SocialGen: Modeling Multi-Human Social Interaction with Language Models

Heng Yu* Juze Zhang* Changan Chen Tiange Xiang Yusu Fang
Juan Carlos Niebles Ehsan Adeli

Stanford University

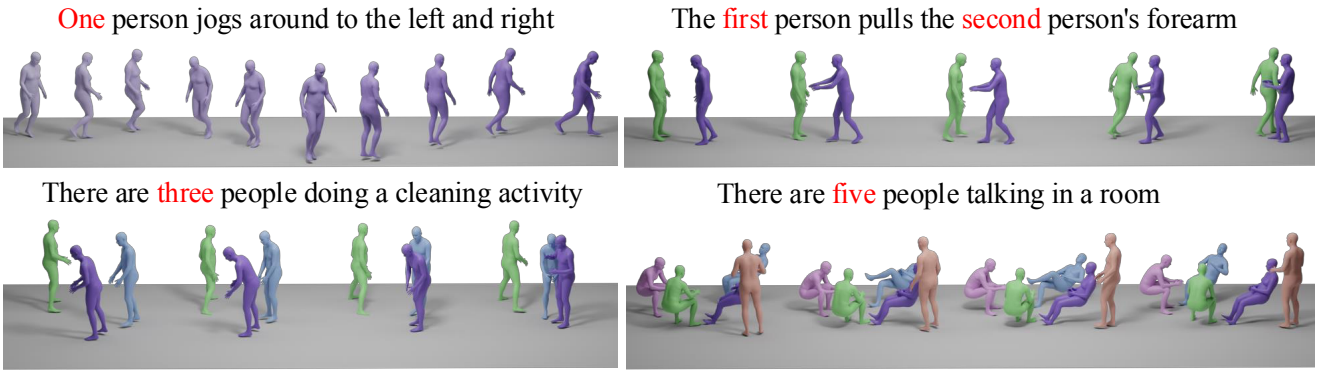

Figure 1. SocialGen generates social interaction motion for varying numbers of individuals while also supporting diverse motion-language tasks such as captioning, completion, and reaction generation.

Abstract

Human interactions in everyday life are inherently social, involving engagements with diverse individuals across various contexts. Modeling these social interactions is fundamental to a wide range of real-world applications. In this paper, we introduce SocialGen, the first unified motion-language model capable of modeling interaction behaviors among varying numbers of individuals, to address this crucial yet challenging problem. Unlike prior methods that are limited to two-person interactions, we propose a novel social motion representation that supports tokenizing the motions of an arbitrary number of individuals and aligning them with the language space. This alignment enables the model to leverage rich, pretrained linguistic knowledge to better understand and reason about human social behaviors. To tackle the challenges of data scarcity, we curate a comprehensive multi-human interaction dataset, ***SocialX****, enriched with textual annotations. Leveraging this dataset, we establish the first comprehensive benchmark for multi-human interaction tasks. Our method achieves state-of-the-art performance across motion-language tasks, setting a new standard for multi-human interaction modeling. Our dataset and source code will be made publicly available.*

*indicates equal contribution

1. Introduction

Modeling human behavior in social interactions are pivotal for transformative applications such as VR, gaming, healthcare, and robotics. In the entertainment field, realistic multi-human interactions enhance immersion. In healthcare, understanding patient-physician dynamics improves telemedicine and therapy. In robotics, socially aware interaction modeling enables seamless collaboration in human-centered settings. These applications highlight the significance of modeling multi-human social scenarios.

Recent years have witnessed remarkable progress in human motion understanding [5, 33, 82] and generation [15, 39, 62]. However, most existing work focuses on single-person [5, 33, 62] or dyadic (two-person) interactions [32, 38], often relying on hard-coded interaction models that cannot generalize to scenarios with arbitrary numbers of participants. Real-world interactions, by contrast, are dynamic and frequently involve multiple individuals with complex dependencies. Addressing these challenges necessitates a scalable approach capable of modeling social interactions for any number of humans, supporting both understanding and generation.

To overcome the limitations in more general interaction modeling, we propose a novel motion-language framework that generates motions for any number of individuals and enables various motion-related tasks. The key to our approach is a novel motion representation **XH3D**, specifically

designed to encode and decode complex multi-human interactions effectively. We tokenize motion, spatial, and text data using modality-specific tokenizers, creating a unified vocabulary that integrates text, individuals, and relative spatial configurations. These tokens are processed by a motion-language model, which outputs sequences of modality tokens. Our framework includes a pre-training stage to spatially and temporally capture human interactions, aligning modalities in the latent space, followed by post-training with diverse task instructions to enhance generalization. This unified approach enables motion generation for any number of individuals and supports diverse tasks like captioning, completion, and reaction generation.

While several motion datasets for multi-human social interactions exist [78, 79, 93], they lack language descriptions, limiting their utility for tasks involving language input or output. To address this, we introduce **SocialX**, a curated dataset enriched with detailed captions generated using GPT-4o. By employing a multi-step prompting scheme, we ensure accurate and faithful descriptions of multi-human social interactions, enabling benchmarking for generation and other motion-language tasks.

To validate our model, we first evaluate the effectiveness of our motion tokenizer on InterHuman and InterX datasets, demonstrating superior reconstruction and better capture of inter-human dynamics compared to existing representations. We further benchmark our model on SocialX, where it outperforms state-of-the-art methods across multiple metrics, generating smoother and more natural human motions. Additionally, our method supports various other motion-related tasks. In a nutshell:

- We propose a novel approach to modeling multi-human social interactions by leveraging language models while also supporting various downstream motion-related tasks.
- We introduce a novel motion representation for scenarios involving varying numbers of participants, enabling scalable and effective modeling of social interactions.
- We develop **SocialX**, a curated dataset paired with a benchmark, encompassing diverse scenarios with human pose data and language descriptions, setting a new standard for evaluating multi-human interaction models.

We believe that our framework and benchmark will provide a strong foundation for future research in multi-human interaction modeling.

2. Related Work

Human Motion Generation and Understanding: Generating and understanding human motion are key to human-centric AI. Recent works explore realistic and diverse motion generation using multimodal inputs, including textual descriptions [15, 22, 23, 39, 62, 74] actions [8, 21, 58], music or audio [5, 54, 76, 96, 99], incomplete motion [3, 7, 48, 51, 87, 94], and other control signals [18, 31, 59, 73, 80] Among these, text-to-motion generation is gaining attention for its intuitive and naturally use-friendly input. Diffusion-based models [7, 35, 50, 74, 81, 92, 97] and vector-quantized approaches [23, 24, 60, 77, 90] have demonstrated promising results in generating isolated human motion. However, they struggle to capture broader motion patterns, particularly in complex social interactions. Recent works [14, 19, 33, 41, 47, 95, 99] unify tasks or modalities, but lack multi-human handling. In contrast, our work delves into a broader set of scenarios, explicitly tackling social group motion generation.

Human-Human Interaction Modeling: Most motion generation methods focus on single humans, while interactions add complexity by requiring inter-individual and spatio-temporal modeling. Recent works on human-human interaction modeling mainly focuse on two-person scenarios, addressing (1) *reaction synthesis*, where a reactor responds to an actor's motion [9, 16, 43, 86], and (2) *interaction generation*, where all individuals' motions are generated simultaneously [4, 38, 69, 72]. Some studies extend to multi-person settings [13, 70], but they focus on motion generation and lack language-related comprehensive capabilities, limiting generalization to larger groups and open-domain prompts (see Tab. 5 in the Appendix).

