InterSyn: Interleaved Learning for Dynamic Motion Synthesis in the Wild

Yiyi Ma[*1] Yuanzhi Liang[2] Xiu Li [1†] Chi Zhang [2] Xuelong Li[2†]
Shenzhen International Graduate School, Tsinghua University[1]
Institute of Artificial Intelligence, China Telecom[2]

Abstract

We present Interleaved Learning for Motion Synthesis (InterSyn), a novel framework that targets the generation of realistic interaction motions by learning from integrated motions that consider both solo and multi-person dynamics. Unlike previous methods that treat these components separately, InterSyn employs an interleaved learning strategy to capture the natural, dynamic interactions and nuanced coordination inherent in real-world scenarios. Our framework comprises two key modules: the Interleaved Interaction Synthesis (INS) module, which jointly models solo and interactive behaviors in a unified paradigm from a first-person perspective to support multiple character interactions, and the Relative Coordination Refinement (REC) module, which refines mutual dynamics and ensures synchronized motions among characters. Experimental results show that the motion sequences generated by InterSyn exhibit higher text-to-motion alignment and improved diversity compared with recent methods, setting a new benchmark for robust and natural motion synthesis. Additionally, our code will be open-sourced in the future to promote further research and development in this area. Project website: https://myy888.github.io/InterSyn/

1. Introduction

Given the crucial role of motion information in applications such as visual perception [21, 23, 24], visual generation [3, 44], and 3D synthesis [13, 40, 42], text-to-motion (T2M) synthesis [11, 18, 36] has garnered increasing attention. Despite significant progress on benchmarks like HumanML3D [10], generating realistic and diverse human motion in the wild remains challenging. Current motion generators [10, 22, 31, 37] often excel on specific datasets yet struggle to generalize to the varied and complex motions observed in real-world interactions. While some works [6, 20, 45] have explored general motion generation for multiple characters, the generalization issue is still largely unresolved.

A promising direction for achieving motion generalization is to simulate real human learning. In everyday life, human motion arises from two intertwined components: semantic expression and social interaction. Motions for self-expression, such as walking and talking, are fundamental to individual development and well-being [15–17, 25, 33, 38], while social motions like hugging or gesturing are primarily acquired through observation, imitation, and active engagement with others [1, 8, 29]. The dynamic nature of human learning, further supported by theories such as situated learning [7, 12, 32, 39] and Cultural-Historical Activity Theory [5, 14, 28], as well as studies on peer group interactions [2, 4, 19], suggests that semantic and social motions mutually reinforce each other, providing a natural blueprint

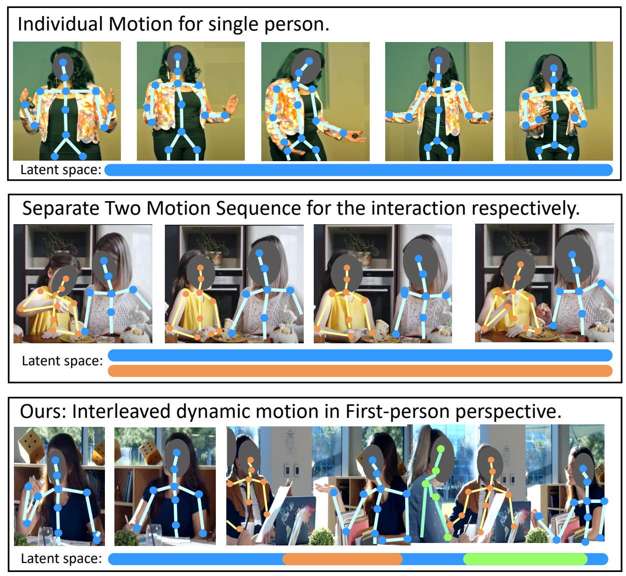

Figure 1. Unlike conventional approaches that process single-person and interaction motions separately, real human behavior exhibits a dynamic interleaving of solo and multi-person actions. Our InterSyn framework learns to generate these integrated motion sequences, closely emulating the fluid and natural dynamics observed in real-world interactions. For multi-human motion, InterSyn takes separate descriptions as inputs to better capture distinct roles and coordination.

[*]This work was performed at TeleAI.

for more generalizable motion synthesis.

However, existing motion synthesis pipelines suffer from two critical limitations. First, they treat semantic and social motions as separate tasks—semantic actions such as walking or talking are typically learned in isolation (e.g., in HumanML3D [10]), while social interactions are modeled independently (e.g., in InterGen [20]). As shown in Fig 1, this separation restricts the ability of models to generate the dynamic behaviors seen in natural settings. Second, many methods fail to capture the mutual cues of human motion, where individuals continuously adapt their movements in response to subtle signals from others. These dynamic adaptations are essential for producing vivid motion expressions.

To this end, we propose a novel framework, Interleaved Learning for Motion Synthesis (InterSyn), which learns to generate dynamic motion sequences by seamlessly integrating single-person and interaction motions. InterSyn comprises two key modules: the Interleaved Interaction Synthesis (INS) and the Relative Coordination Refinement (REC). The INS module jointly models single-person and multi-person motions in a unified learning paradigm, enabling the learning of motion that akin to the real human being, who who change dynamically with solo motion and interaction motion seamlessly. Meanwhile, the REC module refines the synthesized interactions by leveraging the mutual dynamics and nuanced interdependencies among multiple characters, ensuring that the generated motions are both semantically coherent and socially synchronized.

Our extensive experiments demonstrate that InterSyn significantly outperforms existing methods in generating high-quality interaction motions. Quantitative evaluations on multiple datasets reveal substantial improvements in key metrics such as semantic coherence and synchronization between interacting individuals; for example, our method outperforms recent work [6] by 40% in terms of FID. Qualitative results further illustrate that our approach produces natural and nuanced motion sequences that capture the complex dynamics of real-world interactions. These findings confirm that by jointly modeling and refining solo and interactive dynamics, InterSyn effectively addresses the challenges of interaction motion synthesis, setting a new standard for robust and realistic motion generation.

Our contributions are threefold:

- We introduce a novel framework, **InterSyn**, that jointly models single-person and interaction motions within a unified dynamic sequence, enabling the generation of realistic motions that naturally interleave individual and multi-person behaviors.
- We propose **Iterative Interaction Coordination (INC)** and **Relative Coordination Refinement (REC)**. INC facilitates dynamic motion generation by interleaving individual and interactive behaviors, while REC refines these interactions through relative coordination constraints that capture mutual dynamics among characters.
- Extensive experiments demonstrate that our approach significantly outperforms recent methods in generating diverse and realistic interactions, confirming the effectiveness of interleaved learning for natural motion synthesis.

