Large-Scale Multi-Character Interaction Synthesis

Ziyi Chang
ziyi.chang@durham.ac.uk
Durham University
Durham, United Kingdom

He Wang
he_wang@ucl.ac.uk
University College London
London, United Kingdom

George Alex Koulieris
georgios.a.koulieris@durham.ac.uk
Durham University
Durham, United Kingdom

Hubert P. H. Shum*
hubert.shum@durham.ac.uk
Durham University
Durham, United Kingdom

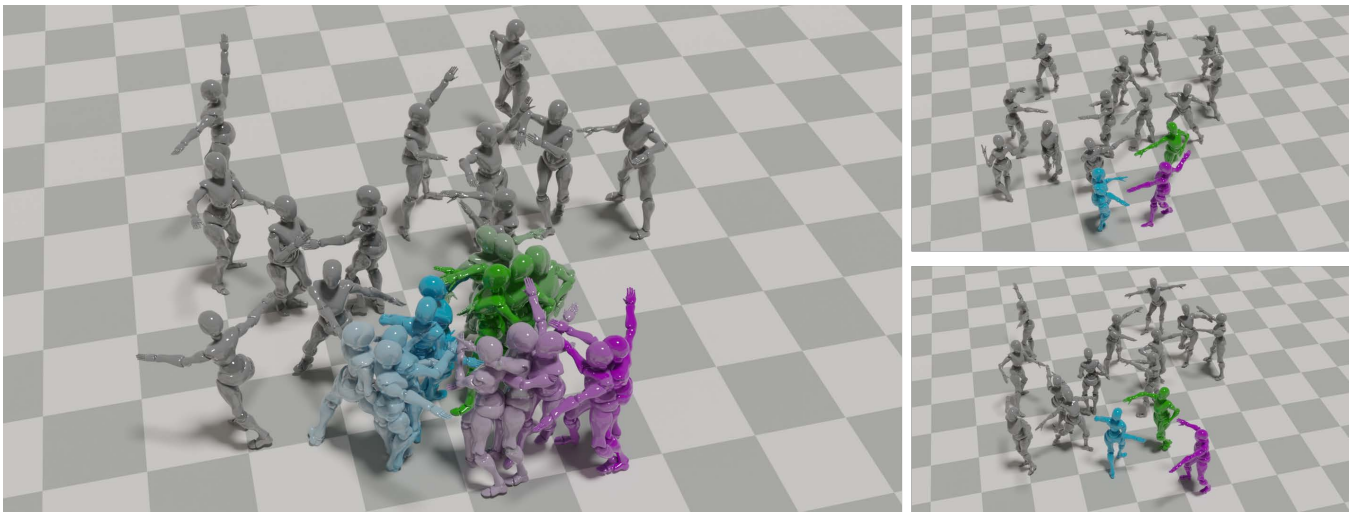

Figure 1: Multi-character interaction coordinated with transition planning. (Left) We highlight the three (currently interacting) characters with blue, purple, and green colors, while others are grey. We show their movements in past frames. The more saturated the color, the more recent the frame. (Upper right) We generate and display the key frame of the transition where the blue and purple characters proceed to have a coordinated interaction. (Lower right) The key frame of transition where the blue and green characters proceed to have a coordinated interaction.

ABSTRACT

Generating large-scale multi-character interactions is a challenging and important task in character animation. Multi-character interactions involve not only natural interactive motions but also characters coordinated with each other for transition. For example, a dancing interaction involves characters that are dancing with partners and also characters that are coordinated to transit to new partners based on spatial and temporal observations. We term such transitions as coordinated interaction for multi-character interactions and decompose it into interaction synthesis and transition planning. Previous methods of single-character animation do not consider interactions that are critical for multiple characters. Deep-learning-based interaction synthesis usually focuses on two characters and does not consider transition planning. Optimization-based interaction synthesis relies on manually designing objective functions that may not generalize well. While crowd simulation involves more characters, their interactions are sparse and passive. We identify two challenges to multi-character interaction synthesis, including the lack of data and the transition planning with close and dense interactions. Existing datasets either do not have multiple characters or do not have close and dense interactions. Planning transitions for multi-character close and dense interactions requires both spatial and temporal consideration. We propose a conditional generative pipeline, including a coordinatable multi-character interaction space for interaction synthesis and a transition planning network to plan transitions. Our experiments demonstrate the effectiveness of our proposed pipeline for multi-character interaction

*Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
SIGGRAPH Conference Papers '25, August 10–14, 2025, Vancouver, Canada
© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

synthesis and the applications facilitated by our method show the scalability and transferability.

CCS CONCEPTS

• **Computing methodologies → Animation**.

KEYWORDS

Motion Synthesis, Multi-person Interaction

ACM Reference Format:

Ziyi Chang, He Wang, George Alex Koulieris, and Hubert P. H. Shum. 2025. Large-Scale Multi-Character Interaction Synthesis. In *Proceedings of Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers '25 (SIGGRAPH Conference Papers '25).* ACM, New York, NY, USA, 10 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Generating large-scale multi-character interactions is a challenging and important task in character animation. For example, a dancing interaction involves characters that are dancing with partners and simultaneously planning future potential interactions with new partners based on spatial and temporal observations. We term such behaviors as *coordinated interaction* in multi-character interactions and decompose them into interaction synthesis and transition planning. The former requires a natural multi-character interaction space to generate realistic interactions while the latter demands the ability to plan transitions among characters. As the number of characters can vary, such a method is required to be scalable for higher application flexibility.

Multi-character interaction synthesis is an underdeveloped area despite recent advances in character animation, interaction generation, and crowd simulation. Single-character animation focuses on modeling motions for a single character with control [Starke et al. 2022]. However, it does not learn the interaction patterns that are critical for multiple characters. Interaction generation with deep learning focuses more on learning motions for two characters [Liang et al. 2024]. While some optimization-based interaction synthesis methods [Shum et al. 2008b] can produce interactions for multiple characters, they depend on manually designing objective functions that may not generalize well. Crowd simulation involves more than two characters, but their interactions are relatively sparse [Charalambous et al. 2023] and passive [Yue et al. 2024]. Our task of large-scale multi-character interaction synthesis requires active and denser interactions with transition planning and needs to be scalable to multiple characters.

