and cross-style generalization. The current Gaussian prior also lacks explicit talking-style control, limiting controllable diversity. Moreover, global translation is not modeled, making foot–ground contact difficult to constrain and contributing to occasional foot sliding. Future work will explore multi-speaker training, style-conditioned latent priors, and joint global-translation/contact constraints for more diverse and physically plausible motion.

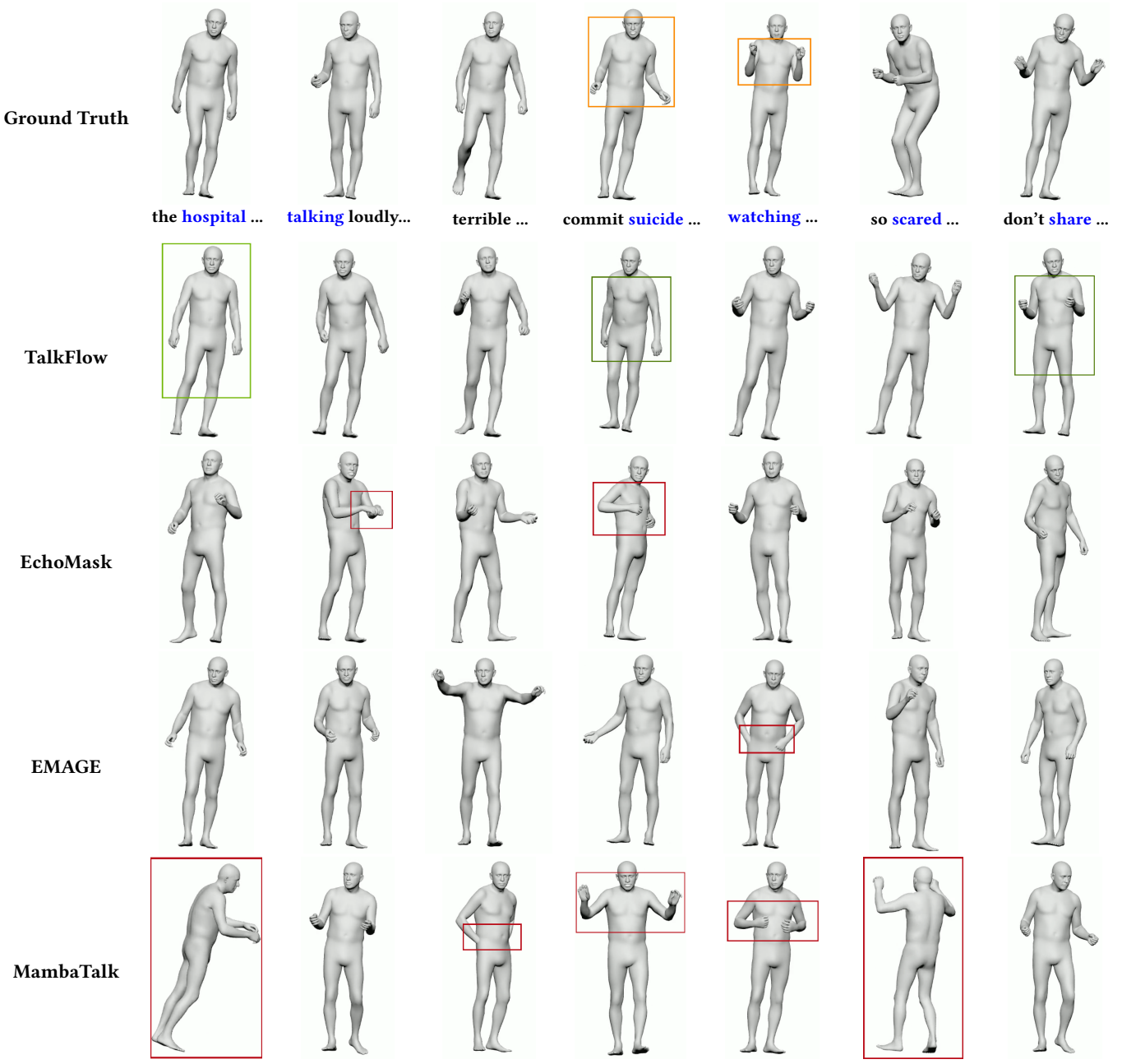


Fig. 5. **Qualitative comparison on representative word-level speech contents.** The first row shows ground-truth motions and the corresponding speech contents, followed by results from different methods. Orange boxes highlight representative ground-truth motion regions, green boxes indicate generated motions closest to the corresponding references, and red boxes mark visible artifacts or semantically inconsistent motions. For EchoMask on "talking loudly," the two hands intersect. For MambaTalk on "terrible" and EMAGE on "watching," the fists intersect the torso; MambaTalk also exhibits abrupt whole-body rotation and imbalance.