2. Related Work

Text-to-motion generation has gained significant attention for its applications in animation, virtual reality, and human-computer interaction. Research in this area has evolved from synthesizing realistic single-person motions to capturing complex multi-person interactions.

Single-Person Motion Generation: Diffusion-based approaches have emerged as powerful paradigms for text-to-motion synthesis. For example, the Motion Diffusion Model (MDM) [36] adapts classifier-free diffusion by directly predicting the motion sample at each step, which facilitates the integration of geometric losses. Other works bypass diffusion by directly regressing pose sequences. MoMask [11] introduces a hierarchical masked modeling framework that represents motion as multi-layer discrete tokens, where a Masked Transformer predicts masked tokens conditioned on text and a Residual Transformer refines these tokens, achieving notable improvements on benchmarks like HumanML3D and KIT-ML. In addition, methods such as OmniControl [43] enable controllable motion synthesis via advanced conditioning techniques that allow users to specify attributes like speed and direction, while TAAT [41] and emotion-enriched approaches leverage large language models (LLMs) to incorporate arbitrary texts and subjective motion details.

Multi-Person Motion Generation: Moving beyond isolated actions, multi-person motion generation focuses on capturing the interactions and relationships among individuals. SMooDi [45] addresses stylized motion synthesis by combining content text with reference style sequences using a pre-trained text-to-motion model and a lightweight style adaptor. FreeMotion [6] overcomes the limitations of fixed agent numbers by unifying single- and multi-person synthesis via a conditional motion distribution, enabling number-free generation with precise spatial control. Other approaches emphasize long-term, socially-aware interactions. SATA [27] employs a scene-aware social transformer that fuses motion and scene information within a diffusion framework, while InterGen [20] tailors diffusion-based generation to two-person interactions using cooperative transformer denoisers with mutual attention and a global relation representation.

3. Preliminary

Diffusion models generate human motion through a probabilistic framework that iteratively denoises Gaussian noise into realistic sequences, extending their success in text-to-

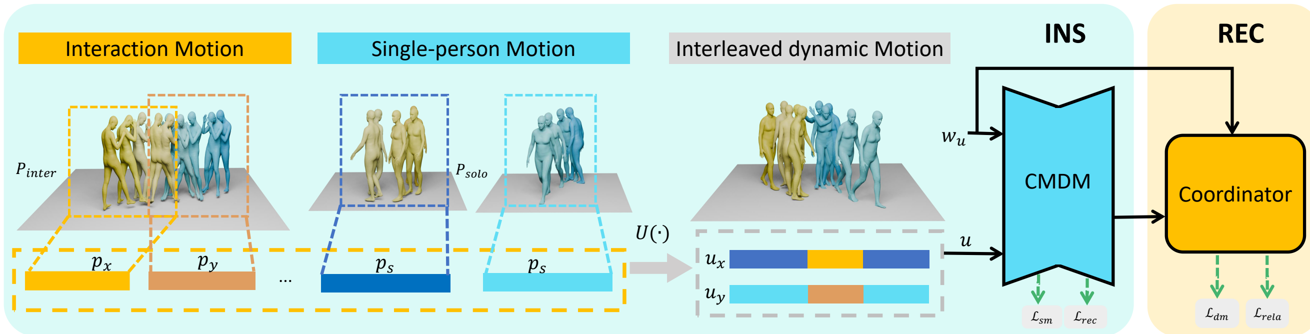

Figure 2. Overview of our interleaved learning framework (InterSyn) for dynamic motion synthesis. Our method consists of two stages: (1) Interleaved Interaction Synthesis (INS), which unifies single-person and interaction motions using a Conditional Motion Diffusion Model to generate a flexible, realistic motion schema, and (2) Relative Coordination Refinement (REC), which employs a Coordinator network to refine interactions based on mutual relationships among characters, yielding vivid and nuanced motion generation.

image synthesis [37]. Unlike traditional autoregressive methods, diffusion models synthesize full motion sequences holistically, addressing temporal coherence and physical plausibility.

The process involves a forward noising stage that corrupts ground-truth motion $x_0^* \in \mathbb{R}^{N\times D}$ (with N frames and D-dimensional joint parameters) into Gaussian noise x_T, and a learned reverse process that reconstructs x_0 from x_t. At each diffusion step t, the reverse transition is governed by:

$$P_\theta(x_{t-1}|x_t) = \mathcal{N}\left(\mu_t(\theta), \sigma_t I\right) \tag{1}$$

where $\mu_t(\theta)$ combines the predicted clean motion $x_0(\theta)$ and noisy input x_t:

$$\mu_t(\theta) = \sqrt{\alpha_{t-1}}x_0(\theta) + \sqrt{1-\alpha_{t-1}-\sigma_t^2}\cdot x_t \tag{2}$$

Here, $\alpha_t \in (0,1)$ controls noise scheduling, with $\beta_t = 1-\alpha_t$, and $\sigma_t^2 = \beta_t(1-\alpha_{t-1})/(1-\alpha_t)$. The model directly predicts x_0 rather than incremental noise, optimizing via $\mathcal{L} = \|x_0 - x_0^*\|_2^2$.

During inference, motion is synthesized by recursively denoising $x_T \sim \mathcal{N}(0, I)$, leveraging cross-attention layers to align text embeddings with spatiotemporal motion features. This framework ensures physically plausible outputs while preserving semantic alignment with text prompts.

4. Interleaved Learning for Motion Synthesis

4.1. Interleaved Interaction Synthesis

Human motion comprises both individual movements (for self-expression) and multi-person interactions (for social engagement). In natural settings, people fluidly transition between these modes, adapting their actions to the context. To capture this behavior, we propose the Interleaved Interaction Synthesis (INS) framework, which jointly models single-person and interaction motions in a unified paradigm. Unlike existing methods that treat these motion types independently, INS learns them together, enabling seamless transitions between solo and interactive actions and closely mimicking the fluidity of real human movement.

We begin by processing and unifying all single-person and interaction motions from a first-person perspective. Specifically, we initialize a motion *bucket* u to represent the interleaved sequence, where

$$u = (u_x, u_y) \in \mathbb{R}^{2\times T\times K\times C} \tag{3}$$

with T denoting the number of time steps, K the number of keypoints per motion, and C the dimensionality of each keypoint. To construct u that adaptively interleaves single-person and multi-person motions, we first randomly sample a single-person motion p_s from a diverse dataset [36].