Large-scale multi-character interaction synthesis faces two challenges. The first one is the lack of data. Existing datasets for interactions [Liang et al. 2024] focus on two characters, and thus do not consider coordinated interactions, and existing datasets for crowd simulation [Zhong et al. 2022] do not contain dense and close interactions. Capturing such a dataset for our task would be time-consuming and labor-intensive, which becomes unmanageable as the number of characters scales up. The second challenge is to plan dense and close interactions with spatial and temporal observations for multiple characters. Scheduling suitable interactions for multiple characters is a highly correlated problem. In the temporal domain, previous coordination could heavily influence the interactions that follow, and in the spatial domain, the difficulty of planning increases with the increasing number of characters.

We propose a generative pipeline for large-scale multi-character interaction synthesis, including a coordinatable interaction space to generate natural interactions for multiple characters and a transition planning network to coordinate potential transitions among multiple characters. Considering the absence of data, we divide multi-character interactions into several two-character groups that are modeled by a pre-trained two-character interaction diffusion model. With division, we can generalize the natural two-character interaction manifold to an interaction space for multiple characters in the absence of data. Additionally, our division is agnostic to the number of characters and is scalable to multiple characters. To plan transitions for coordinated interactions, we propose a planning network to predict high-level transition plans that are represented as re-grouping choices. Thus, our transition planning network is motion-agnostic, which is transferable to other motion types. Considering the lack of data, we propose to train our method through reinforcement learning. Specifically, our coordinatable interaction space works as the environment to generate coordinated interactions, and the transition planning is the policy network. We define transition smoothness and transition diversity as the rewards to encourage a diverse plan.

We train our method with a two-character dancing subset from the InterHuman dataset [Liang et al. 2024] and evaluate it by synthesizing a larger number of characters and transferring to other motion types. We calculate transition smoothness and hip distance as two metrics for comparison. Our experiments show that our method is scalable and transferable.

We summarize our two contributions below:

- We propose a framework to synthesize large-scale multiple characters by decomposing their coordinated interactions into interaction synthesis and transition planning.
- We propose a method of combining a pre-trained two-character diffusion model and a transition planning network to learn the coordinated interactions via deep reinforcement learning without the requirement for data.

2 RELATED WORK

2.1 Deep Learning-based Motion Synthesis

Deep learning has been extensively utilized for single-character motion synthesis, emphasizing high-fidelity motion details [Zhou et al. 2023] with motion control. Action labels have been popular [Chang et al. 2023a; Xu et al. 2023b] to generate various categories of actions. To achieve finer controllability, texts have been explored as a control signal [Petrovich et al. 2022]. Text-to-motion synthesis aims to integrate language representation into pose representation. For example, a unified text-motion joint space has been modeled [Tevet et al. 2022a] for better representation. More recently, many methods leverage diffusion models [Tevet et al. 2022b] and design conditional mechanisms to control the generation. They either fine-tune a diffusion model on conditional data [Xie et al. 2024] or leverage classifier guidance for post-hoc control without requiring conditional data. Other types of control signals, such as physical constraints [Yuan et al. 2023] and objects,[Li et al. 2023] have also been explored. While the aforementioned methods are effective

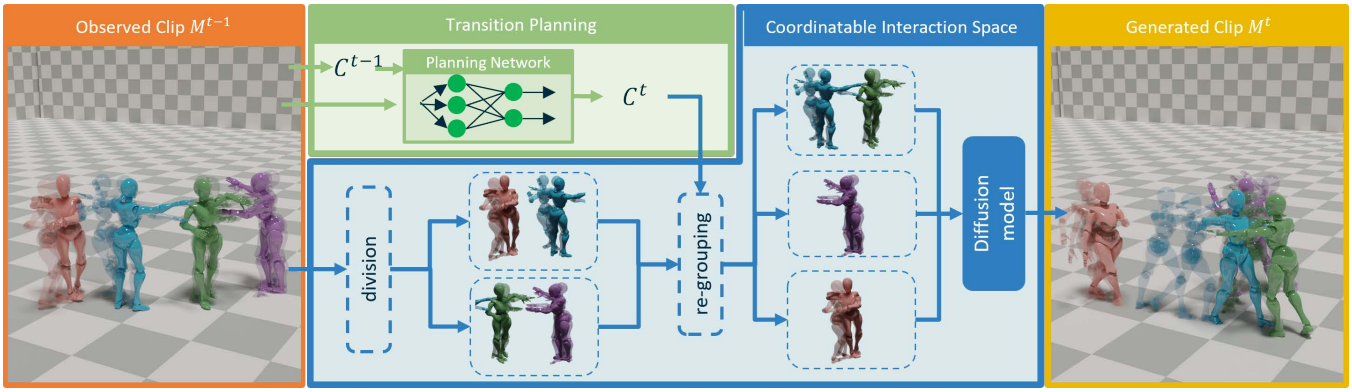

Figure 2: Framework overview. Our pipeline is an autoregressive conditional generative model to plan transitions and synthesize interactions for multiple characters. It has two components: The first component divides multiple characters into groups and contains a pre-trained diffusion-based model to autoregressively generate interactions for each group. The second component predicts a transition plan based on the observed interactions and serves as a conditional signal for the interaction synthesis.

for modeling character motions, they primarily focus on motion synthesis for a single character. Adding control to such methods does not generalize well to multi-character interactions.

2.2 Interaction Synthesis

A popular line of work is based on optimization. It optimizes individual character motions with spatial-temporal constraints [Kwon et al. 2008; Liu et al. 2006] and game theory [Shum et al. 2007, 2008a, 2010] to connect motions smoothly. Optimization is usually based on motion patches, where a patch includes a short interaction of multiple characters [Kim et al. 2012; Shum et al. 2008b; Won et al. 2014; Yersin et al. 2009]. However, a goal function usually needs to be designed manually to perform optimization and may not be well generalized to different motion patterns.

Another line of interaction synthesis is based on modeling the interaction distribution with deep learning methods. The interaction is further defined differently. Some methods simplify interaction synthesis as reaction synthesis [Xu et al. 2024] where the interaction includes active characters performing actions and passive characters responding to such actions. Some methods synthesize character interactions with a large interaction dataset that is usually designed to have two characters by default. They learn the individual motion distribution by modeling the dependency between two characters [Liang et al. 2024; Xu et al. 2023b]. Another line of work synthesizes future interactions based on the observed interactions [Guo et al. 2022; Xu et al. 2023a]. They extract interaction relationships from observed motions and predict future interactions. Others leverage deep reinforcement learning to learn movements for two characters [Won et al. 2021; Zhang et al. 2023]. However, these methods focus more on two characters and cannot be easily extended to multi-character interactions.