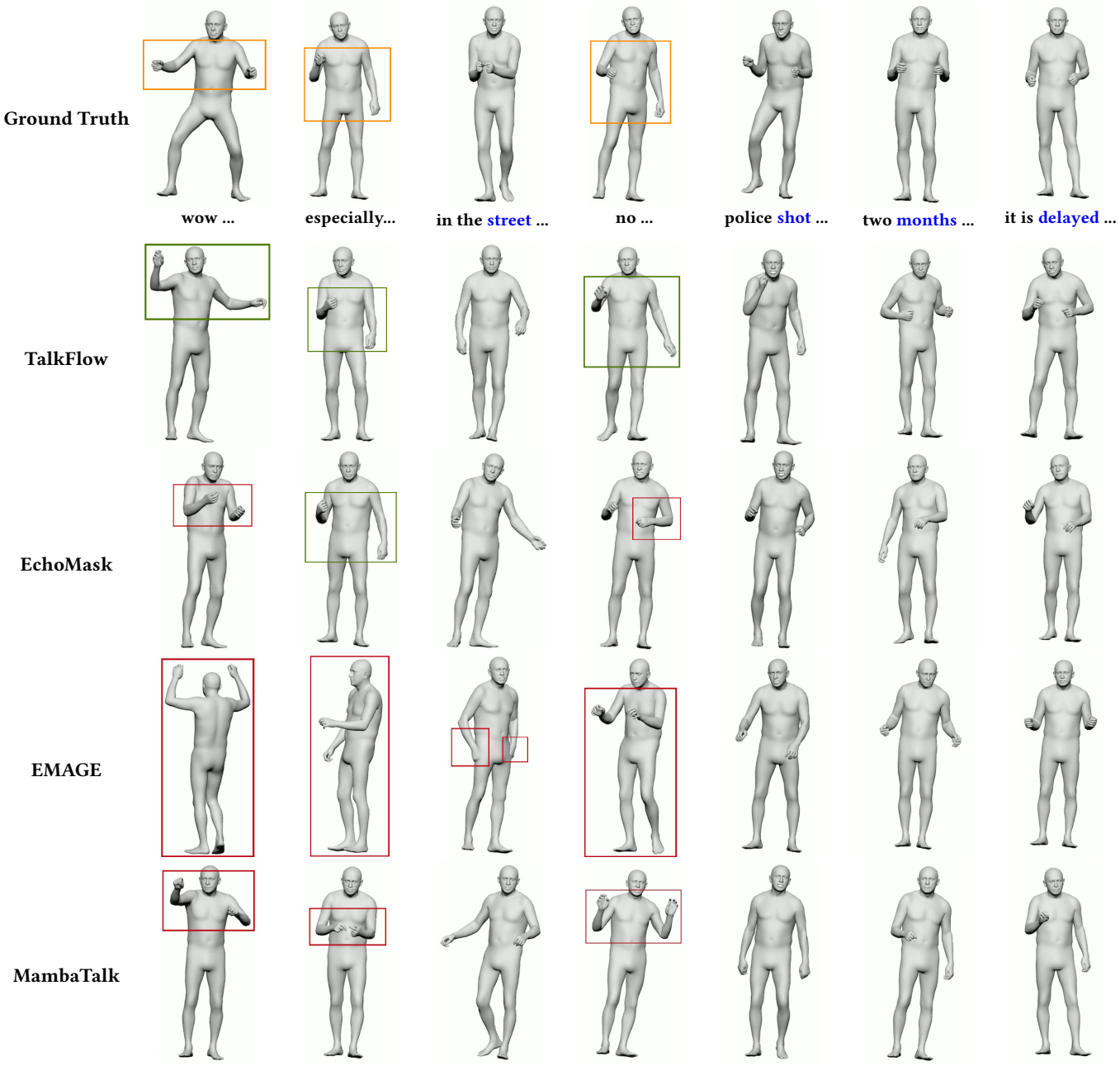


Fig. 6. **Qualitative comparison on representative word-level speech contents.** The first row shows ground-truth motions and the corresponding speech contents, followed by results from different methods. Orange boxes highlight representative ground-truth motion regions, green boxes indicate generated motions closest to the corresponding references, and red boxes mark visible artifacts or semantically inconsistent motions. In the selected examples, EMAGE exhibits abrupt whole-body rotation on "wow" and "especially," together with visible mesh deformation on "in the street." These cases illustrate that larger-amplitude motions may still be accompanied by mesh deformation or temporal jitter.

References

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

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics* 5 (2017), 135–146.

Justine Cassell, Hannes Högni Vilhjálmsson, and Timothy Bickmore. 2001. Beat: the behavior expression animation toolkit. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*. 477–486.

