Mitigating Error Accumulation in Co-Speech Motion Generation via Global Rotation Diffusion and Multi-Level Constraints

Xiangyue Zhang[1] *, **Jianfang Li** *[1†], **Jianqiang Ren** [1] , **Jiaxu Zhang** [2]

[1] Tongyi Lab, Alibaba Group [2] Nanyang Technological University

Abstract

Reliable co-speech motion generation requires precise motion representation and consistent structural priors across all joints. Existing generative methods typically operate on local joint rotations, which are defined hierarchically based on the skeleton structure. This leads to cumulative errors during generation, manifesting as unstable and implausible motions at end-effectors. In this work, we propose GlobalDiff, a diffusion-based framework that operates directly in the space of global joint rotations for the first time, fundamentally decoupling each joint's prediction from upstream dependencies and alleviating hierarchical error accumulation. To compensate for the absence of structural priors in global rotation space, we introduce a multi-level constraint scheme. Specifically, a joint structure constraint introduces virtual anchor points around each joint to better capture fine-grained orientation. A skeleton structure constraint enforces angular consistency across bones to maintain structural integrity. A temporal structure constraint utilizes a multi-scale variational encoder to align the generated motion with ground-truth temporal patterns. These constraints jointly regularize the global diffusion process and reinforce structural awareness. Extensive evaluations on standard co-speech benchmarks show that GlobalDiff generates smooth and accurate motions, improving the performance by 46.0% compared to the current SOTA under multiple speaker identities. Project page: https://xiangyue-zhang.github.io/GlobalDiff

Introduction

Holistic co-speech motion generation, which synchronizes body posture (Chhatre et al. 2024; Zhang et al. 2024a), hand gestures (Li et al. 2021a; Liu et al. 2022b), and facial expressions (Li et al. 2025; Peng et al. 2023) with speech, is an increasing focus in the field of computer vision. It plays a vital role in enabling virtual characters to communicate naturally and convincingly by modeling the full spectrum of non-verbal cues. This capability facilitates their deployment in avatars, interactive games, live-streaming platforms, and human-robot collaboration.

Recent advancements in generative modeling, particularly diffusion-based approaches (Zhu et al. 2023; Alexanderson et al. 2023; Yang et al. 2023; He et al. 2024; Chen et al. 2024a), have notably improved the expressiveness and naturalness of generated gestures. Typically, these methods operate within a hierarchical local joint rotation space, wherein each joint's orientation is defined relative to its parent joint in a kinematic chain (e.g., SMPL-X (Pavlakos et al. 2019)). While intuitive and structurally consistent, this local formulation inherently suffers from cumulative errors due to its recursive propagation mechanism. *Minor inaccuracies in root or intermediate joint predictions can cause significant errors in the end-effectors, as shown in Figure 1*, thereby substantially degrading the quality and stability of distal joints such as the hands, fingers, and feet—especially during prolonged and expressive motion sequences.

To address this limitation, we explore a new design direction: performing diffusion-based motion generation directly in the global joint rotation space, which we term GlobalDiff. By removing recursive dependencies among joints, global rotations fundamentally mitigate the problem of hierarchical error propagation, leading to more stable and coherent motion, even in long, fine-grained co-speech gestures.

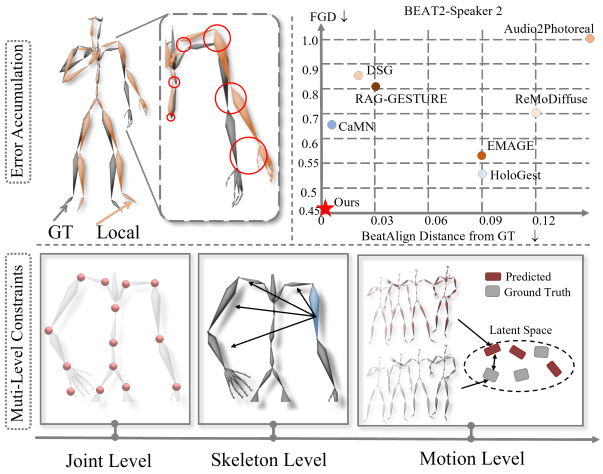

Figure 1: **Overview of our motivation and solution.** Local rotation diffusion leads to error accumulation in distal joints. Global rotation diffusion avoids this but lacks structural priors (top left). We address this with constraints at the joint, skeleton, and motion levels to ensure coherent and reasonable motion (bottom).

*These authors contributed equally.

†Corresponding author.

However, this shift to global representation introduces a new challenge: the loss of natural structural constraints. Unlike local rotations, which implicitly preserve joint relationships through the hierarchical skeleton structure, global rotations treat each joint independently. Without additional guidance, this independence may lead to physically implausible poses or broken kinematic chains. Thus, while global rotation solves error accumulation, it also weakens structural consistency—a tradeoff that must be addressed to fully realize its benefits.

Motivated by these observations, we introduce a multi-level structure constraint designed to reinstate the lost structural coherence progressively. Inspired by human kinematic constraints, we incorporate constraints at three interconnected levels: joint-level, skeleton-level, and motion-level.

At the joint level, we propose to use a set of virtual anchor points attached to each joint to represent the joint-level structure. By aligning the transformed positions of these anchors—rotated by the predicted global rotation—with those derived from GT rotations, we resolve rotation ambiguity and achieve fine-grained supervision of joint orientation.

At the skeleton level, we propose to use an Angular Matrix (AM) computed from pairwise angles between all bone directions to represent the skeleton-level structure. By aligning the predicted AM with that from the GT motion, we constrain global bone relationships and enforce anatomically coherent and physically plausible skeleton configurations.

At the motion level, we propose to use temporal embeddings extracted via a multi-scale variational encoder (Kingma, Welling et al. 2013; Ma et al. 2025) to represent the motion-level structure. By aligning the temporal features of the generated and ground-truth sequences, we ensure dynamic consistency, rhythm synchronization, and smooth temporal transitions throughout the motion.