2.3 Crowd Simulation

Crowd simulation methods can generally be categorized into two streams: macroscopic and microscopic methods [Pelechano et al. 2016]. Macroscopic methods recognize characters as active matter. They aim to synthesize the density of such active matter in both the spatial and temporal domains. By such an abstraction, they do not focus on the different motions of individual characters.

Microscopic methods for crowd simulation focus on synthesizing individual movements for multiple characters. Trajectories are widely used to represent the movements of crowds. Many recent methods use reinforcement learning to learn instantaneous actions such as the velocity or acceleration of each character. These methods usually focus on the navigation of crowd characters [Charalambous et al. 2023; Panayiotou et al. 2022] by designing rewards for collision avoidance and target reaching. Some learn spatial-temporal features from large trajectories of crowd characters [Mao et al. 2023; Xiang et al. 2024]. Human skeletal joint trajectories have also been considered [Jeong et al. 2024] in a manually crafted dataset where several independent groups are concatenated in a scene and their motions are rather static. Although these methods consider more than two characters, their interactions among characters are sparse in time and space.

3 METHODOLOGY

3.1 Problem Formulation

Our goal is to generate large-scale multi-character interactions without available data. In the absence of data, we decompose the coordinated interactions into interaction synthesis for multiple characters and transition planning for future interactions. The former requires a multi-character interaction space to synthesize natural interactions, while the latter demands the ability to plan transitions with close and dense interactions for multiple characters.

We represent interactions with full-body poses and motion clips. We learn full-body poses for more realistic synthesis because full-body poses may provide social cues [Fiore et al. 2013] for realistic interaction synthesis and transition planning. Moreover, to consider coordinated interactions in the temporal domain, we represent motions by short clips that contain a few consecutive frames to provide contextual information that could reflect character intentions for transitions when compared with poses in a single frame.

Specifically, we denote multi-character interactions as

$$M_{1:N}^{1:T} = \left[M_{1:N}^{1}, M_{1:N}^{2}, \cdots, M_{1:N}^{t}, \cdots, M_{1:N}^{T}\right], \tag{1}$$

where T is the total number of motion clips and N is the total number of characters. The t-th clip $M_{1:N}^{t} = \left[m_1^t, m_2^t, \cdots, m_n^t, \cdots, m_N^t\right]$ consists of motions for N characters where each clip contains w frames and m_n^t is the t-th clip of the n-th character. We follow the representations in [Liang et al. 2024] and represent a motion clip by global positions, local rotations, and velocities of each joint.

3.2 Pipeline Overview

Our pipeline is an autoregressive conditional generative model to synthesize coordinated interactions for multiple characters. We design our method to be autoregressive to allow characters to dynamically plan their future transitions based on the observed interactions between themselves and other characters. This is more flexible than planning everything ahead.

Overall, our framework $\mathcal{F}$, as illustrated in Fig. 2, consists of two components for synthesizing multi-character coordinated interaction. Specifically, the first component, namely the coordinatable interaction space, provides a natural interaction space for multiple characters, and the second component, namely transition planning, predicts future transition plans. The second component predicts a plan to control interactions that arise from the first component.

The first component, as shown in Fig. 3, divides multiple characters into several groups of two characters. With the absence of data, such division allows us to simplify the multi-character interaction space by a learned two-character space. Concretely, we leverage a pre-trained two-character diffusion model to simplify the multi-character interaction space. We group multiple characters by their indices to select corresponding historical motions as conditions for generation. The diffusion model generates the interactions for each group autoregressively. Additionally, by dividing multiple characters into two-character groups, our multi-character interaction space is agnostic to the number of characters.

The second component, as shown in Fig. 4 predicts a future transition plan C^t based on the observed interactions $M_{1:N}^{t-1}$. We design the plan to be high-level re-grouping choices for all characters. By re-grouping characters, the newly grouped characters are coordinated to interact with each other in the next motion clip. We leverage a virtual character in a two-character group if the transition plan for this group contains only one existing character.

Specifically, our method follows an autoregressive conditional generative formulation:

$$M_{1:N}^{t} = \mathcal{F}_\theta(M_{1:N}^{t-1}, \epsilon_{1:N}^{t}, C^t), \tag{2}$$

where $\mathcal{F}$ represents our autoregressive generative pipeline, $\epsilon_{1:N}^{t}$ is the sampled standard Gaussian noise for generation, $M_{1:N}^{t-1}$ is the last observed interaction clip for all characters $1, 2 \cdots, N$, $M_{1:N}^{t}$ is the next interaction clip for all characters, C^t is the transition plan for next motion clip $M_{1:N}^{t}$, and θ represents all trainable parameters.

3.3 Coordinatable Multi-Character Interaction Synthesis

A natural and coordinatable interaction space for multiple characters is vital for coordinated interaction synthesis. It requires the interaction space to have a realistic interaction manifold and to be controllable for coordination. Without available data, we consider simplification by a divide-and-conquer paradigm.

We propose to simplify the interaction space for multiple characters by dividing multiple characters into two-character groups. Multiple character interactions are usually composed of several groups that contain fewer characters [Jeong et al. 2024] and each group can be modeled from an existing model. Theoretically, any division is feasible if a group can be modeled. For our implementation, we divide characters into two-character groups and leverage an existing two-character diffusion model to generate two-character interactions. Through our division, we leverage its ability to synthesize natural interactions for each group. Meanwhile, the division facilitates integrating the transition plan as the conditional signal by re-grouping two characters in an existing group for new interaction groups. Additionally, group division allows our interaction space to be flexible to synthesize varying numbers of characters.

We propose to generalize a two-character diffusion model for multiple characters by autoregressively synthesizing groups. We design our generative method to be based on a two-character diffusion model due to its high capacity to model complex data space [Chang et al. 2023b] and flexible conditional mechanism without being trained on the conditional data [Dhariwal and Nichol 2021]. To autoregressively synthesize the next motion clip for a two-character group, we leverage classifier guidance from the already generated other groups and the observed motion clip for this group. As shown in Fig. 3, the red arrow indicates that we condition on the clips of other groups already generated.