Multimodal Language Models: Large language models (LLMs) such as T5 [65], Flan-T5 [10], and LLaMA [75], along with others [11, 12], have demonstrated exceptional capabilities in comprehension and generation tasks. Recently, researchers have extended LLMs to multimodal domains, integrating inputs from images [17, 29, 37], videos [42, 84], audio [17, 26], and other modalities [6, 27], paving the way for multimodal large language models (MLLMs) [83, 89]. For 3D motion, MotionGPT [33] frames single-human motions as a language modeling task, which was extended by [5, 47] to include more modalities.

Building on these advancements, our work introduces the first motion-language model tailored for multi-human social interaction modeling, enabling the generation and other tasks among multiple individuals.

Human Motion Datasets: Datasets are fundamental for advancing motion-related synthesis tasks. Single-human datasets [22, 40, 61, 63] have been instrumental in developing motion-related models but face challenges when generalizing to multi-human scenarios. While datasets like InterGen [38], InterX [85], and InterAct [30] provide two-person interactions with text descriptions, they are limited in addressing broader multi-human contexts.

Several multi-human motion datasets, including 4DAssociation [93], UMPM [78], 3DPW [79], MuPoTS-3D [52], ExPI [25], CMU-Panoptic [34], You2Me [53], BEDLAM [2], and HOI-M^3 [91], have been introduced. However, many of these datasets lack text descriptions, limiting their utility in multimodal understanding. Recent

datasets like LAION-Pose and WebVid-Motion [70] include pose/motion and text pairs extracted from in-the-wild images and videos [68], but their annotations and descriptions are constrained by automated pipelines.

In this paper, we introduce *SocialX*, a novel dataset with accurate SMPL-compatible [44] annotations and high-quality text descriptions. *SocialX* captures interactions involving varying numbers of individuals and establishes a benchmark for multi-human motion modeling, offering valuable resources for advancing the field.

3. Method

We introduce SocialGen, a unified motion-language framework that models interaction motions for varying group sizes by aligning motion and language spaces. As shown in Figure 2, SocialGen consists of two key components, similar to [33]: a motion tokenizer that discretizes motion data and a motion-aware language model that aligns motion and language tokens. This alignment enables motion modeling using pre-trained language models and textual descriptions.

3.1. Motion Representation

Existing Representations Existing motion representations primarily cater to either single-human or two-human.

- **H3D**: Designed for single-human motions, H3D [22] provides a canonical representation by transforming joint positions and velocities into the root frame. A pose is defined as $(r^a, r^x, r^z, r^y, \mathbf{j}^p, \mathbf{j}^v, \mathbf{j}^r, \mathbf{c}^f)$, where $r^a \in \mathbb{R}$ is the root angular velocity along the Y-axis, $(r^x, r^z) \in \mathbb{R}^2$ are root linear velocities on the XZ-plane, and $r^y \in \mathbb{R}$ is root height. The local joint positions, velocities, and rotations are given by $\mathbf{j}^p \in \mathbb{R}^{3(j-1)}$, $\mathbf{j}^v \in \mathbb{R}^{3j}$, and $\mathbf{j}^r \in \mathbb{R}^{6(j-1)}$, with j denoting the number of joints. Binary foot-ground contact features $\mathbf{c}^f \in \mathbb{R}^4$ are obtained by thresholding heel and toe joint velocities [22]. With a 6D rotation format [98] and the SMPL skeleton [44], the pose has 263 dimensions. While effective for single-human motion, H3D lacks the ability to model multi-human interactions.
- **IH**: The IH format [38] models multi-human interactions by encoding absolute global joint trajectories in the world frame. A pose is represented as $(\mathbf{j}_g^p, \mathbf{j}_g^v, \mathbf{j}^r, \mathbf{c}^f)$, where $\mathbf{j}_g^p \in \mathbb{R}^{3j}$ and $\mathbf{j}_g^v \in \mathbb{R}^{3j}$ are global joint positions and velocities, and $\mathbf{j}^r$ and $\mathbf{c}^f$ match the H3D definitions. With the SMPL skeleton, the pose dimension is 262. However, IH suffers from reconstruction artifacts during motion tokenization due to its high-dimensional global position.
- **IX**: Directly using SMPL [44] parameters, IX is a straightforward representation employed by datasets like InterX [85]. While conceptually simple, it faces similar challenges as IH, such as jittering and artifacts during tokenization, owing to its high-dimensional encoding.

Proposed XH3D Although the IH and IX formats are capable of representing multi-human scenarios, they struggle to achieve high-quality reconstruction during motion tokenization. This is because the token codebook typically consists of very few tokens (usuallly hundreds), while the root global position and 6D continuous SMPL representation are both high-dimensional. Consequently, IH and IX formats often suffer from jittering and artifacts after motion tokenization, as demonstrated in Section 5. To address this issue, we propose XH3D, an extension of the H3D format for multi-human scenarios. For each individual, XH3D adopts the same tuple $(r^a, r^x, r^z, r^y, \mathbf{j}^p, \mathbf{j}^v, \mathbf{j}^r, \mathbf{c}^f)$ used in H3D as the motion representation. Additionally, to capture relative spatial relationships between humans, we introduce a tuple (s^x, s^z, s^θ) representing the first-frame relative pose between two motions. Here, $(s^x, s^z) \in \mathbb{R}^2$ denotes the relative position on the root XZ-plane, and $s^\theta \in \mathbb{R}$ represents the relative orientation. For multiple humans, one is randomly chosen as the reference, and the relative pose of each remaining human is computed accordingly. This randomness also serves as data augmentation. We assert that XH3D is scalable to any number of individuals and achieves superior tokenization performance compared to existing formats, as demonstrated in Table 2.

3.2. Motion Tokenizer

To represent motion as discrete tokens, we employ a 3D human motion tokenizer based on the Vector Quantized Variational Autoencoder (VQVAE) framework [19, 33, 47, 77, 95, 99]. The motion tokenizer consists of an encoder $\mathcal{E}$, a decoder $\mathcal{D}$, and a learnable codebook $\mathcal{Z} = \{z_1, z_2, \ldots, z_K\}$ containing K discrete latent embedding vectors of dimension d. The encoder extracts latent motion features that are quantized into discrete tokens using the nearest entries from the codebook. The decoder reconstructs the motion sequence from these quantized tokens, enabling a compact and discrete representation suitable for downstream tasks. We independently tokenize each individual using VQVAE, following a similar approach to [33].

The training process optimizes a combination of reconstruction, embedding, and commitment losses to ensure high-quality motion representation following [33, 95]. The loss function is defined as:

$$\mathcal{L}_{\text{VQ}} = \underbrace{\|\hat{\mathbf{m}} - \mathbf{m}\|_1}_{\mathcal{L}_r} + \underbrace{\|\text{sg}[\mathbf{z}] - \mathbf{e}\|_2^2}_{\mathcal{L}_e} + \underbrace{\beta\|\mathbf{z} - \text{sg}[\mathbf{e}]\|_2^2}_{\mathcal{L}_c},$$

where $\mathbf{m}$ is the original motion sequence, $\hat{\mathbf{m}}$ is the reconstructed motion, $\mathbf{z}$ are the latent embeddings, $\mathbf{e}$ are the codebook entries, $\text{sg}[\cdot]$ denotes the stop-gradient operation, and β is a weighting factor for the commitment loss.