Then, for multi-person interactions, let

$$P_i = \{p_1, p_2, \ldots, p_N\}$$

denote the set of motions from N individuals. We randomly select a pair of interaction motions (p_x, p_y) to interleave with p_s. These motions are combined into a unified sequence using a function $U(\cdot)$:

$$u = U(p_x, p_y, p_s, t_i, t_s) \tag{4}$$

where t_i and t_s are the start time indices for the interaction motion (p_x, p_y) and the solo motion p_s, respectively. The sequence u_x can begin with either p_x or p_s (ensuring $t_i + t_s > 0$ and $t_i \cdot t_s = 0$) so that exactly one segment starts at the first time step. The $U(\cdot)$ manages motion alignment, smooth transitions, and orientation adjustments between segments. Based on this, "interleave" means blending multi-person interaction into single-person motions using a function that alternates between solo and interactive actions.

Due to the mismatch of skeletons in the fusion of two datasets, we need to apply rotation and translation to the skeleton sk_2 relative to sk_1 in order to make their expression of the motion consistent. In forward kinematics (FK), the goal is to determine the global position of each joint in a skeletal structure based on its local transformations, which consist of rotations and translations. Let the global position of the k-th joint in skeleton sk_1 be denoted as $j_k^{sk_1}$, and in skeleton sk_2 as $j_k^{sk_2}$. Transforming the motion from one skeleton to another requires adjusting both the bone lengths and rotations of the joints.

The process starts with the root joint, which serves as the starting point for all subsequent transformations. We assume that the global position of the root joint is the same in both skeletons. For each joint k, the relative position from its parent joint in both skeletons is adjusted by scaling the bone length and applying the rotational difference. In forward kinematics, starting from the root node and extending to the child nodes, the transformation from sk_2 to sk_1 can be expressed as:

$$j_k^{sk_1} = R_i \cdot j_{k-1}^{sk_2} + T_i(j_k^{sk_2} - j_{k-1}^{sk_2}) \tag{5}$$

where R_i is the rotation matrix for the i-th joint relative to the parent joint j_{i-1} in sk_2 and sk_1, and T_i is the translation matrix corresponding to the bone length difference between the i-th joint in sk_2 and sk_1 relative to the parent joint j_{i-1}. More details about the $U(\cdot)$ can be found in the appendix.

Moreover, text descriptions for p_s, p_x, and p_y are denoted as w_s, w_x, and w_y. We concatenate w_s and w_x into a combined description w_u for u.

To synthesize the dynamic motion u, we employ a Conditional Motion Diffusion Model (CMDM), denoted as M_s. As shown in Fig. 2, M_s is a Transformer-based diffusion network [36] that incorporates multiple denoising blocks. The model takes the interaction time steps t_i and t_s as conditioning inputs and includes time embedding layers to encode t_i and t_s, as well as a text embedding layer to encode w_u. During training, at time step t, we add noise to the motion u to obtain a noisy motion u^t, and concatenate the resulting embeddings, which are then fed into the diffusion model:

$$\hat{u} = (\hat{u_x}, \hat{u_y}) = M_s(u, w_u, t_i, t_s) \tag{6}$$

By conditioning on these temporal signals, the diffusion model learns to generate a continuous motion sequence that smoothly interleaves individual and interactive actions.

We optimize INS using the following loss function:

$$\mathcal{L}_I = \lambda_1 \mathcal{L}_{\text{rec}} + \lambda_2 \mathcal{L}_{\text{smooth}} \tag{7}$$

where $\mathcal{L}_{\text{rec}}$ is the reconstruction loss of the CMDM for p_u, and $\mathcal{L}_{\text{smooth}}$ enforces smooth transitions at the boundaries (within a window of ± 5 frames) during motion fusion. The hyperparameters λ_i balance these loss terms.

$$\mathcal{L}(u)_{\text{smooth}} = \sum_{t=t_i-5}^{t_i+5} \left| \frac{d}{dt}\,(u(t)) \right| \tag{8}$$

While our framework supports multiple interleaved segments (by introducing additional t_i and t_s), we observe that more than two segments per sequence degrade performance. This is because increasing the number of segments reduces the duration allocated to each motion, potentially truncating key motion phases and hindering semantic capture.

4.2. Relative Coordination Refinement

Interaction motions require both individual semantic coherence (each person's motion is contextually meaningful) and synchronized social coordination (movements are responsive and harmonized). To address this, we propose the Relative Coordination Refinement (REC) module, which refines multi-person interactions by considering spatial relationships and mutual influence among characters.

In REC, a coordinator network M_c (see Fig. 2) refines the interaction predictions. We use the Transformer network architecture to build the coordinator module. The Transformer is primarily used to model the spatial and temporal relationships between multiple motion sequences. During training phase for example, in a two-person interaction as in [20], the first person's predicted motion can be refined using the second's motion, and vice versa. Let $\hat{u_y}$ be the reference motion. The refinement process for $\hat{u_x}$ is as follows:

$$\phi_x = M_c\big(\hat{u}_x,\, \hat{u}_y,\, w_u\big) \tag{9}$$

where $\hat{u_x}$ is the motion reconstructed by M_s via the diffusion process and is further refined by M_c to yield ϕ_x.

Similarly, we process the reference interaction motion u_y by passing it through the diffusion model M_s without applying any temporal offset. We obtain:

$$\phi_y = M_c\big(\hat{u}_y,\, \phi_x,\, w_u\big) \tag{10}$$

We model the problem of handling social coordination as the conditional distribution $P(\hat{u_x}|\hat{u_y})$, where the model needs to learn how to process the action sequence $\hat{u_x}$ based on the spatial information of $\hat{u_y}$. We simulate the M_c module as a mechanism that performs translation and rotation on the spatial positions of $\hat{u_x}$'s actions, as well as a rationalization of its trajectory, ensuring that $\hat{u_y}$ can interact appropriately with $\hat{u_x}$.