These multi-level constraints and global rotation representation collectively guide the model to generate expressive, physically plausible motion from speech.

We validate GlobalDiff on several co-speech motion benchmarks. Experimental results show that it produces high-quality motion with improved structural consistency and expressiveness, outperforming existing approaches.

Our contributions are summarized as follows:

- We propose GlobalDiff, a diffusion-based co-speech motion generation framework that uses global joint rotations to resolve hierarchical error accumulation.
- We design a multi-level structure constraint mechanism, consisting of joint-level virtual anchor points, skeleton-level bone consistency, and motion-level temporal coherence, progressively restoring structural plausibility.
- We demonstrate that GlobalDiff achieves superior motion stability, anatomical correctness, and expressiveness compared to prior state-of-the-art methods on standard benchmarks.

Related Work

Holistic Co-speech Motion Generation. Co-speech motion generation aims to produce speech-aligned body movements. Early methods used GANs and diffusion models to improve realism and diversity (Habibie et al. 2021; Ahuja, Lee, and Morency 2022; Ahuja et al. 2023; Zhu et al. 2023; Yang et al. 2023; Zhi et al. 2023). Later work introduced semantic control through hierarchical designs or prompts (Qi et al. 2024; Liu et al. 2022b; Liang et al. 2022; Chen et al. 2024a). Recent approaches target holistic motion—jointly modeling face, hands, and torso—and fall into two main groups: VQ-VAE-based and diffusion-based.

Among VQ-VAE-based methods, TalkSHOW (Yi et al. 2023) handles face separately, while EMAGE (Liu et al. 2024a) masks latent tokens and splits the body into four parts. ProbTalk (Liu et al. 2024b) leverages PQ-VAE with decoding for rhythmic precision. SemTalk (Zhang et al. 2025a) and EchoMask (Zhang et al. 2025b) adopt RVQ-VAE (Guo et al. 2024; Lee et al. 2022) with frame-level semantic focus. However, VQ-VAE-based methods suffer from the limited generation diversity. In this work, we use the diffusion model as our baseline to generate diverse motions.

Diffusion-based Motion Generation. Diffusion models have become a dominant framework in human motion generation, particularly in text-to-motion and co-speech motion tasks, due to their strong generative capacity. For co-speech motion, DiffuseStyleGesture (Yang et al. 2023) pioneered expressive audio-conditioned gesture synthesis, incorporating rhythm alignment and style control. DiffSHEG (Chen et al. 2024b) extended these approaches by jointly modeling facial and body gestures efficiently. GestureDiffuCLIP (Ao, Zhang, and Liu 2023) further introduced multimodal style control using CLIP embeddings. HoloGest (Cheng and Huang 2025) leveraged motion priors and decoupled body parts during diffusion. LivelySpeaker (Zhi et al. 2023) proposed a two-stage method, generating semantic-aware gestures from text before rhythm refinement via audio-driven diffusion. RAG-Gesture (Mughal et al. 2025) introduced retrieval-augmented generation for explicit semantic alignment. However, previous methods overlook the error accumulation problem caused by local rotations and the forward kinematics process, resulting in significant inconsistencies in the motion of end-effectors. In this work,we predict global joint rotations directly and introduce multi-level constraints—at the joint, skeleton, and motion levels—to ensure structural consistency and expressive realism.

Method

Preliminary

Local and Global Joint Rotations. Based on the articulated skeletons, previous methods represent each joint's motion in its local coordinate frame. The local rotation $R_k^{\text{local}} \in \text{SO}(3)$ defines the orientation of joint k relative to its parent in the skeletal hierarchy. In contrast, the global rotation $R_k^{\text{global}} \in \text{SO}(3)$ expresses its absolute orientation in world coordinates. Typically, global rotations are recursively recovered by composing local rotations along the parent-child chain of the skeletal hierarchy:

$$R_k^{\text{global}} = R_{k-1}^{\text{global}} R_k^{\text{local}}, \quad R_1^{\text{global}} = R_1^{\text{local}}, \tag{1}$$

and the global position is:

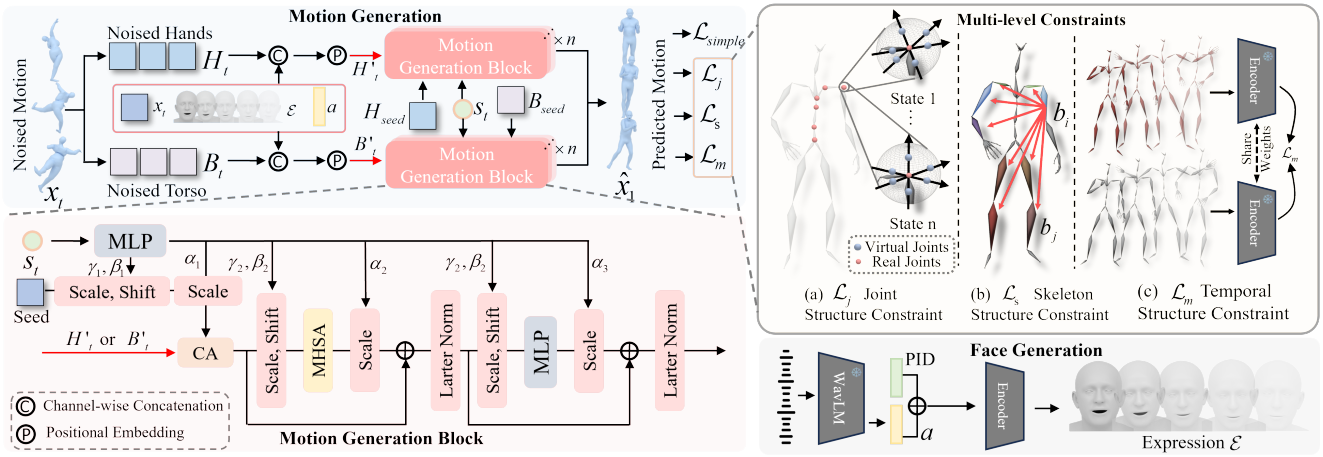

Figure 2: **Overview of the GlobalDiff Framework.** Our model generates consistent and expressive co-speech motion using the global rotation diffusion augmented with multi-level structural constraints. The diffusion model is conditioned on seed pose and prosodic features and predicts body motion through stacked motion generation blocks. To enforce structural plausibility, we introduce: (a) a Joint structure constraint using virtual anchor points to disambiguate orientations; (b) a Skeleton structure constraint that enforces angular consistency across adjacent bones by aligning the angular matrices; and (c) a Temporal structure constraint based on a shared multi-scale VAE encoder to preserve temporal dynamics. Facial expressions are generated in parallel from prosody using a transformer encoder.