Additionally, we follow [90] to incorporate L1 smooth loss for reconstruction and velocity regularization to improve motion generation quality. Exponential moving average (EMA) is applied to update the codebook during train-

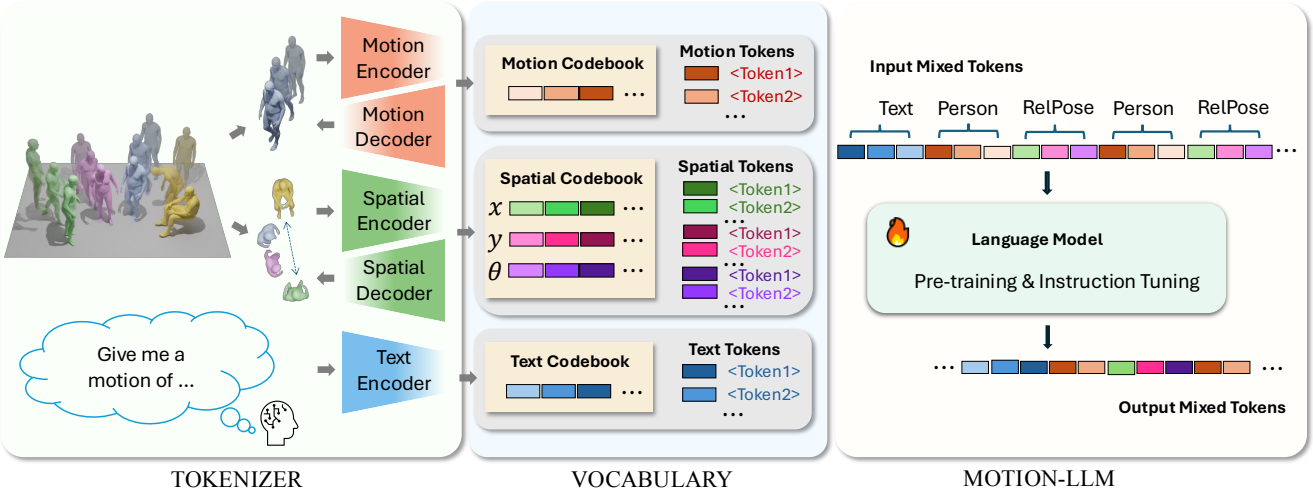

Figure 2. Overview of SocialGenFramework. SocialGenunifies motion generation, understanding, and other motion-language tasks for multi-human social interactions. The tokenizer encodes motion, spatial, and text features into tokens via respective codebooks, forming a unified vocabulary of motion, spatial, and text Tokens. These tokens, combined into mixed input tokens representing text, individuals, and relative spatial configurations, are processed by the Motion-LLM, which leverages pre-training and instruction tuning to generate mixed tokens, enabling versatile modeling of complex social interactions.

ing [66], ensuring stable and efficient codebook utilization. This approach enables efficient motion tokenization, facilitating the representation of motion as discrete tokens for seamless integration with language models and improved performance in motion-related tasks.

For the relative root pose tuple (s^x, s^z, s^θ), we apply uniform binning to discretize each component into K tokens by dividing its range—determined by dataset-wide extrema—into equal intervals. While simple and effective, alternative methods such as non-uniform bins, larger K, or vector quantization could improve performance. However, our focus is on the framework of separately tokenizing the relative pose to handle arbitrary numbers of individuals, while the optimal tokenization strategy remains an open problem beyond this study.

3.3. Motion-Language Model

To seamlessly integrate text and motion modeling, we employ a motion-aware language model built on a pre-trained large language model (LLM) such as T5 [65]. Our approach leverages the proposed motion tokenizer to map a motion sequence $m^{1:M}$ into a sequence of discrete motion tokens $z^{1:L}$, enabling the unified modeling of motion and text within a shared vocabulary [33, 36, 56, 65]. Each motion token z_i corresponds to an index in the motion codebook, including tokens for relative root poses, ensuring compatibility with the text token vocabulary. To achieve this, we extend the original text token vocabulary V_t with motion tokens $V_m = \{\langle \text{Motion}_i \rangle\}_{i=1}^K$ and relative root pose tokens $V_x = \{\langle \text{x}_i \rangle\}_{i=1}^K$, $V_z = \{\langle \text{z}_i \rangle\}_{i=1}^K$, and $V_\theta = \{\langle \theta_i \rangle\}_{i=1}^K$, forming a unified vocabulary $V = \{V_t, V_m, V_x, V_z, V_\theta\}$. This unified representation allows both input and output sequences to seamlessly incorporate natural language and motion tokens (including relative root pose tokens).

Special tokens, such as $\langle \text{Motion}_S \rangle$ and $\langle \text{Motion}_E \rangle$, mark the start and end of a motion sequence. For multiple humans, the tokens $\langle \text{x}_i \rangle$, $\langle \text{z}_i \rangle$, and $\langle \theta_i \rangle$ serve both as separators between different individuals and as representations of their relative poses. The general format for encoding motion tokens in multi-human scenarios is:

$$\langle \text{Motion}_S \rangle \ldots \langle \text{Motion}_{i-1} \rangle \langle \text{x}_i \rangle \langle \text{z}_i \rangle \langle \theta_i \rangle \langle \text{Motion}_i \rangle$$
$$\langle \text{x}_{i+1} \rangle \langle \text{z}_{i+1} \rangle \langle \theta_{i+1} \rangle \langle \text{Motion}_{i+1} \rangle \ldots \langle \text{Motion}_E \rangle$$

where each triplet $\langle \text{x}_i \rangle \langle \text{z}_i \rangle \langle \theta_i \rangle$ uniquely identifies a human and encodes their relative pose. This structure ensures scalability to any number of humans within a scene. These special tokens significantly enhance the model's ability to model multi-human interactions. Unless otherwise stated, motion tokens in the following text include both individual motion and relative pose information.

During training, the model learns to represent the text-motion relationship as an autoregressive sequence generation task. Given an input sequence $X_s = \{x_s^i\}_{i=1}^N$ consisting of text and motion tokens of length N and an output sequence $X_t = \{x_t^i\}_{i=1}^L$ of length L, the model predicts the probability of each token conditioned on the preceding tokens:

$$p_\theta(x_t|x_s) = \prod_i p_\theta(x_t^i|x_t^{<i}, x_s),$$

The training objective maximizes the log-likelihood of the

data distribution:

$$\mathcal{L}_{LM} = -\sum_{i=0}^{L_t-1} \log p_\theta(x_t^i | x_t^{<i}, x_s).$$

At inference time, the model generates text or motion tokens iteratively in an autoregressive manner. For instance, given a text prompt describing a motion, the model predicts a sequence of motion tokens until the $\langle \text{Motion}_E \rangle$ token is generated, marking the completion of the motion sequence. This iterative approach supports flexible motion generation with variable lengths and accommodates varying numbers of humans, adapting to the specific requirements of the input description. By incorporating motion into the language modeling framework, our motion-aware language model bridges the gap between text and motion, delivering robust performance on tasks requiring the joint understanding and generation of text and motion data, particularly in scenarios involving multi-human interaction behaviors.