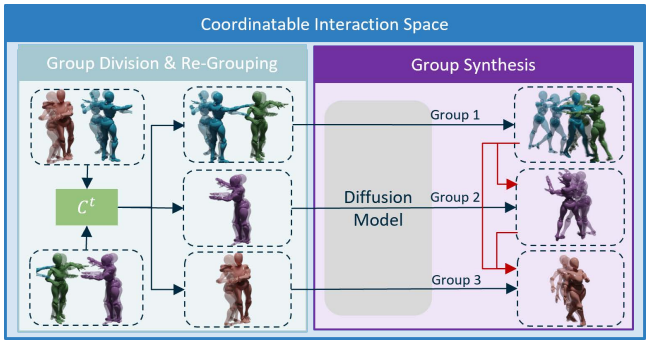

Figure 3: Coordinatable multi-character interaction space by group division. We divide multiple characters into groups and re-group them for potential coordination. The group synthesis generates new motions group by group. The newly generated group is conditioned on the already generated ones, which is indicated by red arrows.

Specifically, when given the observed interaction clip for multiple characters $M_{1:N}^{t-1}$, we generalize a state-of-the-art diffusion model [Liang et al. 2024] for two-character interactions to generate the next interaction clip for all characters. We autoregressively generate a new social group under the condition of existing social groups. These social groups form the next interaction clip for multiple characters $M_{1:N}^{t}$. This generation process is formulated as:

$$M_{1:N}^{t} = \left[g(M_{i,j}^{t-1}, M'),\ for\ (i,j)\ in\ C^t\right], \tag{3}$$

where g is the two-character diffusion model, M' represents groups that have already been generated during this autoregressive procedure, and $M_{i,j}^{t-1}$ represent the observed interaction clip for the i-th and j-th character in a group, C^t is the transition plan to re-grouping observed motions.

We propose to leverage the classifier guidance in the diffusion model to maintain the social distance between the newly generated group $M_{i,j}^t$ and the already generated groups M'. To maintain suitable distances, we leverage a distance function $d(\cdot)$ between different groups. The hip distance follows Proxemics Theory [Rios-Martinez et al. 2015], which states that humans regulate their personal space based on hip position. This promotes the appropriate distances between individuals. Other joints are less representative and would require the pre-trained diffusion model to solve a more complex high-dimensional spatiotemporal constraint. The distance function measures the average of hip-wise distances:

$$d(M_{i,j}^t, M') = \frac{1}{|M'|} \sum_{n'}^{|M'|} \min\left(\|p_{i,j} - p_{n'}\|_2^2 - \tau, 0\right), \tag{4}$$

where τ is the threshold, and $p_{i,j}$ is the hip position of the character i and j. We also constraint the smoothness between the observed clip and the generated clip, which is calculated as:

$$d(M_{i,j}^t, M_{i,j}^{t-1}) = \sum \|acc_{i,j}\|_2^2, \tag{5}$$

where acc represents the acceleration of all joints. Therefore, the constraint function is:

$$d = d(M_{i,j}^t, M') + d(M_{i,j}^t, M_{i,j}^{t-1}) \tag{6}$$

Our design for the coordinatable interaction space is summarized in Algorithm 1.

3.4 Transition Planning

Multi-character interactions contain not only realistic interaction details but also the ability to plan suitable transitions. Their transition is planned on the basis of the currently observed interactions of themselves and others. After a transition is planned, they perform the next actions, which are synthesized by the aforementioned interaction space. Therefore, transition planning contains high-level decisions on interaction groups for multiple characters.

To achieve the planning ability, we design a conditional mechanism via re-grouping for the previously introduced multi-character interaction space. While considering all characters for coordination is theoretically feasible, we implement transition planning locally within four characters to strike a balance between the complexity of grouping choices and the number of potential transition candidates. We balance the trade-off between transition representativeness and learning complexity. A character evaluating all others as potential transition partners improves representativeness, but increases training complexity. Conversely, considering only the nearest character removes the need for learning, but limits transition diversity.

The condition signal in our case is the indices of characters to be grouped together. To address the lack of ground truth data, we use re-grouping as high-level semantic control, which avoids the computational burden of determining low-level joint movements in a high-dimensional spatiotemporal space. Thus, our planning network is agnostic to motion types. Specifically, the planning

ALGORITHM 1: Coordinate-able Multi-Character Interaction Space

Data: Re-grouping choice (i, j) from a transition plan C^t, motion mask m for motion inpainting, observed motions $M_{1:N}^{t-1}$, already generated groups M', a pre-trained model $g(\cdot)$, distance function $d(\cdot)$
Result: a group $M_{i,j}^t$
$M_{i,j}^{t-1} \leftarrow M_{1:N}^{t-1}[(i,j)]$ ▷ Re-grouping by transition plan;
$u \leftarrow U$ ▷ U is total number of diffusion timestep;
$\epsilon^t \leftarrow \mathcal{N}(0, I)$ ▷ Sample a random noise;
$x^u \leftarrow \epsilon^t$;
while $u \neq 0$ **do**
  $x^0 = g(x^u, u)$ ▷ Diffusion predicts xstart;
  $x^0 = m \otimes M_{i,j}^{t-1} + (1 - m) \otimes x^0]$ ▷ Masking for inpainting;
  $x^0 = x^0 + \nabla_x d$ ▷ Classifier guidance;
  $u \leftarrow u - 1$;
end
$M_{i,j}^t \leftarrow x^0$;
$M' \leftarrow M' \cup M_{i,j}^t$

network takes the motions of considered characters and returns a re-grouping choice as a transition plan, which is formulated as:

$$C^t = f_\theta(M_{i,j,i',j'}^t). \tag{7}$$

Due to the lack of an available dataset for our goal, we formulate transition planning as a Markov Decision Process and learn our planning network with a deep reinforcement framework, as depicted in Figure 4. The MDP is represented by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{R}, E)$ where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $\mathcal{R}$ is a scalar reward function, and E is the environment model.

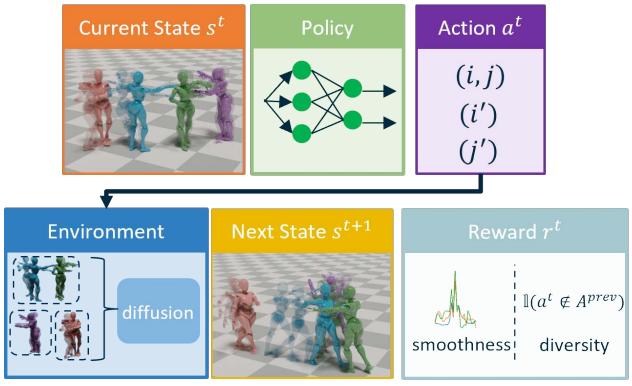

Figure 4: The planning network is learned as a policy network via deep reinforcement learning. The action is a transition plan that contains a high-level grouping choice.