$$q_k = R_{k-1}^{\text{global}}(t_k - t_{k-1}) + q_{k-1}, \tag{2}$$

where t_k is the position of joint k at rest state (T-pose). This process, known as Forward Kinematics (FK), has been widely studied and applied in (Kucuk and Bingul 2006; Aberman et al. 2020; Zhang et al. 2024b). By unrolling this recursion, we can express R_k^{global} and q_k as cumulative products and sums:

$$R_k^{\text{global}} = R_1^{\text{local}} R_2^{\text{local}} \cdots R_k^{\text{local}}, \tag{3}$$

$$\begin{aligned} q_k = &\left(\prod_{i=1}^{k-1} R_i^{\text{local}}\right)(t_k - t_{k-1}) \\ &+ \left(\prod_{i=1}^{k-2} R_i^{\text{local}}\right)(t_{k-1} - t_{k-2}) \\ &+ \cdots + q_1. \end{aligned} \tag{4}$$

This shows that the global position of a joint depends on a chain of matrix multiplications and additions through all its ancestors in the skeleton. The deeper the joint is in the kinematic tree, the more transformations are involved.

Error Accumulation in Local Methods. Existing diffusion-based methods supervise the generated motion using positional losses (e.g., L2 distance between joint or vertex positions). To compute positions, they first predict local joint rotations R_k^{local} and apply Forward Kinematics.

However, FK recursively composes transformations along the kinematic chain, as in equation (4). As a result, even small errors in earlier local rotations (e.g., R_i^{local}) are multiplied and propagated forward, resulting in increasingly large deviations in joint positions at greater depths in the hierarchy, as shown in Figure 1. This effect is especially severe at distal joints such as hands and feet, where accumulated errors lead to unstable or anatomically implausible motion—a phenomenon we term *hierarchical error accumulation*. Moreover, backpropagation through the FK chain involves deep, as in equation 4, nonlinear transformations, causing gradient instability and hindering effective training.

Our Global Rotation Prediction. To avoid the recursive composition of local rotations, we directly predict each joint's global rotation R_k^{global}. All joints are therefore defined in a shared world frame, eliminating depth-dependent multiplication of transformations. Following the explicit path-sum form derived in the previous subsection, the global position of joint k is written as

$$q_k = q_{\text{root}} + \sum_{(i \to j) \in \pi(k)} R_i^{\text{global}}(t_j - t_i), \tag{5}$$

where $\pi(k)$ denotes the unique parent-child path from the root to joint k, and $t_j - t_i$ is the rest-pose offset between consecutive joints along this path. This form matches the unfolded FK formulation but replaces local rotations with predicted global rotations.

Because this position computation is additive along $\pi(k)$ and avoids recursive rotation composition, each joint receives a direct and stable gradient with respect to its global rotation. This removes hierarchical error accumulation and improves robustness when training with positional supervision. Although direct global rotation prediction reduces the natural structural coupling enforced by standard FK, we introduce multi-level structure constraints in Section to restore these geometric relations.

Overall Structure

As shown in Figure 2, our model operates under the conditional flow matching (CFM) framework and takes as input a noised motion sequence x_t, audio features a, speaker identity, and a short seed motion clip. It predicts the clean global joint rotations and translations $x_1 \in \mathbb{R}^{T \times (J \times 6 + 3)}$, representing the full-body motion in 6D rotation format.

Audio and Expression Conditioning. Given a raw audio waveform, we extract high-level acoustic features $a \in \mathbb{R}^{T\times d_a}$ using a pretrained WavLM encoder (Chen et al. 2022). These features are then combined with the speaker identity vector PID and passed through a shallow Transformer encoder to directly estimate facial expressions $\mathcal{E} \in \mathbb{R}^{T\times d_e}$ from the audio and PID embedding. This design is motivated by the near one-to-one correspondence between phoneme sequences and lip movements, allowing a lightweight yet accurate mapping for expression synthesis.

Region-wise Motion Decomposition. To enhance learning capacity, we divide the motion sequence into hand and torso components. The noised global pose input $x_t \in \mathbb{R}^{T\times(J\times 6+3)}$ is decomposed into hand joints $\mathbf{H}_t$ and torso joints $\mathbf{B}_t$ based on a predefined joint partition. We then concatenate each of these with the expression and audio features along the channel dimension:

$$\mathbf{H}'_t = \text{Concat}(x_t, \mathbf{H}_t, \mathcal{E}, a), \quad \mathbf{B}'_t = \text{Concat}(x_t, \mathbf{B}_t, \mathcal{E}, a), \tag{6}$$

yielding input embeddings enriched with multi-modal cues.

Motion Generation Blocks. The refined hand and torso features, $\mathbf{H}'_t$ and $\mathbf{B}'_t$, are passed through separate stacks of Motion Generation Blocks (MGBs), each built as a residual Transformer layer inspired by DiT (Peebles and Xie 2023). Specifically, we apply FiLM-style (Perez et al. 2018) affine transformations at multiple layers, conditioning on the style vector $s_t = \text{Concat}(\text{MLP}(\text{PID}), \text{MLP}(t))$, the flow step t, and the seed motion. Each block also employs cross-attention to integrate style-aware context via a learned key-query mechanism.