3.4. Training Paradigm

As shown in Figure 2, our training process consists of three stages: (1) Motion Tokenizer Training, which learns a motion codebook to represent motion as discrete tokens; (2) Motion-Language Alignment Pre-training, which aligns motion and text modalities using a unified vocabulary and both supervised and unsupervised objectives; and (3) Instruction Tuning, which fine-tunes the model with prompt-based instructions to enhance its performance on diverse motion-relevant tasks.

Motion Tokenizer Training We first train the motion tokenizer using the objective defined by $\mathcal{L}_{\text{VQ}}$. This process maps any human motion sequence $m^{1:M}$ into a sequence of discrete motion tokens $z^{1:L}$, enabling seamless integration with text within the unified model. Once trained, the motion tokenizer remains fixed throughout subsequent stages of the pipeline to ensure stability and consistency. For the relative pose tokenizer, uniform binning is employed, eliminating the need for additional training.

Motion-language Alignment Pre-training The T5 model [65], pre-trained and fine-tuned [10] on natural language datasets, is further pre-trained on our SocialX dataset, which combines text and motion data involving various numbers of people. This pre-training occurs in both unsupervised and supervised settings. To improve the model's generalization ability across diverse tasks [12, 33, 56, 64, 65], we adopt an unsupervised learning objective [33, 65]. Specifically, 15% of the tokens in the input sequence X_s are randomly replaced with a special sentinel token. The target sequence is then constructed by extracting the replaced token spans, delimited by sentinel tokens, with an additional sentinel token marking the end of the target sequence. Further details can be found in [33].

In the supervised setting, the model is trained on paired text-motion datasets through specifically designed tasks. The first task, motion-language alignment, involves generating paired data where the input is either a motion sequence or a text description, and the output is its corresponding counterpart. The second task, motion forecasting, trains the model to predict future motion sequences based on a given initial motion. The third task, reaction synthesis, involves masking the motion of one individual in a multi-human scenario, with the model tasked to generate the missing motion.

Through these unsupervised and supervised pre-training stages, the model learns to generate and understand coherent text and motion representations, effectively bridging the gap between these modalities and achieving robust performance on motion-related tasks.

Post-training with Instruction Tuning To enhance the generalization and instruction-following capabilities of our motion-language model, we construct a diverse set of tasks focusing on motion captions, human count, frame count, and motion length, leveraging our SocialX dataset. Specifically, we define 25 core motion-related tasks, including motion completion, motion generation from text, motion captioning, motion prediction, and human number estimation from motion data. For each task, we create a variety of instruction templates using GPT-4 [1], resulting in a dataset containing thousands of unique task instructions. Additional examples of these templates can be found in the Supp.

4. Dataset

To create the benchmark for multi-person interactions, we enriched existing motion datasets with detailed text descriptions. This led to **SocialX**, the first dataset capturing diverse social interactions across varying group sizes and scenarios, including home, workplace, and recreational settings.

4.1. Datset Sources

Our **SocialX** dataset integrates and extends multiple existing datasets, including HumanML3D [22], InterHuman [38], InterX [85], and HOI-M^3 [91], CMU Panoptic [34], and MuPoTS-3D [52], as summarized in Table 8 in the Appendix. To further enrich SocialX, we synthesize multi-human interactions by combining motions from single-person [22] and two-person [38, 85] datasets while ensuring no physical conflicts. Although these synthesized interactions may not always form meaningful full-group scenarios, they help the model learn participant counts and sub-group interactions. SocialX also provides detailed text descriptions and a unified pose representation, making it the first large-scale dataset for multi-human interaction modeling. See the Appendix for details.

4.2. Caption Generation

For datasets lacking text descriptions, we use GPT-4o-mini [55] to generate content descriptions from videos. Frames are sampled at 1 fps, and videos are segmented into 20-second clips. When multi-view images (e.g., in HOI-M3) are available, they are utilized to enhance caption accuracy and reduce occlusion effects. Descriptions are generated at two levels, with manual filtering applied to ensure clean and accurate text. See the Appendix for details.

Scene Description We prompt GPT-4o-mini to generate detailed descriptions of human motion and interactions in the video clips. While the focus is on motion, scene-level descriptions may also include human-object interactions, appearances, and other contextual details.

Human Interaction Description To create captions focused purely on human interactions, we refine the scene-level descriptions using GPT-4o-mini. This process removes unrelated information, ensuring the captions concentrate solely on human motion and interactions. The result is a detailed, behavior-level description.

4.3. Motion Data Postprocessing

To ensure consistency across all datasets, we preprocess the motion data into a unified format compatible with our framework. For datasets that do not provide SMPL parameters, we first estimate these parameters by fitting the SMPL model [44] to the motion data. This process involves optimizing the pose and shape parameters to best align the motion capture data with the SMPL representation.

Once the SMPL parameters are obtained, we process the motion of each individual following the methodology introduced in HumanML3D [22]. Specifically, we extract individual motions and normalize them to a standard reference. Additionally, as discussed in Section 3.1, we tokenize the relative positions between individuals to encode interperson spatial relationships effectively. This step is crucial for capturing social interactions and enables our model to generalize across multi-human scenarios.

We utilize the validation and test sets from HumanML3D [22], InterHuman [38], InterX [85], and HOI-M^3 [91]. The remaining datasets, including those with limited multi-human interaction scenarios such as MuPoTS-3D [52] and the Haggling sequences from CMU Panoptic [34], are incorporated into the training set. To ensure consistency across multi-human datasets, all sequences are downsampled to 20 fps, and longer sequences are segmented into shorter clips, following the captioning preprocessing strategy described earlier. By integrating data from diverse sources, our training set captures a broad range of motion scenarios, enhancing the robustness and generalization capability of our model. By utilizing this pipeline, all motion data are transformed into our standardized format, allowing for seamless integration and consistent processing in our framework. More details are in the Appendix.

5. Experiment

5.1. Evaluation Metrics

To evaluate our model, we employ metrics assessing motion quality, diversity, multimodality, and text-motion alignment. Briefly, these include Frechet Inception Distance (FID) for motion quality, Diversity (DIV) and Multimodality (MM) for variation and distinctiveness, and R-Precision and Multi-Modal Distance (MM Dist) for text-motion alignment. Detailed descriptions of these metrics are provided in the Appendix. To evaluate the reconstruction error of VQ-VAE using different motion representations, we report the mean per joint position error (MPJPE) and the Procrustes Analysis MPJPE (PA-MPJPE) [20] for global/local errors in millimeters, the mean per joint acceleration error (Accel) for temporal quality as previous work [8, 46, 88]. Additional implementation details can be found in the Appendix.

5.2. Tokenization Evaluation

To assess the effectiveness of our proposed pose representation, we evaluate IH, IX, and our XH3D pose representations on the InterHuman and InterX datasets independently as shown in Table 2. The results demonstrate that our XH3D representation achieves superior performance in terms of joint position accuracy, measured by MPJPE and PA-MPJPE. Importantly, it significantly outperforms other representations in temporal quality, as reflected by the Accel metric, which is crucial for ensuring motion smoothness.