Specifically, a state $s \in \mathcal{S}$ in our setting consists of clips of characters of interest:

$$s := M_{i,j,i',j'}. \tag{8}$$

In our case, an action $a \in \mathcal{A}$ is defined to be a transition plan:

$$a := C, \tag{9}$$

where C represents the re-grouping choice for the four characters.

We leverage the previously defined interaction space as the environment for reinforcement learning. The transition function $E : (s^t, a^t) \rightarrow s^{t+1}$ describes the probability that we transition to state $s^{t+1} \in \mathcal{S}$ given that we executed action $a^t \in \mathcal{A}$ in state $s^t \in \mathcal{S}$. The interaction synthesis module introduced in Section 3.3 works as the environment model in our reinforcement learning setting:

$$E := g(\cdot). \tag{10}$$

To serve as the environment model, our generative model takes as inputs the current state M^t and generates the next state M^{t+1}:

$$s^{t+1} = g\left(\epsilon^t, s^t\right), \tag{11}$$

where $\epsilon^t \sim \mathcal{N}(0, I)$. The first state M^1 is generated by setting the current state to be empty.

The planning network is formulated as the policy network:

$$\pi := f_\theta(\cdot). \tag{12}$$

It takes four characters' motions as the current state and predicts the next state of these characters:

$$a^t = \pi(s^t). \tag{13}$$

A reward function $R : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow R$ and the reward $r := R(s^t, a^t, s^{t+1})$ evaluate a transition (s^t, a^t, s^{t+1}) given the agent task. In our case the task is to interact with all the other three characters as smoothly as possible. The reward function is designed to advocate for such considerations. The smoothness reward is defined as:

$$r_{smooth} = e^{-\|acc^t - acc^{t+1}\|_2^2}, \tag{14}$$

where *acc* corresponds to the acceleration of motions during the transition. In particular, we consider ten frames to calculate the acceleration. The second one is the diversity to encourage more different transition choices. This diversity is calculated as:

$$r_{div} = \begin{cases} 1, & \text{if } a^t \text{ is novel} \\ 0, & \text{otherwise} \end{cases} \tag{15}$$

Therefore, we have the following reward function:

$$r = r_{smooth} + r_{div}. \tag{16}$$

4 EXPERIMENTS

We run a series of experiments to (i) investigate the effectiveness of our method by employing two metrics and (ii) validate the scalability and transferability of our method to three applications.

We evaluate our method with transition smoothness (TS) [Barquero et al. 2024] and hip distance (HD) for comparison. Following [Barquero et al. 2024], transition smoothness calculates the change of acceleration, and we report its maximum change (peak jerk) as a metric. For hip distance, we calculate the average distance between the hip of a character and those of other characters. This metric is designed to indicate whether characters overlapped in the same position. We report the average value over all characters and frames, which is calculated by

$$HD = \frac{2}{N(N-1)F} \sum_{f=1}^{F} \sum_{i,j \in N} \|h_i^f - h_j^f\|_2^2, \tag{17}$$

where N is the set of all characters, F is the number of frames, $\frac{2}{N(N-1)F}$ is an averaging term, and h_i^f and h_j^f are the hip positions for characters i and j at frame f.

4.1 Quantitative Comparison

As currently there is no method for implementing our task, we use InterGen [Liang et al. 2024] as the baseline for our comparison, which is the state-of-the-art method for interaction synthesis. We also implement an adaptation to InterGen for our comparison, and the adapted InterGen is denoted as InterGen†. For our adaptation, we implement exactly the same coordinatable interaction space to allow InterGen to generate multiple characters. For transition planning, as InterGen does not have such planning, we randomly sample coordinations and use them as the control signal. To calculate the metrics, we repeat the generation and report the average value. Table 1 shows the results of our comparison. Our method has the best transition smoothness, which indicates that the characters generated by our method transit with fewer artifacts.

Table 1: Comparison with interaction synthesis models. † represents our implementation of the coordinatable interaction space in the original method. TS denotes transition smoothness and HD, the hip distance.

Methods	TS ↓	HD
InterGen	0.073	0.567
InterGen†	0.117	1.578
Ours	**0.071**	1.963

The InterGen demonstrates close TS performance compared with ours. However, the generated characters are heavily overlapped with each other and the resulting motion is visibly worse. We hypothesize that this better transition smoothness score achieved by InterGen, compared to ours, is because the transition occurs within a much smaller distance between characters and thus the TS value becomes smaller. This is also further corroborated by the hip distance being much higher in InterGen than ours. The InterGen Figure 5 shows that the characters heavily overlap each other.

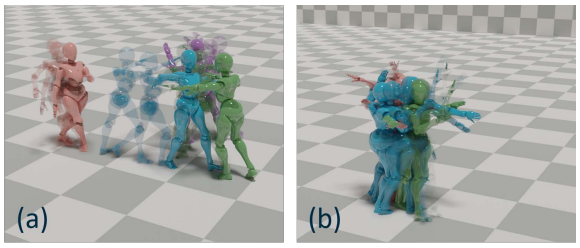

Figure 5: (a) An example result from our method. (b) An example from InterGen where characters heavily overlap.

Figure 6 shows the density of the HD values for the three methods. The two modes in our density indicate that the characters are not overlapped and are clearly transited with the learned planning network. InterGen† does not have the ability to plan transitions, which leads to an averaged distance density with only one mode. InterGen has a curve shape similar to that of InterGen† because

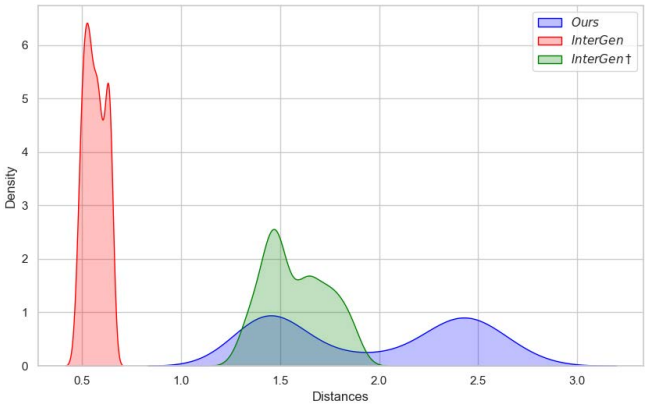

Figure 6: The density of hip distance for the three methods evaluated. The two modes in our hip distance density demonstrate minimal character overlap and clear transitions. InterGen† does not have the ability of transition planning, leading to an averaged distance density with a single mode. InterGen has a similar curve shape with InterGen† as both of them do not have transition planning. Its much smaller mode value indicates that characters heavily overlap.

both do not have transition planning. The mode value of InterGen is much smaller than the other two, which indicates that the characters are heavily overlapped.