Therefore, when ϕ_x represents a reasonable interaction action relative to $\hat{u_y}$, we fine-tune $\hat{u_y}$ conditionally to obtain ϕ_y. Based on the functional definition of the M_c module, ϕ_y should be infinitesimally close to $\hat{u_y}$. Based on this, we define the relative coordination loss as:

$$\mathcal{L}_{\text{rela}}(\phi_y, \hat{u_y}) = \|\phi_y - \hat{u_y}\|_2 \tag{11}$$

The REC loss is formulated as:

$$\mathcal{L}_R = \lambda_3 \mathcal{L}_{\text{rela}}(\phi_y, \hat{u_y}) + \lambda_4\, \mathcal{L}_{\text{dm}}(\phi_x, \hat{u_y}) \tag{12}$$

where $\mathcal{L}_{\text{dm}}$ is the masked joint distance map loss from [20].

Finally, the overall loss function is:

$$\mathcal{L} = \mathcal{L}_I + \mathcal{L}_R \tag{13}$$

5. Experiment

Datasets. We conducted experiments on the **HumanML3D** [10] single-human dataset and the **InterHuman** [20] dual-human dataset. The HumanML3D dataset re-annotates the AMASS dataset [26] and the HumanAct12 dataset [9], offering a total of 44,970 annotations across 14,616 motion sequences. The InterHuman dataset is the first dual-human motion dataset with text annotations, containing 6,022 different actions across various human activity types, each action accompanied by 16,756 unique textual descriptions, using 5,656 different words.

Evaluation Metrics. We follow the evaluation protocol outlined by [10]. To evaluate the naturalness of the generated motions, we utilize the Frechet Inception Distance (FID). R-Precision is employed to assess the alignment between the generated motions and the text prompt, while Diversity quantifies the variation within the generated motions. Multimodality (MModality) captures the diversity present within a single text prompt. Finally, Multi-modal distance (MM Dist) measures the distance between the motion and text feature representations.

5.1. Implementation Details

During training, we adopt an alternating training strategy using interleaved data and HumanML3D data. Specifically, within a single learning process, the model computes the reconstruction loss $\mathcal{L}_{\text{rec}}$ not only for the single-person data but also for the data obtained by fusing the single-person data with InterHuman. This approach ensures that the CMDM model learns more efficient denoising capabilities from the single-person data training, as the interleaved data u inherently contains noise from the data processing. For the REC module, we use transformer.

The diffusion timesteps are set to 1,000, and the DDIM [35] sampling strategy is used during the inference stage. In the first stage, the INS module is trained, and the overall loss function is L_{ins}. In the second stage, the parameters of the INS module are frozen, and the INC module is trained, with the total loss being L. The hyperparameters are set as $\lambda_1 = 1$, $\lambda_2 = 0.1$, $\lambda_3 = 1$, and $\lambda_4 = 0.5$. Both training stages use a learning rate of $1e-5$, with 2,000 training epochs for each stage. We use the frozen CLIP-ViT-L-14 model as the text encoder. Our model trains in 31 hours on a single H100 GPU with a batch size of 256 and 44 GB of memory. For the data ratio issue between the two datasets, please refer to the appendix for details.

For the function $U(\cdot)$, we adopt the data representation method from HumanML3D, where each frame of motion is represented as a 263-dimensional feature vector. All single-person and dual-person data are uniformly normalized for initial positions, orientations, and skeletal frameworks, and the data are processed into a recursive form.

To make a fair comparison, we benchmarked FreeMotion [6], which is also dedicated to tackling the issues of number-free motion generation and human-to-human interaction. FreeMotion trains its number-free motion generation model by using InterHuman dataset. As shown in Table 1, we adopted the same training data format as FreeMotion and conducted tests on the relevant metrics. At the same time, since the FreeMotion framework cannot train on single-human datasets, we used $U(\cdot)$ function at the input stage, consistent with InterSyn, for processing and training.

5.2. Quantitative Results

Comparisons with state-of-the-art models: We first compare the interaction performances in InterHuman test set. As in Table 1, our method achieves state-of-the-art performance across all metrics, demonstrating superior text-to-motion synthesis capabilities. With R Precision scores of 0.335, 0.479, and 0.584 for top-1, top-2, and top-3 respectively, and a low FID of 6.332, InterSyn excels in translating textual semantics into physically plausible interactions. The INS module effectively resolves ambiguities in action descriptions by jointly modeling individual and interactive motions, ensuring precise alignment between text prompts and spatiotemporal dynamics. Notably, our approach outperforms InterGen by achieving a 25.3% improvement in Top-1 R Precision and a 50.6% reduction in FID, highlighting its superior capability in capturing nuanced coordination. Furthermore, our framework achieves near-real diversity (7.763 vs. 7.748) and optimal MModality (1.601), confirming its ability to generate varied yet natural interactions. The REC module enforces interaction aware constraints while preserving motion variety, and compared to single-person methods such as MDM, InterSyn reduces MM Dist by 37.0%, underscoring its effectiveness in synthesizing geometrically coherent multi-actor motions. Overall, these results validate that interleaved learning of solo and interactive dynamics—coupled with explicit coordination refinement—is critical for generating socially synchronized motions that accurately reflect textual intent while maintaining human-like naturalness.

Evaluation in Dynamic Environments: Human motion synthesis must address the inherent complexity of real-world scenarios where solo actions and social interactions coexist and dynamically interleave. To rigorously evaluate this capability, we construct an interleaved motion benchmark unifying HumanML3D and InterHuman test sets, reflecting natural transitions between individual and interactive behaviors. To avoid alignment issues from fused datasets, we use a hybrid strategy: single-person and multi-person evaluators are applied to their respective segments, and for mixed test sequences, we compute a frame-wise weighted average of the matching scores. Temporal masking is ap-