5.3. Quantitative Comparison

Table 1 presents a quantitative comparison of our method against previous approaches. Existing evaluation models, such as those in [32, 33], are trained on datasets with either single-person or two-person motions using different motion representations (e.g., IH or IX). For two-person interactions, we cannot directly use evaluation models like [32] due to representation differences. We adapted these methods for multi-person settings, with details and results provided in the Appendix. For single-person motions, our model performs similarly to MotionGPT, though single-person generation is not the focus of this paper.

Since no existing evaluation model can assess text-to-motion generation for an arbitrary number of individuals, we train an evaluation model on SocialX following the approach in [22]. To ensure compatibility with varying numbers of individuals, we modify the input and output dimensions of [22] to support up to N individuals (setting $N = 5$ in our experiments) while keeping all other configurations unchanged. Our evaluation model is designed to assess motion generation across different numbers of individuals, providing a more comprehensive evaluation of a model's ability

Dataset	Method	R Precision↑			FID↓	MM Dist↓	Diversity→	MModality↑
		Top 1	Top 2	Top 3				
SocialX	Ground Truth	$0.542^{\pm.002}$	$0.752^{\pm.002}$	$0.853^{\pm.002}$	$0.0002^{\pm.0001}$	$3.676^{\pm.006}$	$14.334^{\pm.114}$	-
	MDM [74]	$0.281^{\pm.003}$	$0.563^{\pm.003}$	$0.672^{\pm.003}$	$4.260^{\pm.006}$	-	-	-
	T2M [22]	$\mathbf{0.409}^{\pm.003}$	$\underline{0.599}^{\pm.004}$	$\underline{0.707}^{\pm.002}$	$10.694^{\pm.092}$	$\mathbf{5.081}^{\pm.010}$	$13.647^{\pm.121}$	$0.824^{\pm.047}$
	TM2T [23]	$0.390^{\pm.003}$	$0.584^{\pm.003}$	$0.696^{\pm.002}$	$2.349^{\pm.007}$	$5.479^{\pm.013}$	$\underline{14.109}^{\pm.130}$	$2.873^{\pm.073}$.
	T2M-GPT [90]	$0.403^{\pm.003}$	$0.588^{\pm.003}$	$0.706^{\pm.004}$	$10.163^{\pm.007}$	$\underline{5.097}^{\pm.005}$	$13.359^{\pm.121}$	$0.9186^{\pm.051}$.
	Ours - Pre-training	$0.396^{\pm.004}$	$0.593^{\pm.002}$	$0.699^{\pm.003}$	$\underline{1.795}^{\pm.005}$	$5.662^{\pm.012}$	$13.898^{\pm.139}$	$\underline{3.028}^{\pm.077}$
	Ours - Instruction Tuning	$\underline{0.403}^{\pm.003}$	$\mathbf{0.601}^{\pm.002}$	$\mathbf{0.717}^{\pm.002}$	$\mathbf{1.403}^{\pm.005}$	$5.389^{\pm.011}$	$\mathbf{14.213}^{\pm.147}$	$\mathbf{3.263}^{\pm.098}$

Table 1. **Quantitative evaluation on our SocialX test sets.** ± indicates a 95% confidence interval and → means the closer to ground truth the better. **Bold face** indicates the best result, while underscore refers to the second best.

Dataset	Method	MPJPE↓	PAMPJPE↓	ACCEL↓
InterHuman	IH	100.30	71.94	26.73
	IX	72.52	49.58	44.00
	XH3D	**65.42**	**43.52**	**8.84**
InterX	IH	100.32	51.49	13.89
	IX	90.19	44.53	31.85
	XH3D	**56.36**	**34.97**	**5.71**

Table 2. Comparison of MPJPE, PAMPJPE, and ACCEL across methods for InterHuman and InterX datasets. ↓ indicates lower values are better. Bold numbers indicate the best results. Our motion representation achieves the best reconstruction quality.

to generate multi-human motion beyond fixed-sized groups. To ensure fair comparison, we retrain several representative open-source baseline methods on our dataset and motion representation and compare them with our approach.

Notably, our method significantly outperforms on metrics such as FID, Diversity, and MModality. However, the improvements in R Precision and MM Dist are less pronounced. We attribute this to the evaluation methodology adopted from [22], which is tailored for text-to-motion scenarios involving one or two individuals. This limitation poses challenges for effectively training and assessing models designed to handle varying group sizes, particularly in terms of motion-language embedding alignment, which R Precision and MM Dist emphasize. We present supporting evidence in the Appendix. Despite this, metrics like FID, Diversity, and MModality remain robust as they focus exclusively on motion embeddings. Qualitative results further highlight our approach, demonstrating its effectiveness in generating high-quality, multi-human motions.

5.4. Qualitative Results

We compare text-to-motion generation results between our method and previous approaches in Figure 3. The results show that motions generated by other methods are less aligned with textual descriptions and sometimes exhibit penetration issues (see InterGen results). In contrast, our method produces smoother and more accurate motions across different numbers of individuals. It effectively captures nuanced behaviors and spatial relationships, faithfully reflecting the given prompts.

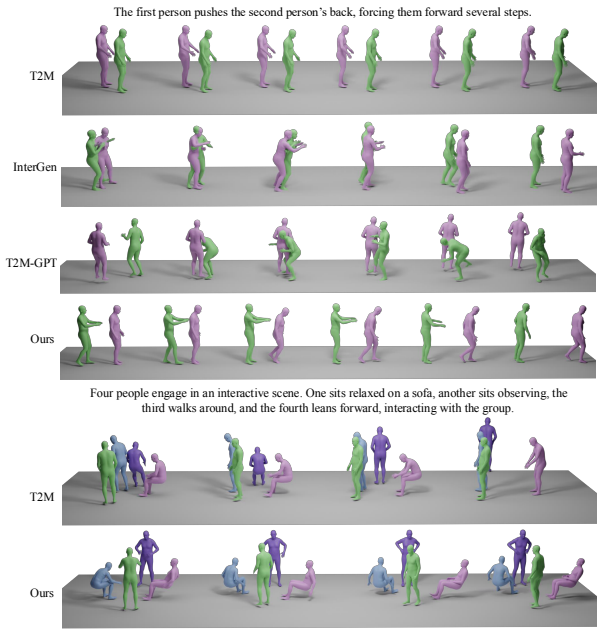

Figure 3. Text-to-motion Generation. Our method produces smoother, more accurate motions that align more precisely with textual descriptions compared to prior approaches.

5.5. Other Motion-relevant Tasks

With a shared space for motion and language tokens, our method naturally supports various motion-related tasks, demonstrating its versatility.

Motion Captioning Motion captioning generates textual descriptions from input. Figure 4 demonstrates that our model effectively captures key actions in social scenarios.