Flow Matching Objective. Following (Tevet et al. 2022), our model adopts a simplified flow-matching formulation that directly learns to predict the clean motion sample $x_1 \sim p_1$ from an intermediate sample $x_t = (1-t)x_0 + tx_1$, rather than estimating the continuous velocity field along the flow path. Given x_t, the model is trained to output the corresponding target x_1 using a simple regression loss. The conditioning signal c consists of the expression features $\mathcal{E}$, high-level audio feature a, and the initial seed motion. The flow-matching loss is defined as:

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{t,x_0\sim p_0,x_1\sim p_1} \left\| f_\theta(x_t, c) - x_1 \right\|^2, \tag{7}$$

where f_θ is the model's prediction network, and x_t is obtained by linearly interpolating between x_0 and x_1.

Multi-Level Constraints

While our global rotation prediction removes error accumulation from recursive kinematic chains, it also discards the hierarchical structure inherent in local representation. As a result, directly learning global rotations can lead to physically implausible or unstable motions. To mitigate this, we introduce multi-level structure constraints at the joint, skeleton, and motion levels to restore structural coherence.

Joint Structure Constraint. Although our method learns global joint rotations through flow matching objectives, this rotation-based supervision alone does not provide joint structure information, e.g., skeleton length. To incorporate geometric information, MDM (Tevet et al. 2022) suggests adding spatial position supervision of the joint $\mathcal{L}_{pos}$. However, $\mathcal{L}_{pos}$ does not sufficiently constrain the relationship between rotation and position, because simply constraining the position of the joint will cause multiple valid rotations to produce the same joint position, as shown in 2, especially for the distal nodes. As can be seen from Equation (2), the position of the end node does not involve its own rotation. Therefore, we introduce joint-level constraints using virtual anchor points to introduce geometric priors into the learning process. These virtual points not only provide explicit rotation guidance, but also implicitly encode bone length and spatial structure, helping the model learn meaningful relationships between joint rotations and bone geometry.

To this end, besides $\mathcal{L}_{pos}$, we introduce a joint structure constraint based on *virtual anchor points*. For each joint k, we define N non-coplanar points $\{v_k^n\}_{n=1}^N$ in its local frame. Under predicted rotation R_k^{global}, they are transformed to:

$$\hat{v}_k^n = R_k^{\text{global}} \cdot v_k^n, \tag{8}$$

and compared with ground-truth rotated anchors:

$$\tilde{v}_k^n = R_k^{\text{gt}} \cdot v_k^n. \tag{9}$$

Since the anchors span 3D space, matching $\hat{v}_k^n$ to $\tilde{v}_k^n$ uniquely constrains rotation.

We define the supervision loss as:

$$\mathcal{L}_j = \frac{1}{KN} \sum_{k=1}^{K} \sum_{n=1}^{N} \|\hat{v}_k^n - \tilde{v}_k^n\|_2^2. \tag{10}$$

This constraint injects geometric priors into training and stabilizes rotation prediction, especially in expressive regions like hands and elbows. Although distal joints are more prone to rotational ambiguity, we apply this constraint to all joints to regularize global rotation learning across the entire skeleton and improve anatomical coherence.

Skeleton Structure Constraint. While the joint-level constraint enforces local rotational fidelity, it alone cannot capture the global anatomical structure of the human skeleton. In particular, human motion is governed by interdependent skeletal geometry, where the orientation of one bone is implicitly constrained by the orientations of all others. Prior works that supervise each joint independently fail to model this global coordination, leading to implausible artifacts such as asymmetric bending, unnatural twisting, and broken kinematic continuity.

To model this high-level structural regularity, we propose a skeleton-level constraint based on a pairwise Angular Matrix (AM) that captures angular relations between all bone pairs. For each bone defined by a joint pair (k, j) with valid positional data, we define the unit bone direction vector in global space as:

$$b_{k\to j} = \frac{q_j - q_k}{\|q_j - q_k\|_2}, \tag{11}$$

where q_k and q_j are the global positions of joints k and j.

Next, we define an Angular Matrix $\mathcal{A} \in \mathbb{R}^{K\times K}$, where each entry represents the cosine similarity between bone $b_{k\to j}$ and every other bone $b_{k'\to j'}$:

$$\mathcal{A}_{kj,k'j'} = b_{k\to j}^\top b_{k'\to j'}. \tag{12}$$

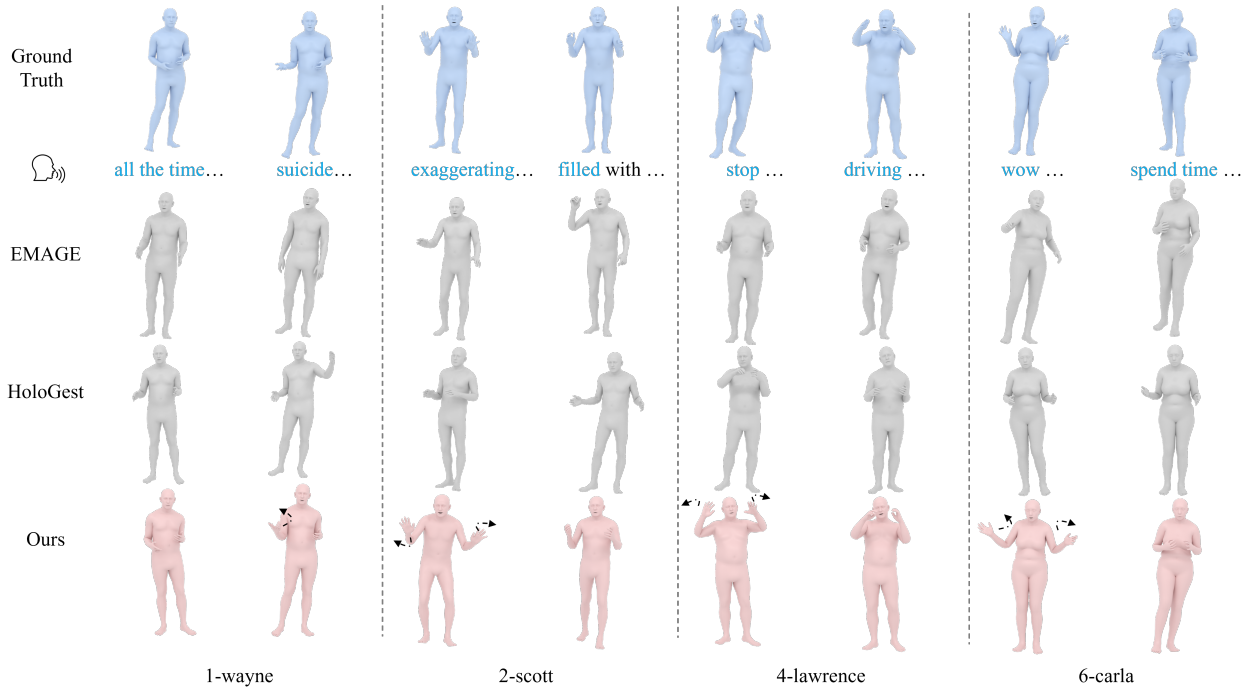

Figure 3: **Visual comparison.** Our GlobalDiff produces semantically meaningful gestures that align closely with the spoken content and speaker identity. For example, for the phrase "stop," Our GlobalDiff generates symmetric and contextually appropriate hand gestures near the head, conveying a clear intent. In contrast, HoloGest and EMAGE often result in unbalanced or semantically incoherent motions, such as asymmetric arms or ambiguous limb orientations.

This matrix encodes all pairwise bone orientation relations, thus capturing long-range skeletal dependencies and relative angular consistency between any pair of limbs, even if they are not directly connected in the kinematic tree.

Given predicted joint positions $\{q_k\}$ and ground-truth positions $\{\tilde{q}_k\}$, we compute their corresponding angular matrices $\mathcal{A}$ and $\tilde{\mathcal{A}}$. The skeleton structure constraint is defined as the mean squared error over all valid bone pairs:

$$\mathcal{L}_s = \frac{1}{|\mathcal{B}|} \sum_{(k,j),(k',j')\in\mathcal{B}} \left\| \mathcal{A}_{kj,k'j'} - \tilde{\mathcal{A}}_{kj,k'j'} \right\|_2^2, \quad (13)$$

where $\mathcal{B}$ is the set of all valid bone pairs with well-defined direction vectors.

Temporal Structure Constraint. While joint-level and skeleton-level constraints ensure spatial plausibility within individual frames, they do not capture the temporal structure of co-speech motion, which is inherently rhythmic and synchronized with the speech. Without explicit modeling of temporal dynamics, generated motions may appear erratic, lack rhythmic coherence, or fail to align with prosodic patterns. To address this, we introduce a motion-level constraint that enforces temporal consistency by aligning the dynamics of the generated motion sequence with those of the ground truth. Specifically, both the predicted motion $\hat{X} = \{\hat{x}_t\}_{t=1}^T$ and the reference motion $X = \{x_t\}_{t=1}^T$ are encoded into temporal embeddings using a shared multi-scale VAE $g(\cdot)$ to capture both short- and long-term dependencies.

We compute a direct perceptual alignment loss based on mean squared error (MSE) between the temporal embeddings:

$$z^{\text{gen}} = g(\hat{X}), \quad z^{\text{gt}} = g(X), \quad (14)$$

$$\mathcal{L}_m = \left\| z^{\text{gen}} - z^{\text{gt}} \right\|_2^2. \quad (15)$$

This formulation encourages the model to match global motion dynamics in latent space while maintaining simplicity and stability in training.

Experiments

Experimental Setup

Datasets. For training and evaluation, we adopt the BEAT2 dataset (Liu et al. 2024a), which provides approximately 60 hours of 3D motion data paired with speech from 25 speakers (12 female, 13 male). It contains 1,762 dialogue sequences, each lasting around 65 seconds on average. While prior methods typically report results on a fixed subset of speakers, we evaluate our model on both the full test set and the subset corresponding to speaker "Scott" to support fair comparison with single-speaker setups and assess generalization across diverse speakers, as recommended by (Mughal et al. 2025).

Implementation Details. Our model is trained on four NVIDIA V100 GPUs for 1,000 epochs with a batch size of 128, taking approximately 17 hours. We use the ADAM optimizer with a learning rate of 1e-4. We empirically set the number of virtual nodes to 6. Following (Liu et al. 2024a), we train the model starting with an 8-frame seed pose. For streaming, the last 8 frames from the previous clip are reused as the seed for the next, enabling long-form generation with only the initial 8 frames required.

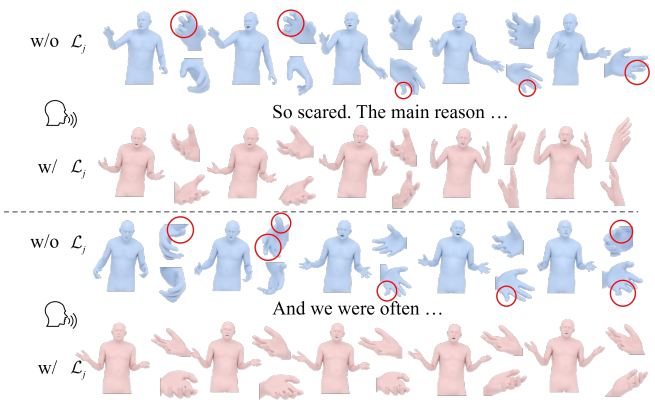

Figure 4: **Qualitative study on the effect of $\mathcal{L}_j$.** Without $\mathcal{L}_j$, finger poses often appear anatomically implausible.

Metrics. To evaluate the quality of generated gestures, we use several metrics targeting different aspects of performance. For overall realism, we adopt the Fréchet Gesture Distance (FGD) (Yoon et al. 2020), which compares the statistical distribution of generated gestures to that of the ground truth (GT). Motion diversity (Diversity) is measured using (Li et al. 2021a), defined as the mean L1 distance between pairs of generated clips. Temporal coordination between speech and motion is assessed with the Beat Alignment (BeatAlign) metric (Li et al. 2021b), which reflects how well gesture timing aligns with the rhythm of the accompanying audio. For facial accuracy, we report vertex MSE (Yang et al. 2023), measuring deviations between predicted and ground-truth facial vertices.