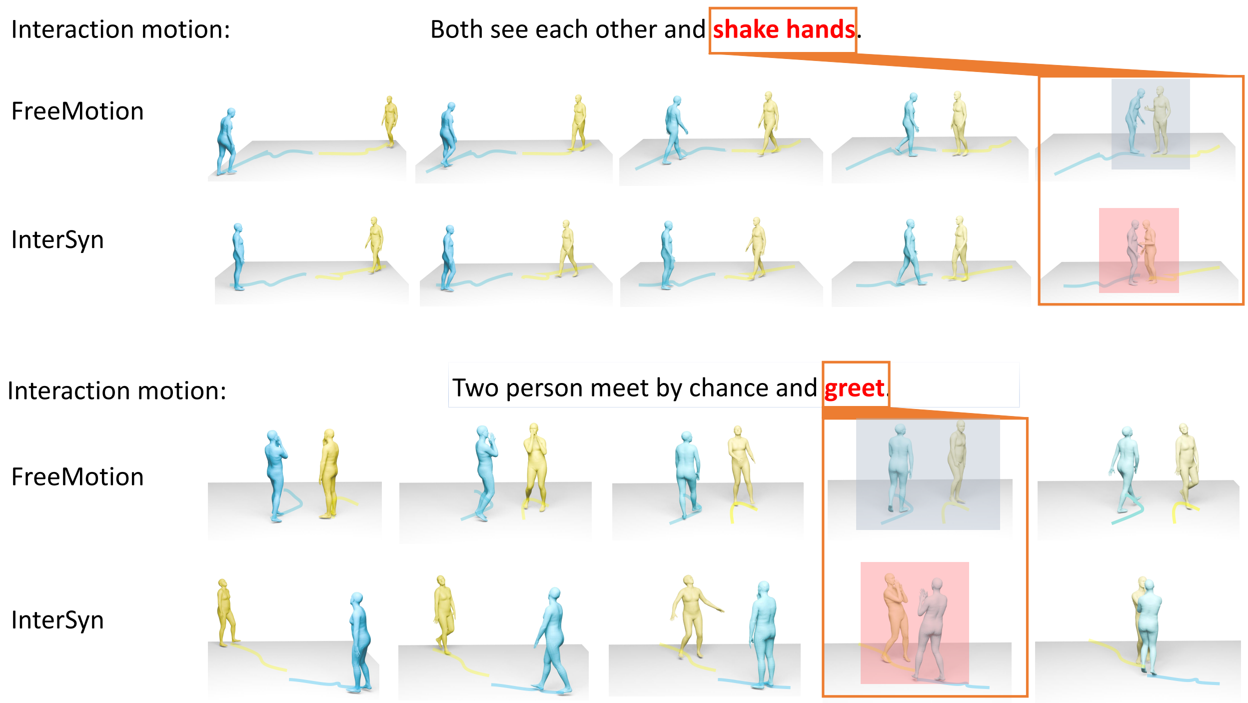

Figure 3. Example of generated interaction motions. By learning from interleaved motions that closely mirror real human dynamics, our method produces more vivid and natural interactions. Significant interaction regions are highlighted in blue for [6] and in red for our approach.

Method	R Precision ↑ Top 1	Top 2	Top 3	FID ↓	MM Dist ↓	Diversity →	MModality ↑
Real	$0.452^{\pm.008}$	$0.610^{\pm.009}$	$0.710^{\pm.008}$	$0.273^{\pm.007}$	$3.755^{\pm.008}$	$7.748^{\pm.064}$	-
TEMOS [30]	$0.224^{\pm.010}$	$0.316^{\pm.013}$	$0.450^{\pm.018}$	$17.375^{\pm.043}$	$5.342^{\pm.015}$	$6.939^{\pm.071}$	$0.535^{\pm.014}$
T2M [10]	$0.238^{\pm.012}$	$0.325^{\pm.010}$	$0.464^{\pm.014}$	$13.769^{\pm.072}$	$4.731^{\pm.013}$	$7.046^{\pm.022}$	$1.387^{\pm.076}$
MDM [37]	$0.153^{\pm.012}$	$0.260^{\pm.009}$	$0.339^{\pm.012}$	$9.167^{\pm.056}$	$6.125^{\pm.018}$	$7.602^{\pm.045}$	$\mathbf{2.355}^{\pm\mathbf{.080}}$
ComMDM [34]	$0.223^{\pm.009}$	$0.334^{\pm.008}$	$0.466^{\pm.010}$	$7.069^{\pm.054}$	$5.212^{\pm.021}$	$7.244^{\pm.038}$	$1.822^{\pm.052}$
InterGen [20]	$0.264^{\pm.006}$	$0.392^{\pm.005}$	$0.472^{\pm.005}$	$13.404^{\pm.200}$	$3.882^{\pm.001}$	$7.77^{\pm.030}$	$1.451^{\pm.034}$
FreeMotion [6]	$0.326^{\pm.003}$	$0.462^{\pm.006}$	$0.544^{\pm.006}$	$6.740^{\pm.130}$	$\mathbf{3.848}^{\pm\mathbf{.002}}$	$7.828^{\pm.130}$	$1.226^{\pm.046}$
Ours	$\mathbf{0.335}^{\pm\mathbf{.008}}$	$\mathbf{0.479}^{\pm\mathbf{.006}}$	$\mathbf{0.584}^{\pm\mathbf{.008}}$	$\mathbf{6.332}^{\pm\mathbf{.210}}$	$3.856^{\pm.005}$	$\mathbf{7.763}^{\pm\mathbf{.052}}$	$1.601^{\pm.047}$

Table 1. Quantitative comparisons on the InterHuman test set. Our method significantly outperforms competing approaches across various metrics.

plied to boundary frames to prevent bias from transitions. As shown in Table 2, our method achieves a 6.4% improvement in Top 1 R Precision and 42.1% reduction in FID over FreeMotion. These gains stem from two points: (1). InterSyn jointly encodes solo and interactive motions within a unified latent space, enabling seamless transitions between action types. This eliminates the modality bias observed in FreeMotion, which processes individual and group motions through separate pathways, leading to inconsistent semantics during dynamic interleaving (e.g., abrupt posture breaks when switching from 'walking alone' to 'shaking hands'). (2). Our method reduces MM Dist by 5.1% through adaptive kinematic constraints that preserve spatial relationships during motion transitions. For instance, it maintains appropriate