Motion Completion The Motion Completion task involves filling in missing motion segments based on given parts. Given only the start, it becomes motion forecasting (Figure 5), while providing both start and end forms in-between completion (Figure 6). The results demonstrate that our

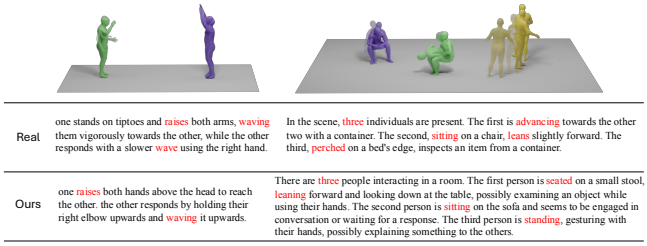

Figure 4. Motion to Text Generation. 'Real' is the labeled caption in the dataset. Our model effectively captures key actions and generates meaningful descriptions.

model effectively handles motion completion for varying numbers of individuals.

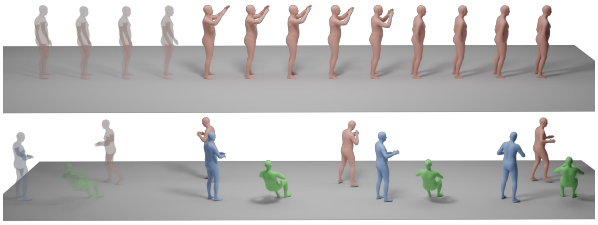

Figure 5. Motion Forecasting. Transparency indicates the initial motion used as a condition. Our model supports any number of individuals.

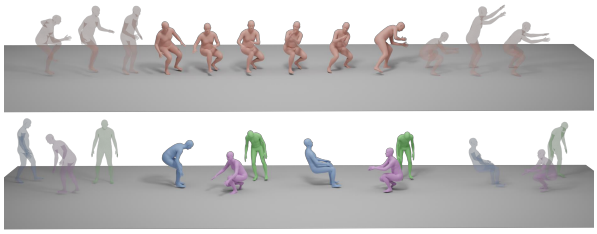

Figure 6. Motion Completion. Transparency indicates the initial and ending motions used as conditions.

Reaction Generation Reaction Generation is a multi-person motion task that involves generating reaction motions based on existing motions. Unlike two-person scenarios, where one motion is used to generate another, our method can generate reactions for any number of individuals, as shown in Figure 7.

5.6. Ablation Study

Motion Tokenizer Codebook Size We explore the impact of varying codebook sizes for VQVAE-based tokenization on our SocialX dataset, with the results presented in Table 3. The results indicate that increasing the codebook size improves the VQVAE's ability to reconstruct motion. After balancing reconstruction performance and codebook size, we select 512 as the codebook size in our experiment.

Data Augmentation To enhance the model's robustness, we incorporate a data augmentation strategy during both the pre-training and post-training stages. This strategy involves randomly shuffling the order of humans in interactions, ensuring that the model's understanding of interactions remains invariant to participant order. This random shuffling not only makes the reference human for computing relative poses variable but also randomizes the order of the remaining participants. Consequently, the model learns to capture interaction dynamics independent of ordering, improving its generalization to diverse scenarios. To assess the effectiveness of this augmentation, we conduct an ablation study. As shown in Table 4, this strategy significantly enhances the diversity of the generated outputs.

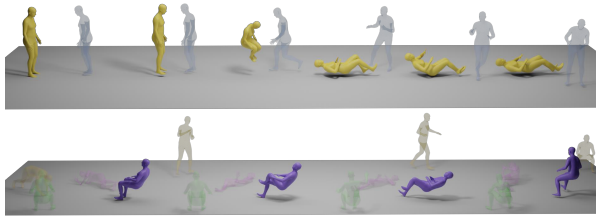

Figure 7. Reaction Generation. Transparency indicates the motions of the persons used as conditions. Our model accommodates any number of individuals as input conditions.

Dataset	#Token	MPJPE↓	PAMPJPE↓	ACCEL↓
SocialX	128	80.01	48.40	7.41
	256	82.17	45.60	7.27
	512	67.25	38.82	6.86
	1024	65.26	35.81	6.97

Table 3. Ablation on VQVAE codebook size.The codebook size impacts model learning—a smaller size increases errors, while a larger size adds learning complexity.

Method	R Precision↑			FID↓	MM Dist↓	Diversity→	MModality↑
	Top 1	Top 2	Top 3				
Ground Truth	$0.542^{\pm.002}$	$0.752^{\pm.002}$	$0.853^{\pm.002}$	$0.0002^{\pm.0001}$	$3.676^{\pm.006}$	$14.334^{\pm.114}$	-
Ours w/o Aug	$0.373^{\pm.003}$	$0.580^{\pm.003}$	$0.678^{\pm.002}$	$1.794^{\pm.007}$	$5.732^{\pm.013}$	$13.817^{\pm.136}$	$3.141^{\pm.105}$
Ours	$0.403^{\pm.003}$	$0.601^{\pm.002}$	$0.717^{\pm.002}$	$1.403^{\pm.005}$	$5.389^{\pm.011}$	$14.213^{\pm.147}$	$3.263^{\pm.098}$

Table 4. Ablation on data augmentation. The augmentation strategy of random shuffling enhances the model's interaction learning.

6. Conclusion and Discussion

We introduce SocialGen, a unified framework for multi-human motion generation, alongside SocialX, the first large-scale social interaction dataset. While our approach improves generation quality, challenges like evaluation metrics and open-domain adaptability remain. Overall, our work advances motion-language research by fostering scalable multi-human interaction modeling.

7. Acknowledgments

This project was partially supported by NIH grant R01AG089169, Massachusetts AI & Technology Center (MassAITC) P30AG073107, and the Stanford University Jaswa Innovator Awards. The authors would also like to thank Chaitanya Patel, Mengyi Shan and Bowei Chen for their valuable input and discussions.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 5

[2] Michael J Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. Bedlam: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8726–8737, 2023. 2

[3] Judith Butepage, Michael J Black, Danica Kragic, and Hedvig Kjellstrom. Deep representation learning for human motion prediction and classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6158–6166, 2017. 2

[4] Zhongang Cai, Jianping Jiang, Zhongfei Qing, Xinying Guo, Mingyuan Zhang, Zhengyu Lin, Haiyi Mei, Chen Wei, Ruisi Wang, Wanqi Yin, et al. Digital life project: Autonomous 3d characters with social intelligence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 582–592, 2024. 2

[5] Changan Chen, Juze Zhang, Shrinidhi K Lakshmikanth, Yusu Fang, Ruizhi Shao, Gordon Wetzstein, Li Fei-Fei, and Ehsan Adeli. The language of motion: Unifying verbal and non-verbal language of 3d human motion. *arXiv preprint arXiv:2412.10523*, 2024. 1, 2

[6] Feilong Chen, Minglun Han, Haozhi Zhao, Qingyang Zhang, Jing Shi, Shuang Xu, and Bo Xu. X-llm: Bootstrapping advanced large language models by treating multi-modalities as foreign languages. *arXiv preprint arXiv:2305.04160*, 2023. 2

[7] Ling-Hao Chen, Jiawei Zhang, Yewen Li, Yiren Pang, Xiaobo Xia, and Tongliang Liu. Humanmac: Masked motion completion for human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9544–9555, 2023. 2

[8] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18000–18010, 2023. 2, 6

[9] Baptiste Chopin, Hao Tang, Naima Otberdout, Mohamed Daoudi, and Nicu Sebe. Interaction transformer for human reaction generation. *IEEE Transactions on Multimedia*, 25: 8842–8854, 2023. 2