Qualitative Results

Qualitative Comparisons. As shown in Figure 3, GlobalDiff consistently generates semantically grounded and physically coherent co-speech motions across all speaker identities, outperforming baselines. We compare gestures produced by all three methods on four speakers—Wayne (ID1), Scott (ID2), Lawrence (ID4), and Carla (ID6)—across varied utterances. RAG-GESTURE (Mughal et al. 2025) is excluded due to unavailable public outputs. For the phrase *"exaggerating"*, GlobalDiff produces wide, symmetric arm extensions that clearly express emphasis, whereas EMAGE shows constrained gestures and HoloGest yields flat, non-expressive arm poses. For *"wow"*, our model generates elevated open-hand motions that reflect surprise, while EMAGE collapses the upper limbs and HoloGest shows minimal variation. When expressing *"driving"*, GlobalDiff synthesizes realistic, steering-like motions. In contrast, EMAGE and HoloGest often fall back on generic, low-effort gestures. For *"all the time"*, our model maintains consistent chest-level hand motions across speakers, preserving semantic clarity, while EMAGE introduces lateral imbalance and HoloGest shows reduced motion fidelity. This comparison underscores the advantage of our global rotation strategy and structural constraints in producing expressive, semantically accurate, and identity-consistent gestures.

Joint Structure Constraint. We visualize hand close-ups from motions generated with and without $\mathcal{L}_j$. Without $\mathcal{L}_j$, the model frequently produces anatomically implausible joint rotations—such as unnaturally flipped thumbs or twisted little fingers (highlighted in red)—due to the under-constrained nature of rotation prediction. In contrast, incorporating $\mathcal{L}_j$ guides the model to produce structurally consistent and physically realistic finger orientations by supervising global rotations via virtual anchor points.

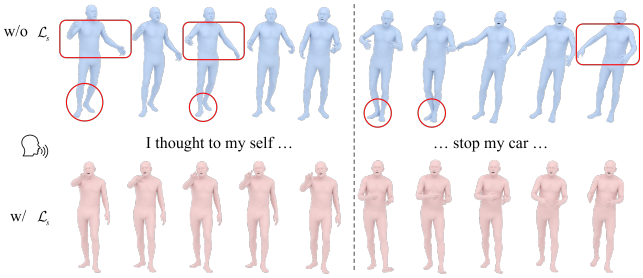

Figure 5: **Qualitative study on the effect of $\mathcal{L}_s$.** Without $\mathcal{L}_s$, motion becomes structurally incoherent, exhibiting unbalanced posture and unstable foot placement.

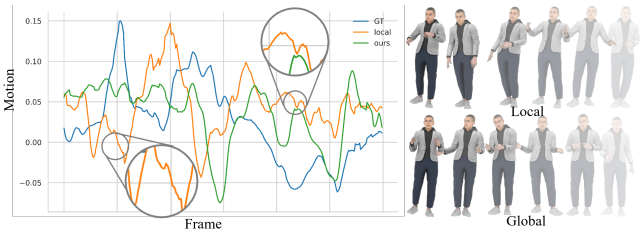

Figure 6: **Ablation Study on Global vs. Local Rotation Prediction.** Our global method produces smoother fingertip motion and more stable body posture compared to the noisy and distorted results from local rotation prediction.

Skeleton Structure Constraint. As shown in Figure 5, without $\mathcal{L}_s$, the generated motions exhibit structural incoherence, such as unnatural leaning, unbalanced steps, and asymmetric limb configurations. These artifacts arise from the lack of explicit constraints on angular relationships between bones. In contrast, incorporating $\mathcal{L}_s$ enforces consistent bone orientations across the body, resulting in smoother and more anatomically coherent movements.

Ablation Study on Global vs. Local Rotation Prediction. Figure 6 presents a qualitative comparison between our global rotation strategy, traditional local rotation methods, and ground truth (GT). On the left, we plot the trajectory of the right middle fingertip over 300 frames from the test sequence "2-scott-0-103-103". The local method suffers from high-frequency oscillations and instability, while our global method produces a smoother trajectory that aligns more closely with GT. On the right, snapshots sampled every 50 frames show that the local method introduces body tilt and severe finger deformation, especially at distal joints. In contrast, our method maintains natural body posture and stable articulation, demonstrating its effectiveness in mitigating hierarchical error accumulation.

Quantitative Results

Comparison with Baselines. Table 1 provides a detailed quantitative comparison between our method and several

	1 Speaker				All Speakers			
	FGD↓	BeatAlign→	Diversity→	MSE↓	FGD↓	BeatAlign→	Diversity→	MSE↓
GT	–	0.703	11.97	–	–	0.477	7.29	–
CaMN (Liu et al. 2022a)	0.604	0.711	9.97	–	0.512	0.200	5.58	–
Audio2Photoreal (Ng et al. 2024)	1.02	0.550	12.47	–	0.849	0.326	6.24	–
ReMoDiffuse (Zhang et al. 2023)	0.702	0.824	12.46	–	1.120	0.218	5.06	–
DSG (Yang et al. 2023)	0.881	0.724	11.49	–	1.174	0.734	11.12	–
HoloGest (Cheng and Huang 2025)	0.534	0.795	14.15	–	0.646	0.803	13.53	–
EMAGE (Liu et al. 2024a)	0.570	0.793	11.41	7.680	0.692	0.284	6.06	6.908
SemTalk (Zhang et al. 2025a)	0.428	0.777	12.91	**6.153**	-	-	-	-
RAG-GESTURE (Mughal et al. 2025)	0.808	0.734	**11.97**	–	0.487	**0.514**	9.94	–
Ours	**0.478**	**0.705**	13.73	6.330	**0.263**	0.404	**8.24**	**4.144**

Table 1: **Comparison with state-of-the-art methods trained on BEAT2.** We demonstrate superior performance, especially when generalizing well across multiple speaker identities. We report MSE$\times 10^{-}8$ for simplify.