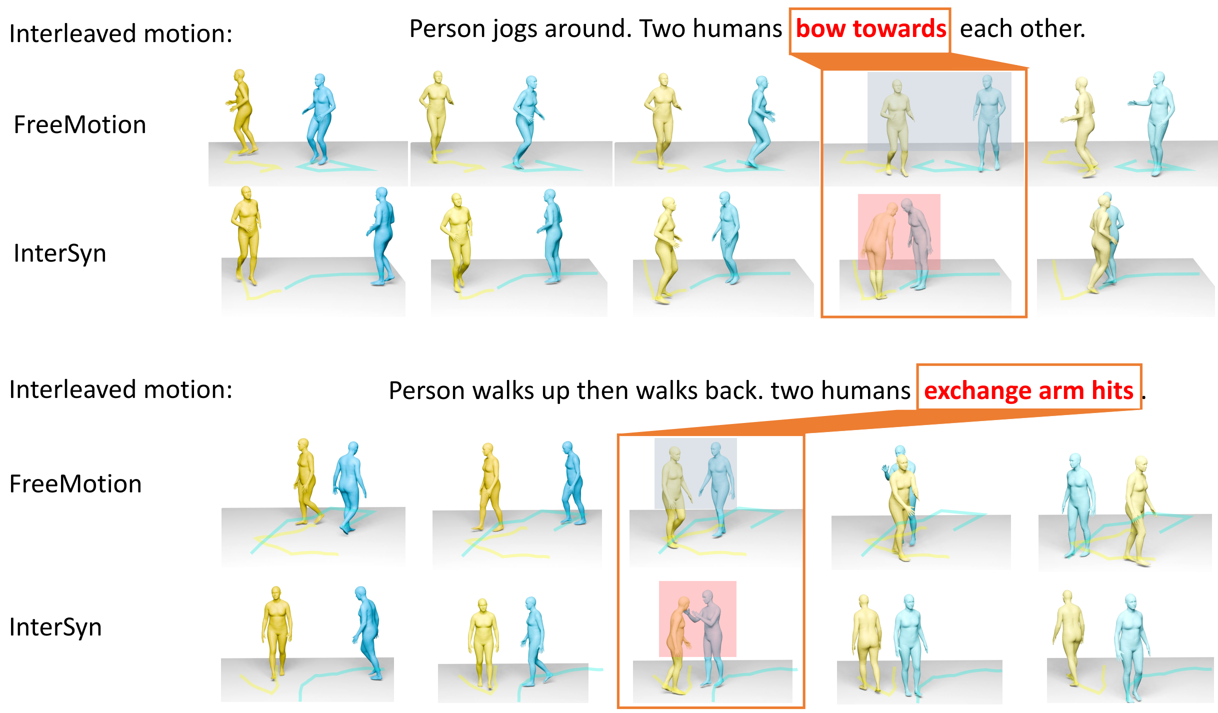

Figure 4. Example of generated interleaved motions. Our method produces realistic and nuanced sequences that closely mimic real-world scenarios by seamlessly integrating solo actions and interactions. The visualized motions highlight improved fidelity and smooth transitions compared to prior work. Significant interaction regions are marked in blue for [6] and in red for our approach.

interpersonal distances when synthesizing sequences like 'sit down $\rightarrow$ converse $\rightarrow$ stand up', where FreeMotion often introduces implausible body collisions.

The method also achieves a 13.3% higher MModality, demonstrating its capacity to generate diverse and text-consistent motions. This is exemplified in scenarios like 'passing an object while walking', where our framework produces varied arm trajectories and walking speeds while ensuring object transfer completion—a nuance that FreeMotion fails to capture due to its decoupled motion modeling. Notably, our approach attains this without sacrificing diversity, proving that joint modeling enhances rather than restricts motion variety. In all, the results validate that interleaved learning of solo and interactive dynamics is essential for synthesizing motions in real-world environments, where actions fluidly blend individual intent and social coordination.

Ablation on Data orgnization of INC: As shown in Table 3, after setting the fixed frame number to 196 for fair comparison, we conducted experiments using different data ratios and sequencing schemes for single- and dual-human interaction data (e.g., s-i-s-i" represents a sequence where the data alternates between single-person and interactive data segments). From the table, we can observe that incorporating more complex and frequently switching data patterns within a limited time window leads to a decline in the model's learning capacity, ultimately resulting in poorer performance. In contrast, the simpler method (s-i-s") not only yields better motion expressiveness but also proves to be more robust and scalable for long-duration interaction modeling tasks.

Ablation for timestep settings in INC: Based on the ablation study results in Table 3, we can conclude that the configuration of $t_s = 0, random, t_i$ performs the best across multiple metrics. It achieves a better balance in Top 1 precision, Diversity, and MModality compared to other settings like $t_i = 0, t_s = 60$ or $t_s = 0, t_i = 60$. Specifically, in terms of Diversity and MModality, $t_s = 0, random, t_i$ shows higher values, indicating that it generates more diverse and multi-modal action sequences. Therefore, $t_s = 0, random, t_i$ is considered the optimal configuration, as it adapts better to different time settings and complex interaction tasks, delivering improved model performance and generation diversity.

Ablations on loss functions: We systematically evaluate the loss function of InterSyn as in Table 3. In REC, removing the coordination module (w/o coordinator) severely degrades performance, with MM Dist increasing by 57.6% and FID

Method	R Precision ↑ Top 1	Top 2	Top 3	FID ↓	MM Dist ↓	Diversity →	MModality ↑
FreeMotion [6]	$0.280^{\pm.011}$	$0.381^{\pm.008}$	$0.508^{\pm.010}$	$0.721^{\pm.016}$	$3.905^{\pm.002}$	$9.244^{\pm.091}$	$1.397^{\pm.066}$
ours	$\mathbf{0.298}^{\pm.010}$	$\mathbf{0.410}^{\pm.007}$	$\mathbf{0.508}^{\pm.008}$	$\mathbf{0.417}^{\pm.012}$	$\mathbf{3.707}^{\pm.003}$	$\mathbf{9.322}^{\pm.030}$	$\mathbf{1.583}^{\pm.029}$

Table 2. Quantitative comparisons on our interleaved motion dataset, constructed by unifying the HumanML3D and InterHuman test sets to better reflect real-world interactions. Our method demonstrates significant improvements across all evaluated metrics.