[10] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024. 2, 5

[11] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. arxiv 2019. *arXiv preprint arXiv:1901.02860*. 2

[12] Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2, 5

[13] 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. 2

[14] Yao Feng, Jing Lin, Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, and Michael J Black. Chatpose: Chatting about 3d human pose. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2093–2103, 2024. 2

[15] Anindita Ghosh, Noshaba Cheema, Cennet Oguz, Christian Theobalt, and Philipp Slusallek. Synthesis of compositional animations from textual descriptions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1396–1406, 2021. 1, 2

[16] Anindita Ghosh, Rishabh Dabral, Vladislav Golyanik, Christian Theobalt, and Philipp Slusallek. Remos: 3d motion-conditioned reaction synthesis for two-person interactions. In *European Conference on Computer Vision*, pages 418–437. Springer, 2024. 2

[17] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190, 2023. 2

[18] Purvi Goel, Kuan-Chieh Wang, C Karen Liu, and Kayvon Fatahalian. Iterative motion editing with natural language. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–9, 2024. 2

[19] Kehong Gong, Dongze Lian, Heng Chang, Chuan Guo, Zihang Jiang, Xinxin Zuo, Michael Bi Mi, and Xinchao Wang. Tm2d: Bimodality driven 3d dance generation via music-text integration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9942–9952, 2023. 2, 3

[20] John C Gower. Generalized procrustes analysis. *Psychometrika*, 40:33–51, 1975. 6

[21] 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. 2

[22] 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. 2, 3, 5, 6, 7, 4

[23] Chuan Guo, Xinxin Zuo, Sen Wang, and Li Cheng. Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In *European Conference on Computer Vision*, pages 580–597. Springer, 2022. 2, 7

[24] 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. 2

[25] Wen Guo, Xiaoyu Bie, Xavier Alameda-Pineda, and Francesc Moreno-Noguer. Multi-person extreme motion prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13053–13064, 2022. 2

[26] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Audioclip: Extending clip to image, text and audio. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 976–980. IEEE, 2022. 2

[27] Jiaming Han, Kaixiong Gong, Yiyuan Zhang, Jiaqi Wang, Kaipeng Zhang, Dahua Lin, Yu Qiao, Peng Gao, and Xiangyu Yue. Onellm: One framework to align all modalities with language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26584–26595, 2024. 2

[28] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 3

[29] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, et al. Language is not all you need: Aligning perception with language models. *Advances in Neural Information Processing Systems*, 36:72096–72109, 2023. 2

[30] Yinghao Huang, Leo Ho, Dafei Qin, Mingyi Shi, and Taku Komura. Interact: Capture and modelling of realistic, expressive and interactive activities between two persons in daily scenarios. *arXiv preprint arXiv:2405.11690*, 2024. 2

[31] Yiming Huang, Weilin Wan, Yue Yang, Chris Callison-Burch, Mark Yatskar, and Lingjie Liu. Como: Controllable motion generation through language guided pose code editing. In *European Conference on Computer Vision*, pages 180–196. Springer, 2024. 2

[32] Muhammad Gohar Javed, Chuan Guo, Li Cheng, and Xingyu Li. Intermask: 3d human interaction generation via collaborative masked modelling. *arXiv preprint arXiv:2410.10010*, 2024. 1, 6, 2

[33] 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*, 36:20067–20079, 2023. 1, 2, 3, 4, 5, 6

[34] Hanbyul Joo, Tomas Simon, Mina Cikara, and Yaser Sheikh. Towards social artificial intelligence: Nonverbal social signal prediction in a triadic interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10873–10883, 2019. 2, 5, 6, 3

[35] Hanyang Kong, Kehong Gong, Dongze Lian, Michael Bi Mi, and Xinchao Wang. Priority-centric human motion generation in discrete latent space. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14806–14816, 2023. 2

[36] T Kudo. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*, 2018. 4

[37] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 2

[38] 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*, pages 1–21, 2024. 1, 2, 3, 5, 6

[39] Angela S Lin, Lemeng Wu, Rodolfo Corona, Kevin Tai, Qixing Huang, and Raymond J Mooney. Generating animated videos of human activities from natural language descriptions. *Learning*, 1(2018):1, 2018. 1, 2

[40] Jing Lin, Ailing Zeng, Shunlin Lu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-x: A large-scale 3d expressive whole-body human motion dataset. *Advances in Neural Information Processing Systems*, 36: 25268–25280, 2023. 2

[41] Jing Lin, Yao Feng, Weiyang Liu, and Michael J Black. Chathuman: Language-driven 3d human understanding with retrieval-augmented tool reasoning. *arXiv preprint arXiv:2405.04533*, 2024. 2

[42] Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. World model on million-length video and language with ringattention. *arXiv e-prints*, pages arXiv–2402, 2024. 2

[43] Yunze Liu, Changxi Chen, and Li Yi. Interactive humanoid: Online full-body motion reaction synthesis with social affordance canonicalization and forecasting. *arXiv preprint arXiv:2312.08983*, 2023. 2

[44] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866. 2023. 3, 6

[45] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 2

[46] Shunlin Lu, Ling-Hao Chen, Ailing Zeng, Jing Lin, Ruimao Zhang, Lei Zhang, and Heung-Yeung Shum. Humantomato: Text-aligned whole-body motion generation. *arXiv preprint arXiv:2310.12978*, 2023. 6

[47] Mingshuang Luo, Ruibing Hou, Zhuo Li, Hong Chang, Zimo Liu, Yaowei Wang, and Shiguang Shan. M^3gpt: An advanced multimodal, multitask framework for motion comprehension and generation. *Advances in Neural Information Processing Systems*, 2024. 2, 3

[48] Hengbo Ma, Jiachen Li, Ramtin Hosseini, Masayoshi Tomizuka, and Chiho Choi. Multi-objective diverse human motion prediction with knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8161–8171, 2022. 2

[49] 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. 2

[50] Lorenzo Mandelli and Stefano Berretti. Generation of complex 3d human motion by temporal and spatial composition

of diffusion models. *arXiv preprint arXiv:2409.11920*, 2024. 2

[51] Julieta Martinez, Michael J Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2891–2900, 2017. 2

[52] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3d pose estimation from monocular rgb. In *2018 International Conference on 3D Vision (3DV)*, pages 120–130. IEEE, 2018. 2, 5, 6, 3

[53] Evonne Ng, Donglai Xiang, Hanbyul Joo, and Kristen Grauman. You2me: Inferring body pose in egocentric video via first and second person interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9890–9900, 2020. 2

[54] Evonne Ng, Javier Romero, Timur Bagautdinov, Shaojie Bai, Trevor Darrell, Angjoo Kanazawa, and Alexander Richard. From audio to photoreal embodiment: Synthesizing humans in conversations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1001–1010, 2024. 2

[55] OpenAI. Gpt-4o-mini: Advancing cost-efficient intelligence, 2024. 6, 1

[56] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022. 4, 5

[57] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 2

[58] Mathis Petrovich, Michael J Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer vae. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10985–10995, 2021. 2