Method	**FGD↓**	**BeatAlign→**	**Diversity→**
GT	–	0.703	11.97
Ours(local)	0.594	0.578	9.33
Ours(global)	0.592	0.693	13.08
+ $\mathcal{L}_j$	0.574	0.665	**12.30**
+ $\mathcal{L}_j$ + $\mathcal{L}_s$	0.517	0.593	13.78
+ $\mathcal{L}_j$ + $\mathcal{L}_s$ + $\mathcal{L}_m$	**0.478**	**0.705**	13.73

Table 2: **Ablation study on the effectiveness of each component within the GlobalDiff on the speaker 2.**

state-of-the-art baselines on both single-speaker and multi-speaker co-speech motion generation. Our method consistently achieves the best performance across nearly all metrics. Specifically, we obtain the lowest FGD and comparable MSE, indicating superior spatial fidelity and motion reconstruction accuracy. On BeatAlign, our method performs comparably with the top baselines, reflecting robust temporal alignment with speech rhythm. Although RAG-GESTURE attains the closest diversity with GT on the single-speaker setting, our method maintains competitive diversity scores while preserving structural integrity and semantic consistency. These results highlight the strength of our global rotation modeling and multi-level structural constraints in producing accurate, expressive, and speaker-agnostic co-speech motion.

Ablation Study on Components. Table 2 presents the results of an ablation study evaluating the impact of each major component in our GlobalDiff pipeline. Starting from a local-rotation baseline with $\mathcal{L}_{pos}$, we observe limited performance in both FGD and BeatAlign, highlighting the limitations of hierarchical propagation. Switching to global rotations improves rhythm alignment and motion diversity. Introducing $\mathcal{L}_j$ further enhances BeatAlign and diversity, indicating better semantic expressiveness and orientation precision. Adding $\mathcal{L}_s$ improves FGD substantially, reflecting enhanced anatomical plausibility. Finally, incorporating $\mathcal{L}_m$ brings all metrics to their best levels, demonstrating that our multi-level structure design jointly contributes to realism, rhythm alignment, and expressive richness.

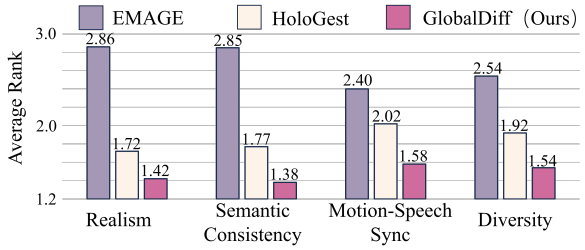

Figure 7: **Results of the user study.** GlobalDiff received higher preference across all evaluation metrics, with particularly strong scores in realism and semantic consistency.

User Study. We conducted a user study with 10 video samples and 28 participants, evaluating realism, semantic consistency, and motion-speech synchrony. Participants were asked to rank anonymized and shuffled outputs from GlobalDiff, EMAGE, and HoloGest. As shown in Figure 7, GlobalDiff received higher preference across all metrics.

Conclusion

In this work, we presented GlobalDiff, a novel diffusion-based framework for holistic co-speech motion generation that addresses the fundamental issue of hierarchical error accumulation inherent in local joint rotation approaches. By leveraging global joint rotations, GlobalDiff effectively decouples the prediction process across joints, significantly reducing cumulative errors, especially at distal limbs. To overcome the challenge of structural inconsistency arising from the loss of implicit hierarchical constraints, we introduced a progressive multi-level constraint scheme. Our joint-level constraint employs virtual anchor points for precise orientation guidance, the skeleton-level constraint ensures angular coherence among bones, and the motion-level constraint enforces temporal consistency through a multi-scale variational encoder. Extensive evaluations on standard co-speech benchmarks demonstrate that GlobalDiff not only achieves state-of-the-art performance in terms of motion stability, structural coherence, and expressiveness but also establishes a robust baseline for future research in structurally-aware global rotation diffusion methods.

Acknowledgments. This work was supported by Alibaba Research Intern Program. The numerical calculation is supported by Tongyi Lab, Alibaba Group.

References

Aberman, K.; Li, P.; Lischinski, D.; Sorkine-Hornung, O.; Cohen-Or, D.; and Chen, B. 2020. Skeleton-aware networks for deep motion retargeting. *ACM Transactions on Graphics (TOG)*, 39(4): 62–1.

Ahuja, C.; Joshi, P.; Ishii, R.; and Morency, L.-P. 2023. Continual learning for personalized co-speech gesture generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 20893–20903.

Ahuja, C.; Lee, D. W.; and Morency, L.-P. 2022. Low-resource adaptation for personalized co-speech gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20566–20576.

Alexanderson, S.; Nagy, R.; Beskow, J.; and Henter, G. E. 2023. Listen, denoise, action! audio-driven motion synthesis with diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4): 1–20.

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

Chen, B.; Li, Y.; Ding, Y.-X.; Shao, T.; and Zhou, K. 2024a. Enabling synergistic full-body control in prompt-based co-speech motion generation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 6774–6783.

Chen, J.; Liu, Y.; Wang, J.; Zeng, A.; Li, Y.; and Chen, Q. 2024b. 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.

Chen, S.; Wang, C.; Chen, Z.; Wu, Y.; Liu, S.; Chen, Z.; Li, J.; Kanda, N.; Yoshioka, T.; Xiao, X.; et al. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6): 1505–1518.

Cheng, Y.; and Huang, S. 2025. HoloGest: Decoupled Diffusion and Motion Priors for Generating Holisticly Expressive Co-speech Gestures. *arXiv preprint arXiv:2503.13229*.

Chhatre, K.; Athanasiou, N.; Becherini, G.; Peters, C.; Black, M. J.; Bolkart, T.; et al. 2024. Emotional speech-driven 3d body animation via disentangled latent diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1942–1953.