As our method generalizes InterGen with coordinatable interaction space and transition planning, Table 1 also shows that the two proposed components are effective. When compared with InterGen†, our method performs better in transition planning. When we compare InterGen and InterGen†, both of which do not have transition planning, we find that the coordinatable interaction space in InterGen† helps to reduce the overlap in InterGen.

4.2 Extended Applications

We demonstrate the scalability of our method and the transferability of our planning network by implementing three applications. Our method is scalable to multiple characters despite only having access to a two-character dataset. We show that synthesizing more characters via our method can be done by either gradually adding new characters or generating a larger scene as a whole. Our planning network is also transferable to other types of motion because our action space is defined as the high-level grouping decision. We show this transferability by taking boxing motions as an example.

4.2.1 Adding New Character Synthesis. Our method facilitates the application of adding new characters to achieve a larger number of characters. In this experiment, we start with four characters and gradually add other characters. The newly added characters can also be coordinated to interact with existing characters via the re-grouping choice from our planning network. Table 2 shows the performance of adding new characters.

4.2.2 Generating Large Scenes. We generalize our method to synthesize a large number of characters. Although we only have access to a two-character dataset, we generate a lot more characters in one scene. Instead of gradually adding new characters, we choose to generate these characters at once. Table 2 shows the performance in the generation of large scenes.

4.2.3 Other types of motion. As we design the planning network to predict the suitable re-grouping choice without considering the actual movements, the planning network is transferable to other motion types. For our experimental implementation, we report the performance of transferring the planning network trained on dancing motions to boxing motions without any adaptation. Table 2 shows the transition smoothness and the hip distance.

When applying our planning network to boxing motions, the transition smoothness does not get degraded and the hip distance increases a bit. This shows that by designing transition plan to be high-level re-grouping operation, the planning network is not related to future movements of characters. This helps our planning network to be transferable. The hip distance increases slightly because boxing motions usually have larger distances than dancing.

Table 2: Method performance on extended applications. TS denotes transition smoothness and HD, the hip distance.

Methods	TS	HD
Adding New Characters	0.026	2.450
Generating Large Scenes	0.075	3.372
Boxing	0.057	2.155

5 CONCLUSION

We propose a method to synthesize coordinated interactions for multiple characters without any multi-character dataset. We decompose coordinated interactions into interaction synthesis and coordination planning, and propose a coordinatable interaction space and a planning network, respectively. Considering the lack of data, a pre-trained two-character interaction diffusion model is leveraged in our interaction space by decomposing multi-character interactions into interactions of two-character groups. The planning network provides control signals by re-grouping the observed motions. Our experiments show that the motions from our method are smoother, and character limbs penetrate less with one another compared with other methods. We also demonstrate scalability and transferability with extended applications.

While our method can achieve multi-character coordinated interactions without accessing any multi-character dataset, limitations exist, most of which arise due to the lack of multi-character datasets. The first one is that our method relies on the dividing approximation between multiple characters and two characters. The second one is the controllability of diffusion models when there is no such conditioned dataset. Although classifier guidance and motion inpainting are used in our method, control accuracy can be further improved [Karunratanakul et al. 2023] for better generation quality. Additionally, our method relies on the pre-trained two-character interaction diffusion model to synthesize motions for two-character groups. The synthesis quality of two-character groups also influences the quality of our method. One potential direction for future work is collecting data on close interactions between two or more

people. Existing crowd simulation data mainly focus on 2D and recently 3D motions like collision avoidance. With data collected, our method would enhance fine-grained interactions and improve visual realism. For large scenes, an imaginary character is introduced if the total number of characters is not divisible by four and future improvements could include predefined actions such as walking or recycling characters [Shum et al. 2008b] to further enhance transitions. Currently, a 4-character group is formed by a greedy distance-based strategy, and future work could explore other advanced methods such as a first-person receptive field.

ACKNOWLEDGMENTS

This research is supported in part by the EPSRC NortHFutures project (ref: EP/X031012/1).

REFERENCES

German Barquero, Sergio Escalera, and Cristina Palmero. 2024. Seamless Human Motion Composition with Blended Positional Encodings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Ziyi Chang, Edmund J. C. Findlay, Haozheng Zhang, and Hubert P. H. Shum. 2023a. Unifying Human Motion Synthesis and Style Transfer with Denoising Diffusion Probabilistic Models. In *Proceedings of the 2023 International Conference on Computer Graphics Theory and Applications* (Lisbon, Portugal) *(GRAPP '23)*. SciTePress, 64–74. https://doi.org/10.5220/0011631000003417

Ziyi Chang, George A Koulieris, and Hubert PH Shum. 2023b. On the design fundamentals of diffusion models: A survey. *arXiv preprint arXiv:2306.04542* (2023).

Panayiotis Charalambous, Julien Pettre, Vassilis Vassiliades, Yiorgos Chrysanthou, and Nuria Pelechano. 2023. GREIL-Crowds: Crowd Simulation with Deep Reinforcement Learning and Examples. *ACM Transactions on Graphics (TOG)* 42, 4 (2023), 1–15.

Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems* 34 (2021), 8780–8794.

Stephen M Fiore, Travis J Wiltshire, Emilio JC Lobato, Florian G Jentsch, Wesley H Huang, and Benjamin Axelrod. 2013. Toward understanding social cues and signals in human–robot interaction: effects of robot gaze and proxemic behavior. *Frontiers in psychology* 4 (2013), 859.

Wen Guo, Xiaoyu Bie, Xavier Alameda-Pineda, and Francesc Moreno-Noguer. 2022. Multi-person extreme motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13053–13064.