[59] Mathis Petrovich, Or Litany, Umar Iqbal, Michael J Black, Gul Varol, Xue Bin Peng, and Davis Rempe. Multi-track timeline control for text-driven 3d human motion generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1911–1921, 2024. 2

[60] Ekkasit Pinyoanuntapong, Pu Wang, Minwoo Lee, and Chen Chen. Mmm: Generative masked motion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1546–1555, 2024. 2

[61] Matthias Plappert, Christian Mandery, and Tamim Asfour. The kit motion-language dataset. *Big data*, 4(4):236–252, 2016. 2

[62] Matthias Plappert, Christian Mandery, and Tamim Asfour. Learning a bidirectional mapping between human whole-body motion and natural language using deep recurrent neural networks. *Robotics and Autonomous Systems*, 109:13–26, 2018. 1, 2

[63] Abhinanda R Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J Black. Babel: Bodies, action and behavior with english labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 722–731, 2021. 2

[64] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 5

[65] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020. 2, 4, 5

[66] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019. 4

[67] Pablo Ruiz-Ponce, German Barquero, Cristina Palmero, Sergio Escalera, and José García-Rodríguez. in2in: Leveraging individual information to generate human interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1941–1951, 2024. 2

[68] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 3

[69] Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit H Bermano. Human motion diffusion as a generative prior. *arXiv preprint arXiv:2303.01418*, 2023. 2

[70] Mengyi Shan, Lu Dong, Yutao Han, Yuan Yao, Tao Liu, Ifeoma Nwogu, Guo-Jun Qi, and Mitch Hill. Towards open domain text-driven synthesis of multi-person motions. *arXiv preprint arXiv:2405.18483*, 2024. 2, 3

[71] Shanlin Sun, Gabriel De Araujo, Jiaqi Xu, Shenghan Zhou, Hanwen Zhang, Ziheng Huang, Chenyu You, and Xiaohui Xie. Coma: Compositional human motion generation with multi-modal agents. *arXiv preprint arXiv:2412.07320*, 2024. 2

[72] Mikihiro Tanaka and Kent Fujiwara. Role-aware interaction generation from textual description. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15999–16009, 2023. 2

[73] Chen Tessler, Yunrong Guo, Ofir Nabati, Gal Chechik, and Xue Bin Peng. Maskedmimic: Unified physics-based character control through masked motion inpainting. *ACM Transactions on Graphics (TOG)*, 43(6):1–21, 2024. 2

[74] 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. 2, 7

[75] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 2

[76] Jonathan Tseng, Rodrigo Castellon, and Karen Liu. Edge: Editable dance generation from music. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 448–458, 2023. 2

[77] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 2, 3

[78] NP Van der Aa, Xinghan Luo, Geert-Jan Giezeman, Robby T Tan, and Remco C Veltkamp. Umpm benchmark: A multi-person dataset with synchronized video and motion capture data for evaluation of articulated human motion and interaction. In *2011 IEEE international conference on computer vision workshops (ICCV Workshops)*, pages 1264–1269. IEEE, 2011. 2

[79] Timo Von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European conference on computer vision (ECCV)*, pages 601–617, 2018. 2

[80] Weilin Wan, Zhiyang Dou, Taku Komura, Wenping Wang, Dinesh Jayaraman, and Lingjie Liu. Tlcontrol: Trajectory and language control for human motion synthesis. *arXiv preprint arXiv:2311.17135*, 2023. 2

[81] Yin Wang, Zhiying Leng, Frederick WB Li, Shun-Cheng Wu, and Xiaohui Liang. Fg-t2m: Fine-grained text-driven human motion generation via diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22035–22044, 2023. 2

[82] Qi Wu, Yubo Zhao, Yifan Wang, Yu-Wing Tai, and Chi-Keung Tang. Motionllm: Multimodal motion-language learning with large language models. *arXiv preprint arXiv:2405.17013*, 2024. 1

[83] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519*, 2023. 2

[84] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*, 2021. 2

[85] Liang Xu, Xintao Lv, Yichao Yan, Xin Jin, Shuwen Wu, Congsheng Xu, Yifan Liu, Yizhou Zhou, Fengyun Rao, Xingdong Sheng, et al. Inter-x: Towards versatile human-human interaction analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22260–22271, 2024. 2, 3, 5, 6

[86] Liang Xu, Yizhou Zhou, Yichao Yan, Xin Jin, Wenhan Zhu, Fengyun Rao, Xiaokang Yang, and Wenjun Zeng. Regennet: Towards human action-reaction synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1759–1769, 2024. 2

[87] Ye Yuan and Kris Kitani. Dlow: Diversifying latent flows for diverse human motion prediction. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 346–364. Springer, 2020. 2

[88] Ailing Zeng, Lei Yang, Xuan Ju, Jiefeng Li, Jianyi Wang, and Qiang Xu. Smoothnet: A plug-and-play network for refining human poses in videos. In *European Conference on Computer Vision*, pages 625–642. Springer, 2022. 6

[89] Jun Zhan, Junqi Dai, Jiasheng Ye, Yunhua Zhou, Dong Zhang, Zhigeng Liu, Xin Zhang, Ruibin Yuan, Ge Zhang, Linyang Li, et al. Anygpt: Unified multimodal llm with discrete sequence modeling. *arXiv preprint arXiv:2402.12226*, 2024. 2

[90] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Yong Zhang, Hongwei Zhao, Hongtao Lu, Xi Shen, and Ying Shan. Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14730–14740, 2023. 2, 3, 7

[91] Juze Zhang, Jingyan Zhang, Zining Song, Zhanhe Shi, Chengfeng Zhao, Ye Shi, Jingyi Yu, Lan Xu, and Jingya Wang. Hoi-mˆ 3: Capture multiple humans and objects interaction within contextual environment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 516–526, 2024. 2, 5, 6, 3

[92] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2

[93] Yuxiang Zhang, Liang An, Tao Yu, Xiu Li, Kun Li, and Yebin Liu. 4d association graph for realtime multi-person motion capture using multiple video cameras. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1324–1333, 2020. 2

[94] Yan Zhang, Michael J Black, and Siyu Tang. We are more than our joints: Predicting how 3d bodies move. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3372–3382, 2021. 2

[95] Yaqi Zhang, Di Huang, Bin Liu, Shixiang Tang, Yan Lu, Lu Chen, Lei Bai, Qi Chu, Nenghai Yu, and Wanli Ouyang. Motiongpt: Finetuned llms are general-purpose motion generators. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7368–7376, 2024. 2, 3

[96] Zhuoran Zhao, Jinbin Bai, Delong Chen, Debang Wang, and Yubo Pan. Taming diffusion models for music-driven conducting motion generation. In *Proceedings of the AAAI Symposium Series*, pages 40–44, 2023. 2

[97] Wenyang Zhou, Zhiyang Dou, Zeyu Cao, Zhouyingcheng Liao, Jingbo Wang, Wenjia Wang, Yuan Liu, Taku Komura, Wenping Wang, and Lingjie Liu. Emdm: Efficient motion diffusion model for fast and high-quality motion generation. In *European Conference on Computer Vision*, pages 18–38. Springer, 2024. 2

[98] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5745–5753, 2019. 3

[99] Zixiang Zhou and Baoyuan Wang. Ude: A unified driving engine for human motion generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5632–5641, 2023. 2, 3