Method	R Precision ↑ Top 1	Top 2	Top 3	FID ↓	MM Dist ↓	Diversity →	MModality ↑
$s-i-s$	$0.298^{\pm.010}$	$0.410^{\pm.007}$	$0.508^{\pm.008}$	$0.417^{\pm.012}$	$3.707^{\pm.003}$	$9.322^{\pm.030}$	$1.583^{\pm.029}$
$s-i-s-i$	$0.242^{\pm.008}$	$0.301^{\pm.010}$	$0.477^{\pm.004}$	$0.469^{\pm.024}$	$3.958^{\pm.011}$	$9.870^{\pm.056}$	$1.637^{\pm.071}$
$i-s-i-s$	$0.238^{\pm.011}$	$0.319^{\pm.005}$	$0.453^{\pm.003}$	$0.467^{\pm.041}$	$3.880^{\pm.020}$	$9.634^{\pm.038}$	$1.600^{\pm.056}$
$s-i-s-i-s$	$0.115^{\pm.006}$	$0.189^{\pm.003}$	$0.267^{\pm.003}$	$0.638^{\pm.055}$	$4.436^{\pm.010}$	$11.493^{\pm.118}$	$1.837^{\pm.109}$
$i-s-i-s-i$	$0.116^{\pm.010}$	$0.204^{\pm.005}$	$0.318^{\pm.007}$	$0.661^{\pm.024}$	$4.519^{\pm.015}$	$12.945^{\pm.064}$	$1.941^{\pm.050}$
$t_i=0, t_s=60$	$0.301^{\pm.003}$	$0.411^{\pm.009}$	$0.519^{\pm.006}$	$0.422^{\pm.009}$	$3.695^{\pm.007}$	$9.036^{\pm.082}$	$1.416^{\pm.052}$
$t_i=0, random\, t_s$	$0.283^{\pm.012}$	$0.398^{\pm.007}$	$0.503^{\pm.004}$	$0.418^{\pm.020}$	$3.710^{\pm.004}$	$9.271^{\pm.049}$	$1.479^{\pm.037}$
$t_s=0, t_i=60$	$0.310^{\pm.005}$	$0.426^{\pm.005}$	$0.521^{\pm.011}$	$0.423^{\pm.012}$	$3.702^{\pm.005}$	$9.102^{\pm.074}$	$1.399^{\pm.058}$
$t_s=0, random\, t_i$	$0.298^{\pm.010}$	$0.410^{\pm.007}$	$0.508^{\pm.008}$	$0.417^{\pm.012}$	$3.707^{\pm.003}$	$9.322^{\pm.030}$	$1.583^{\pm.029}$
w/o coordinator	$0.103^{\pm.009}$	$0.178^{\pm.008}$	$0.335^{\pm.021}$	$0.847^{\pm.104}$	$5.842^{\pm.018}$	$11.270^{\pm.142}$	$1.837^{\pm.044}$
w/o $\mathcal{L}_{\text{rela}}$	$0.283^{\pm.023}$	$0.393^{\pm.012}$	$0.511^{\pm.015}$	$0.537^{\pm.021}$	$3.838^{\pm.007}$	$9.370^{\pm.072}$	$1.543^{\pm.059}$
w/o $\mathcal{L}_{\text{smooth}}$	$0.295^{\pm.020}$	$0.399^{\pm.008}$	$0.509^{\pm.013}$	$0.431^{\pm.031}$	$3.712^{\pm.013}$	$9.312^{\pm.053}$	$1.544^{\pm.037}$

Table 3. Ablation study evaluating the impact of data organization for interleaved motion, INC timestep settings, and loss function designs.

by 103.1%. This demonstrates the module's critical role in improving the interaction motions. Moreover, disabling $\mathcal{L}_{\text{rela}}$ reduces Top 1 R Precision by 5.0%, revealing its specific function in aligning mutual dynamics of multi-character interactions (e.g., maintaining proper arm positions during "hugging" actions). While ablation of $\mathcal{L}_{\text{smooth}}$ shows minimal metric impact (+3.4% FID), qualitative analysis reveals jittery transitions in motion sequences (e.g., abrupt footplant shifts during gait changes), confirming its role in ensuring natural motion flow.

5.3. Qualitative Evaluation

Visualization of Interaction Motions: Our proposed method demonstrates significant improvements over the baseline FreeMotion approach, as evidenced by the visual results. The visualizations clearly highlight the enhanced performance in generating more natural interactions and better text matching. As shown in Fig. 3 and Fig. 4, the free motion model still exhibits significant deficiencies in generating interactive motions for two-person actions, and its framework does not generalize well to dual-agent motion interaction tasks. In contrast, InterSyn generates motions with higher text matching accuracy, whether using single-person or two-person text inputs. Furthermore, the results demonstrate that InterSyn achieves controllability over the generated outcomes through t_i and t_s, allowing for the customization of motion sequence details and the generation of corresponding actions.

Visualization on Dynamic motions: Using the design structure of the coordinator, the model is able to extend the generation of two-person motions to multiple agents during the generation phase. The final results and discussion can be found in the appendix.

6. Conclusion

In this paper, we introduced Interleaved Learning for Motion Synthesis (InterSyn), a novel framework that unifies the generation of dynamic interaction motions by seamlessly integrating single-person and multi-person dynamics. Our approach leverages the Interleaved Interaction Synthesis (INS) module to jointly model motion, capturing the fluid transitions between solo and interactive behaviors, and the Relative Coordination Refinement (REC) module to fine-tune interdependent interactions, ensuring semantic coherence and social synchronization. Extensive experiments validate that InterSyn produces more realistic and nuanced interaction motions than state-of-the-art methods.

Acknowledgement. This work was partly supported by the Shenzhen Key Laboratory of Next Generation Interactive Media Innovative Technology (Project No. ZDSYS20210623092001004).

References

[1] Albert Bandura and Richard H Walters. *Social learning theory*. Prentice hall Englewood Cliffs, NJ, 1977. 1

[2] Erik W Carter, Lisa S Cushing, Nitasha M Clark, and Craig H Kennedy. Effects of peer support interventions on students' access to the general curriculum and social interactions. *Research and practice for persons with severe disabilities*, 30 (1):15–25, 2005. 1

[3] Minghan Chen, Guikun Chen, Wenguan Wang, and Yi Yang. Hydra-sgg: Hybrid relation assignment for one-stage scene graph generation. *arXiv preprint arXiv:2409.10262*, 2024. 1

[4] Sheldon Cohen. Social relationships and health. *American psychologist*, 59(8):676, 2004. 1

[5] Yrjö Engeström. New forms of learning in co-configuration work. *Journal of Workplace learning*, 16(1/2):11–21, 2004. 1

[6] Ke Fan, Junshu Tang, Weijian Cao, Ran Yi, Moran Li, Jingyu Gong, Jiangning Zhang, Yabiao Wang, Chengjie Wang, and Lizhuang Ma. Freemotion: A unified framework for number-free text-to-motion synthesis. In *European Conference on Computer Vision*, pages 93–109. Springer, 2024. 1, 2, 5, 6, 7, 8

[7] Vaike Fors, Åsa Bäckström, and Sarah Pink. Multisensory emplaced learning: Resituating situated learning in a moving world. *Mind, Culture, and Activity*, 20(2):170–183, 2013. 1

[8] Mitch J Fryling, Cristin Johnston, and Linda J Hayes. Understanding observational learning: An interbehavioral approach. *The Analysis of verbal behavior*, 27:191–203, 2011. 1