Bohong Chen, Yumeng Li, Yinglin Xu, Youyi Zheng, Yanlin Weng, and Kun Zhou. 2026. EchoAvatar: Real-time Generative Avatar Animation from Audio Streams. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers (SIGGRAPH Conference Papers '26)*. Association for Computing Machinery, 29:1–29:12. doi:10.1145/3799902.3811066

Junming Chen, Yunfei Liu, Jianan Wang, Ailing Zeng, Yu Li, and Qifeng Chen. 2024. 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*. 7352–7361.

Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. 2020. Pose2mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose. In *European Conference on Computer Vision*. Springer, 769–787.

Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael J Black. 2019. Capture, learning, and synthesis of 3D speaking styles. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10101–10111.

Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. 2016. Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering. In *Advances in Neural Information Processing Systems*, Vol. 29.

Yingruo Fan, Zhaojiang Lin, Jun Saito, Wenping Wang, and Taku Komura. 2022. Faceformer: Speech-driven 3d facial animation with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 18770–18780.

Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, and Jitendra Malik. 2019. Learning individual styles of conversational gesture. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 3497–3506.

Ziyan Guo, Zeyu Hu, De Wen Soh, and Na Zhao. 2025. Motionlab: Unified human motion generation and editing via the motion-condition-motion paradigm. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 13869–13879.

Ikhsanul Habibie, Weipeng Xu, Dushyant Mehta, Lingjie Liu, Hans-Peter Seidel, Gerard Pons-Moll, Mohamed Elgharib, and Christian Theobalt. 2021. Learning speech-driven 3d conversational gestures from video. In *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents*. 101–108.

Gustav Eje Henter, Simon Alexanderson, and Jonas Beskow. 2020. Moglow: Probabilistic and controllable motion synthesis using normalising flows. *ACM Transactions on Graphics (TOG)* 39, 6 (2020), 1–14.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems* 33 (2020), 6840–6851.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing* 29 (2021), 3451–3460.

Vincent Tao Hu, Wenzhe Yin, Pingchuan Ma, Yunlu Chen, Basura Fernando, Yuki M Asano, Efstratios Gavves, Pascal Mettes, Bjorn Ommer, and Cees GM Snoek. 2023. Motion flow matching for human motion synthesis and editing. *arXiv preprint arXiv:2312.08895* (2023).

Herve Jegou, Matthijs Douze, and Cordelia Schmid. 2010. Product quantization for nearest neighbor search. *IEEE transactions on pattern analysis and machine intelligence* 33, 1 (2010), 117–128.

Taekyung Ki, Dongchan Min, and Gyeongsu Chae. 2025. Float: Generative motion latent flow matching for audio-driven talking portrait. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 14699–14710.

Stefan Kopp and Ipke Wachsmuth. 2004. Synthesizing multimodal utterances for conversational agents. *Computer animation and virtual worlds* 15, 1 (2004), 39–52.

Sergey Levine, Philipp Krähenbühl, Sebastian Thrun, and Vladlen Koltun. 2010. Gesture controllers. In *Acm siggraph 2010 papers*. 1–11.

Jing Li, Di Kang, Wenjie Pei, Xuefei Zhe, Ying Zhang, Zhenyu He, and Linchao Bao. 2021a. Audio2gestures: Generating diverse gestures from speech audio with conditional variational autoencoders. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11293–11302.

Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. 2021b. Ai choreographer: Music conditioned 3d dance generation with aist++. In *Proceedings of the IEEE/CVF international conference on computer vision*. 13401–13412.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. 2022. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747* (2022).

Haiyang Liu, Naoya Iwamoto, Zihao Zhu, Zhengqing Li, You Zhou, Elif Bozkurt, and Bo Zheng. 2022b. Disco: Disentangled implicit content and rhythm learning for diverse co-speech gestures synthesis. In *Proceedings of the 30th ACM international conference on multimedia*. 3764–3773.

Haiyang Liu, Zihao Zhu, Giorgio Becherini, Yichen Peng, Mingyang Su, You Zhou, Xuefei Zhe, Naoya Iwamoto, Bo Zheng, and Michael J Black. 2024b. 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*. 1144–1154.

Haiyang Liu, Zihao Zhu, Naoya Iwamoto, Yichen Peng, Zhengqing Li, You Zhou, Elif Bozkurt, and Bo Zheng. 2022d. Beat: A large-scale semantic and emotional multimodal dataset for conversational gestures synthesis. In *European conference on computer vision*. Springer, 612–630.

Xingchao Liu, Chengyue Gong, and Qiang Liu. 2022a. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003* (2022).

Xian Liu, Qianyi Wu, Hang Zhou, Yinghao Xu, Rui Qian, Xinyi Lin, Xiaowei Zhou, Wayne Wu, Bo Dai, and Bolei Zhou. 2022c. Learning hierarchical cross-modal association for co-speech gesture generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10462–10472.