Jaewoo Jeong, Daehee Park, and Kuk-Jin Yoon. 2024. Multi-agent long-term 3d human pose forecasting via interaction-aware trajectory conditioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1617–1628.

Korrawe Karunratanakul, Konpat Preechakul, Supasorn Suwajanakorn, and Siyu Tang. 2023. Guided Motion Diffusion for Controllable Human Motion Synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2151–2162.

Manmyung Kim, Youngseok Hwang, Kyunglyul Hyun, and Jehee Lee. 2012. Tiling motion patches. In *Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation*. 117–126.

Taesoo Kwon, Young-Sang Cho, Sang I Park, and Sung Yong Shin. 2008. Two-character motion analysis and synthesis. *IEEE transactions on visualization and computer graphics* 14, 3 (2008), 707–720.

Jiaman Li, Jiajun Wu, and C Karen Liu. 2023. Object motion guided human motion synthesis. *ACM Transactions on Graphics (TOG)* 42, 6 (2023), 1–11.

Han Liang, Wenqian Zhang, Wenxuan Li, Jingyi Yu, and Lan Xu. 2024. Intergen: Diffusion-based multi-human motion generation under complex interactions. *International Journal of Computer Vision* (2024), 1–21.

C Karen Liu, Aaron Hertzmann, and Zoran Popović. 2006. Composition of complex optimal multi-character motions. In *Proceedings of the 2006 ACM SIGGRAPH/Eurographics symposium on Computer animation*. 215–222.

Weibo Mao, Chenxin Xu, Qi Zhu, Siheng Chen, and Yanfeng Wang. 2023. Leapfrog diffusion model for stochastic trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5517–5526.

Andreas Panayiotou, Theodoros Kyriakou, Marilena Lemonari, Yiorgos Chrysanthou, and Panayiotis Charalambous. 2022. Ccp: Configurable crowd profiles. In *ACM SIGGRAPH 2022 conference proceedings*. 1–10.

Nuria Pelechano, Jan M Allbeck, Mubbasir Kapadia, and Norman I Badler. 2016. *Simulating heterogeneous crowds with interactive behaviors*. CRC Press.

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

Jorge Rios-Martinez, Anne Spalanzani, and Christian Laugier. 2015. From proxemics theory to socially-aware navigation: A survey. *International Journal of Social Robotics* 7 (2015), 137–153.

Hubert PH Shum, Taku Komura, Masashi Shiraishi, and Shuntaro Yamazaki. 2008b. Interaction patches for multi-character animation. *ACM transactions on graphics (TOG)* 27, 5 (2008), 1–8.

Hubert PH Shum, Taku Komura, and Shuntaro Yamazaki. 2007. Simulating competitive interactions using singly captured motions. In *Proceedings of the 2007 ACM symposium on Virtual reality software and technology*. 65–72.

Hubert PH Shum, Taku Komura, and Shuntaro Yamazaki. 2008a. Simulating interactions of avatars in high dimensional state space. In *Proceedings of the 2008 Symposium on interactive 3D Graphics and Games*. 131–138.

Hubert PH Shum, Taku Komura, and Shuntaro Yamazaki. 2010. Simulating multiple character interactions with collaborative and adversarial goals. *IEEE Transactions on Visualization and Computer Graphics* 18, 5 (2010), 741–752.

Sebastian Starke, Ian Mason, and Taku Komura. 2022. Deepphase: Periodic autoencoders for learning motion phase manifolds. *ACM Transactions on Graphics (TOG)* 41, 4 (2022), 1–13.

Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. 2022a. Motionclip: Exposing human motion generation to clip space. In *European Conference on Computer Vision*. Springer, 358–374.

Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. 2022b. Human motion diffusion model. *arXiv preprint arXiv:2209.14916* (2022).

Jungdam Won, Deepak Gopinath, and Jessica Hodgins. 2021. Control strategies for physically simulated characters performing two-player competitive sports. *ACM Transactions on Graphics (TOG)* 40, 4 (2021), 1–11.

Jungdam Won, Kyungho Lee, Carol O'Sullivan, Jessica K Hodgins, and Jehee Lee. 2014. Generating and ranking diverse multi-character interactions. *ACM Transactions on Graphics (TOG)* 33, 6 (2014), 1–12.

Wei Xiang, YIN Haoteng, He Wang, and Xiaogang Jin. 2024. SocialCVAE: Predicting Pedestrian Trajectory via Interaction Conditioned Latents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 6216–6224.

Yiming Xie, Varun Jampani, Lei Zhong, Deqing Sun, and Huaizu Jiang. 2024. OmniControl: Control Any Joint at Any Time for Human Motion Generation. In *The Twelfth International Conference on Learning Representations*.

Liang Xu, Ziyang Song, Dongliang Wang, Jing Su, Zhicheng Fang, Chenjing Ding, Weihao Gan, Yichao Yan, Xin Jin, Xiaokang Yang, et al. 2023b. ActFormer: A GAN-based transformer towards general action-conditioned 3D human motion generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2228–2238.

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

Qingyao Xu, Weibo Mao, Jingze Gong, Chenxin Xu, Siheng Chen, Weidi Xie, Ya Zhang, and Yanfeng Wang. 2023a. Joint-Relation Transformer for Multi-Person Motion Prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9816–9826.

Barbara Yersin, Jonathan Maïm, Julien Pettré, and Daniel Thalmann. 2009. Crowd patches: populating large-scale virtual environments for real-time applications. In *Proceedings of the 2009 symposium on Interactive 3D graphics and games*. 207–214.

Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. 2023. Physdiff: Physics-guided human motion diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 16010–16021.

Jiangbei Yue, Baiyi Li, Julien Pettré, Armin Seyfried, and He Wang. 2024. Human Motion Prediction Under Unexpected Perturbation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1501–1511.

Yunbo Zhang, Deepak Gopinath, Yuting Ye, Jessica Hodgins, Greg Turk, and Jungdam Won. 2023. Simulation and retargeting of complex multi-character interactions. In *ACM SIGGRAPH 2023 Conference Proceedings*. 1–11.

Jinghui Zhong, Dongrui Li, Zhixing Huang, Chengyu Lu, and Wentong Cai. 2022. Data-driven crowd modeling techniques: A survey. *ACM Transactions on Modeling and Computer Simulation (TOMACS)* 32, 1 (2022), 1–33.