[9] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2motion: Conditioned generation of 3d human motions. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2021–2029, 2020. 5

[10] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5152–5161, 2022. 1, 2, 5, 6

[11] Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. Momask: Generative masked modeling of 3d human motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1900–1910, 2024. 1, 2

[12] Jan Herrington, Ron Oliver, A Herrington, and Heather Sparrow. Towards a new tradition of online instruction: Using situated learning to design web-based units. In *ASCILITE 2000 Conference*, 2000. 1

[13] Hanzhuo Huang, Yuan Liu, Ge Zheng, Jiepeng Wang, Zhiyang Dou, and Sibei Yang. Mvtokenflow: High-quality 4d content generation using multiview token flow. *arXiv preprint arXiv:2502.11697*, 2025. 1

[14] Faraja Teddy Igira and Judith Gregory. Cultural historical activity theory. *Handbook of research on contemporary theoretical models in information systems*, pages 434–454, 2009. 1

[15] Ronald Inglehart and Daphna Oyserman. Individualism, autonomy, self-expression. *The human development sequence. In H. Vinken, J. Soeters & P. Ester (Eds.), Comparing cultures: Dimensions of culture in a comparative perspective*, pages 74–96, 2004. 1

[16] Jaana Isohätälä, Piia Näykki, and Sanna Järvelä. Cognitive and socio-emotional interaction in collaborative learning: Exploring fluctuations in students' participation. *Scandinavian Journal of Educational Research*, 64(6):831–851, 2020.

[17] Carroll E Izard. Four systems for emotion activation: cognitive and noncognitive processes. *Psychological review*, 100 (1):68, 1993. 1

[18] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[19] Richard M Lerner. *Concepts and theories of human development*. Routledge, 2018. 1

[20] Han Liang, Wenqian Zhang, Wenxuan Li, Jingyi Yu, and Lan Xu. Intergen: Diffusion-based multi-human motion generation under complex interactions. *International Journal of Computer Vision*, 132(9):3463–3483, 2024. 1, 2, 4, 5, 6

[21] Yuanzhi Liang, Yalong Bai, Wei Zhang, Xueming Qian, Li Zhu, and Tao Mei. Vrr-vg: Refocusing visually-relevant relationships. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10403–10412, 2019. 1

[22] Yuanzhi Liang, Qianyu Feng, Linchao Zhu, Li Hu, Pan Pan, and Yi Yang. Seeg: Semantic energized co-speech gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10473–10482, 2022. 1

[23] Yuanzhi Liang, Linchao Zhu, Xiaohan Wang, and Yi Yang. A simple episodic linear probe improves visual recognition in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9559–9569, 2022. 1

[24] Yuanzhi Liang, Linchao Zhu, Xiaohan Wang, and Yi Yang. Icocap: Improving video captioning by compounding images. *IEEE transactions on multimedia*, 26:4389–4400, 2023. 1

[25] Mehdi-Antoine Mahfoudi, Alexandre Meyer, Thibaut Gaudin, Axel Buendia, and Saida Bouakaz. Emotion expression in human body posture and movement: A survey on intelligible motion factors, quantification and validation. *IEEE Transactions on Affective Computing*, 14(4):2697–2721, 2022. 1

[26] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5442–5451, 2019. 5

[27] Felix B Mueller, Julian Tanke, and Juergen Gall. Massively multi-person 3d human motion forecasting with scene context. *arXiv preprint arXiv:2409.12189*, 2024. 2

[28] Doris Nussbaumer. An overview of cultural historical activity theory (chat) use in classroom research 2000 to 2009. *Educational review*, 64(1):37–55, 2012. 1

[29] Geir Overskeid. Do we need the environment to explain operant behavior? *Frontiers in psychology*, 9:373, 2018. 1

[30] Mathis Petrovich, Michael J Black, and Gül Varol. Temos: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision*, pages 480–497. Springer, 2022. 6

[31] Mathis Petrovich, Michael J. Black, and Gül Varol. TEMOS: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision (ECCV)*, 2022. 1

[32] Troy D Sadler. Situated learning in science education: socio-scientific issues as contexts for practice. *Studies in science Education*, 45(1):1–42, 2009. 1

[33] Vassilis Sevdalis and Peter E Keller. Perceiving bodies in motion: expression intensity, empathy, and experience. *Experimental brain research*, 222:447–453, 2012. 1

[34] Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit H Bermano. Human motion diffusion as a generative prior. *arXiv preprint arXiv:2303.01418*, 2023. 6

[35] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 5

[36] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022. 1, 2, 3, 4

[37] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2023. 1, 3, 6

[38] Michael Tomasello, Malinda Carpenter, Josep Call, Tanya Behne, and Henrike Moll. Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and brain sciences*, 28(5):675–691, 2005. 1

[39] Bonnie L Utley. Effects of situated learning on knowledge gain of instructional strategies by students in a graduate level course. *Teacher education and special education*, 29(1):69–82, 2006. 1

[40] Jiepeng Wang, Hao Pan, Yang Liu, Xin Tong, Taku Komura, and Wenping Wang. Structre: Rewriting for structured shape modeling. *ACM Transactions on Graphics*, 44(3):1–20, 2025. 1

[41] Runqi Wang, Caoyuan Ma, GuoPeng Li, and Zheng Wang. Taat: Think and act from arbitrary texts in text2motion. *arXiv e-prints*, pages arXiv–2404, 2024. 2

[42] Dianbing Xi, Jiepeng Wang, Yuanzhi Liang, Xi Qiu, Yuchi Huo, Rui Wang, Chi Zhang, and Xuelong Li. Omnivdiff: Omni controllable video diffusion for generation and understanding. *arXiv preprint arXiv:2504.10825*, 2025. 1

[43] Yiming Xie, Varun Jampani, Lei Zhong, Deqing Sun, and Huaizu Jiang. Omnicontrol: Control any joint at any time for human motion generation. In *The Twelfth International Conference on Learning Representations*, 2024. 2

[44] Chi Zhang, Yuanzhi Liang, Xi Qiu, Fangqiu Yi, and Xuelong Li. Vast 1.0: A unified framework for controllable and consistent video generation. *arXiv preprint arXiv:2412.16677*, 2024. 1

[45] Lei Zhong, Yiming Xie, Varun Jampani, Deqing Sun, and Huaizu Jiang. Smoodi: Stylized motion diffusion model. In *European Conference on Computer Vision*, pages 405–421. Springer, 2024. 1, 2