Yifei Liu, Qiong Cao, Yandong Wen, Huaiguang Jiang, and Changxing Ding. 2024a. Towards variable and coordinated holistic co-speech motion generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1566–1576.

Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and music signal analysis in python. *SciPy* 2015 (2015), 18–24.

Evonne Ng, Hanbyul Joo, Liwen Hu, Hao Li, Trevor Darrell, Angjoo Kanazawa, and Shiry Ginosar. 2022. Learning to listen: Modeling non-deterministic dyadic facial motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 20395–20405.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. 2019. 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*. 10975–10985.

Ziqiao Peng, Yihao Luo, Yue Shi, Hao Xu, Xiangyu Zhu, Hongyan Liu, Jun He, and Zhaoxin Fan. 2023. Selftalk: A self-supervised commutative training diagram to comprehend 3d talking faces. In *Proceedings of the 31st ACM International Conference on Multimedia*. 5292–5301.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. 2018. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

Isabella Poggi, Catherine Pelachaud, Fiorella de Rosis, Valeria Carofiglio, and Berardina De Carolis. 2005. Greta. a believable embodied conversational agent. In *Multimodal intelligent information presentation*. Springer, 3–25.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. 3982–3992.

Shuai Shen, Wenliang Zhao, Zibin Meng, Wanhua Li, Zheng Zhu, Jie Zhou, and Jiwen Lu. 2023. Difftalk: Crafting diffusion models for generalized audio-driven portraits animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1982–1991.

Aaron Van Den Oord, Oriol Vinyals, et al. 2017. Neural discrete representation learning. *Advances in neural information processing systems* 30 (2017).

Xinmu Wang, Fumio Honda, Xifeng Gao, Jin Huang, Wei Chen, and Xianfeng David Gu. 2026. HoloDiff: Holistic Talking Human Animation Via Latent Diffusion. *Computer Animation and Virtual Worlds* 37, 4 (2026), e70173. doi:10.1002/cav.70173

Hanwei Wu and Markus Flierl. 2019. Learning product codebooks using vector-quantized autoencoders for image retrieval. In *2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE, 1–5.

Jinbo Xing, Menghan Xia, Yuechen Zhang, Xiaodong Cun, Jue Wang, and Tien-Tsin Wong. 2023. Codetalker: Speech-driven 3d facial animation with discrete motion prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12780–12790.

Chao Xu, Mingze Sun, Zhi-Qi Cheng, Fei Wang, Yang Liu, Baigui Sun, Ruqi Huang, and Alexander Hauptmann. 2024b. Combo: Co-speech holistic 3D human motion generation and efficient customizable adaptation in harmony. *arXiv preprint arXiv:2408.09397* (2024).

Zunnan Xu, Yukang Lin, Haonan Han, Sicheng Yang, Ronghui Li, Yachao Zhang, and Xiu Li. 2024a. Mambatalk: Efficient holistic gesture synthesis with selective state space models. *Advances in Neural Information Processing Systems* 37 (2024), 20055–20080.

Kaixing Yang, Xulong Tang, Ziqiao Peng, Xiangyue Zhang, Puwei Wang, Jun He, and Hongyan Liu. 2025. FlowerDance: MeanFlow for Efficient and Refined 3D Dance Generation. *arXiv preprint arXiv:2511.21029* (2025).

Sicheng Yang, Zhiyong Wu, Minglei Li, Zhensong Zhang, Lei Hao, Weihong Bao, Ming Cheng, and Long Xiao. 2023. Diffusestylegesture: Stylized audio-driven co-speech gesture generation with diffusion models. *arXiv preprint arXiv:2305.04919* (2023).

Hongwei Yi, Hualin Liang, Yifei Liu, Qiong Cao, Yandong Wen, Timo Bolkart, Dacheng Tao, and Michael J Black. 2023. Generating holistic 3d human motion from speech. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 469–480.

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

Xiangyue Zhang, Jianfang Li, Jiaxu Zhang, Ziqiang Dang, Jianqiang Ren, Liefeng Bo, and Zhigang Tu. 2025a. SemTalk: Holistic Co-Speech Motion Generation with Frame-level Semantic Emphasis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 13761–13771.

Xiangyue Zhang, Jianfang Li, Jiaxu Zhang, Jianqiang Ren, Liefeng Bo, and Zhigang Tu. 2025b. Echomask: Speech-queried attention-based mask modeling for holistic co-speech motion generation. In *Proceedings of the 33rd ACM International Conference on Multimedia*. 10827–10836.

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

Wentao Zhu, Xiaoxuan Ma, Dongwoo Ro, Hai Ci, Jinlu Zhang, Jiaxin Shi, Feng Gao, Qi Tian, and Yizhou Wang. 2023. Human motion generation: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46, 4 (2023), 2430–2449.