Guo, C.; Mu, Y.; Javed, M. G.; Wang, S.; and Cheng, L. 2024. Momask: Generative masked modeling of 3d human motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1900–1910.

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

He, X.; Huang, Q.; Zhang, Z.; Lin, Z.; Wu, Z.; Yang, S.; Li, M.; Chen, Z.; Xu, S.; and Wu, X. 2024. Co-speech gesture video generation via motion-decoupled diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2263–2273.

Kingma, D. P.; Welling, M.; et al. 2013. Auto-encoding variational bayes.

Kucuk, S.; and Bingul, Z. 2006. *Robot kinematics: Forward and inverse kinematics*, volume 1. INTECH Open Access Publisher London, UK.

Lee, D.; Kim, C.; Kim, S.; Cho, M.; and Han, W.-S. 2022. Autoregressive image generation using residual quantization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11523–11532.

Li, H.; Dai, J.; Zhao, X.; Zhou, F.; Pan, J.; and Li, L. 2025. Wav2Sem: Plug-and-Play Audio Semantic Decoupling for 3D Speech-Driven Facial Animation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 183–192.

Li, J.; Kang, D.; Pei, W.; Zhe, X.; Zhang, Y.; He, Z.; and Bao, L. 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.

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

Liang, Y.; Feng, Q.; Zhu, L.; Hu, L.; Pan, P.; and Yang, Y. 2022. Seeg: Semantic energized co-speech gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10473–10482.

Liu, H.; Zhu, Z.; Becherini, G.; Peng, Y.; Su, M.; Zhou, Y.; Zhe, X.; Iwamoto, N.; Zheng, B.; and Black, M. J. 2024a. 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.

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

Liu, X.; Wu, Q.; Zhou, H.; Xu, Y.; Qian, R.; Lin, X.; Zhou, X.; Wu, W.; Dai, B.; and Zhou, B. 2022b. 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.

Liu, Y.; Cao, Q.; Wen, Y.; Jiang, H.; and Ding, C. 2024b. Towards variable and coordinated holistic co-speech motion generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1566–1576.

Ma, Y.; Liu, Y.; Zhu, Q.; Yang, A.; Feng, K.; Zhang, X.; Li, Z.; Han, S.; Qi, C.; and Chen, Q. 2025. Follow-Your-Motion: Video Motion Transfer via Efficient Spatial-Temporal Decoupled Finetuning. *arXiv preprint arXiv:2506.05207*.

Mughal, M. H.; Dabral, R.; Scholman, M. C.; Demberg, V.; and Theobalt, C. 2025. Retrieving Semantics from the Deep: an RAG Solution for Gesture Synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 16578–16588.

Ng, E.; Romero, J.; Bagautdinov, T.; Bai, S.; Darrell, T.; Kanazawa, A.; and Richard, A. 2024. From Audio to Photoreal Embodiment: Synthesizing Humans in Conversations. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Pavlakos, G.; Choutas, V.; Ghorbani, N.; Bolkart, T.; Osman, A. A.; Tzionas, D.; and Black, M. J. 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.

Peebles, W.; and Xie, S. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4195–4205.

Peng, Z.; Wu, H.; Song, Z.; Xu, H.; Zhu, X.; He, J.; Liu, H.; and Fan, Z. 2023. Emotalk: Speech-driven emotional disentanglement for 3d face animation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 20687–20697.

Perez, E.; Strub, F.; De Vries, H.; Dumoulin, V.; and Courville, A. 2018. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Qi, X.; Liu, C.; Li, L.; Hou, J.; Xin, H.; and Yu, X. 2024. Emotiongesture: Audio-driven diverse emotional co-speech 3d gesture generation. *IEEE Transactions on Multimedia*, 26: 10420–10430.

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

Yang, S.; Wu, Z.; Li, M.; Zhang, Z.; Hao, L.; Bao, W.; Cheng, M.; and Xiao, L. 2023. DiffuseStyleGesture: Stylized Audio-Driven Co-Speech Gesture Generation with Diffusion Models. *arXiv e-prints*, arXiv–2305.

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

Yoon, Y.; Cha, B.; Lee, J.-H.; Jang, M.; Lee, J.; Kim, J.; and Lee, G. 2020. Speech gesture generation from the trimodal context of text, audio, and speaker identity. *ACM Transactions on Graphics (TOG)*, 39(6): 1–16.

Zhang, F.; Wang, Z.; Lyu, X.; Zhao, S.; Li, M.; Geng, W.; Ji, N.; Du, H.; Gao, F.; Wu, H.; et al. 2024a. Speech-driven personalized gesture synthetics: Harnessing automatic fuzzy feature inference. *IEEE Transactions on Visualization and Computer Graphics*, 30(10): 6984–6996.

Zhang, J.; Huang, S.; Tu, Z.; Chen, X.; Zhan, X.; YU, G.; and Shan, Y. 2024b. TapMo: Shape-aware Motion Generation of Skeleton-free Characters. In *The Twelfth International Conference on Learning Representations*.

Zhang, M.; Guo, X.; Pan, L.; Cai, Z.; Hong, F.; Li, H.; Yang, L.; and Liu, Z. 2023. Remodiffuse: Retrieval-augmented motion diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 364–373.

Zhang, X.; Li, J.; Zhang, J.; Dang, Z.; Ren, J.; Bo, L.; and Tu, Z. 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.

Zhang, X.; Li, J.; Zhang, J.; Ren, J.; Bo, L.; and Tu, Z. 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.

Zhi, Y.; Cun, X.; Chen, X.; Shen, X.; Guo, W.; Huang, S.; and Gao, S. 2023. Livelyspeaker: Towards semantic-aware co-speech gesture generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 20807–20817.

Zhu, L.; Liu, X.; Liu, X.; Qian, R.; Liu, Z.; and Yu, L. 2023. Taming diffusion models for audio-driven co-speech gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10544–10553.