Kanglei Zhou, Hubert PH Shum, Frederick WB Li, and Xiaohui Liang. 2023. Multi-task spatial-temporal graph auto-encoder for hand motion denoising. *IEEE Transactions on Visualization and Computer Graphics* 30, 10 (2023), 6754–6769.

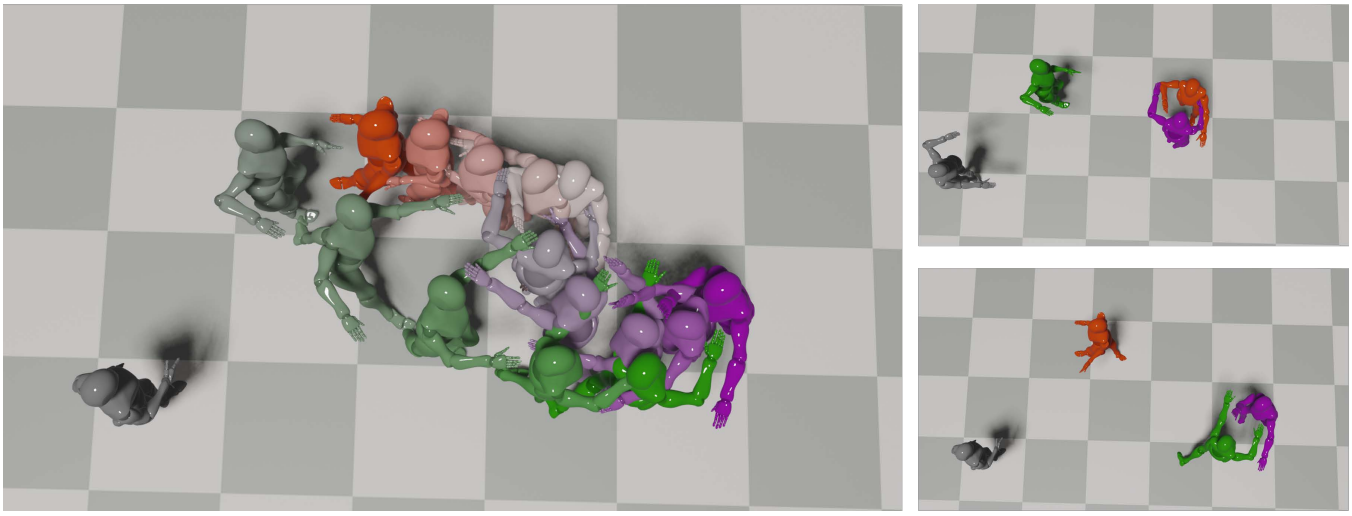

Figure 7: Snapshot of key frames during a transition period for four characters. The more saturated the color, the more recent the frame is. We highlight the three interacting characters with red, purple, and green color while the one in the left corner is grey as this transition only involves the three colored characters. (Left) We show character movements during the transition using key frames. In this transition, the purple character previously interacts with the red one and then is coordinated to interact with the green character. (Right Upper) We generate and display the starting key frame of this transition where the red and purple characters are interacting with each other. (Right Low) The ending key frame of this transition where the purple and green characters are interacting with each other.

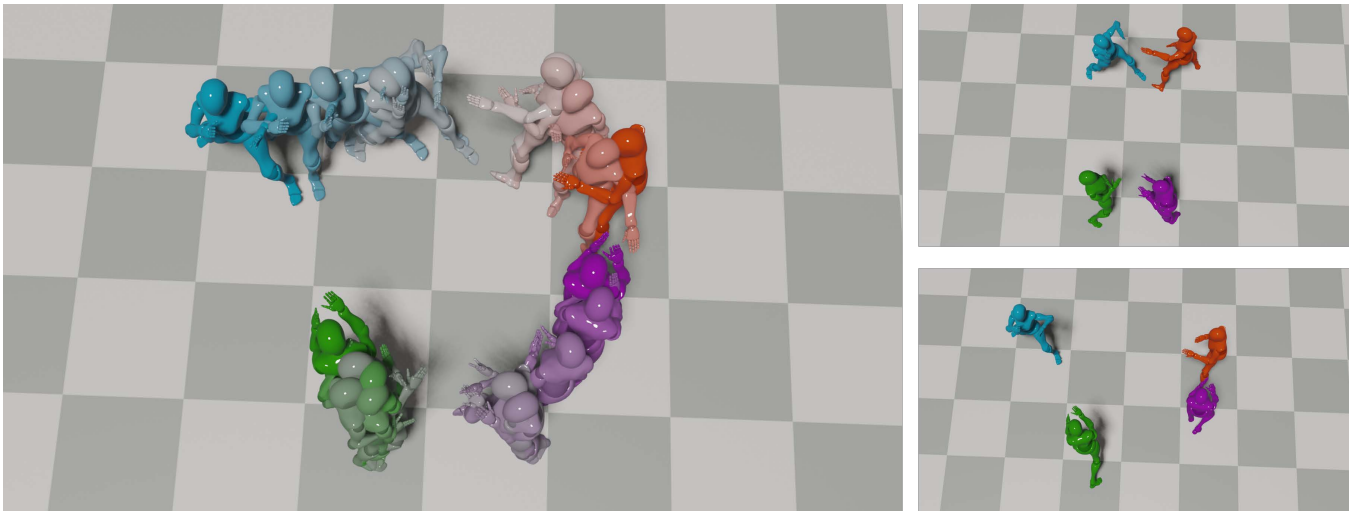

Figure 8: Snapshot of key frames during a transition period in the experiment of transferring to other motion type, e.g., boxing. The more saturated the color, the more recent the frame is. We highlight the four characters with red, purple, blue, and green color. (Left) We show character movements during the transition using key frames. In this transition, the purple and green characters previously interact with each other, and the red and blue ones previously interact with each other. Then the red and purple characters are coordinated to interact, and the blue and green ones are also coordinated to interact. (Right Upper) We generate and display the starting key frame of this transition where the red and blue characters are interacting with each other, and the purple and green ones are also interacting with each other. (Right Low) The ending key frame of this transition where the purple and red characters are interacting with each other, and the blue and green are interacting with each other.

Figure 9: Snapshot of key frames during a transition period in the experiment of adding new characters. The more saturated the color, the more recent the frame is. We highlight the three characters with red, purple, and blue color while others are grey because this transition only involves the three colored characters. The purple and blue characters are more recently added than the red one. (Left) We show their movements during the transition using key frames. In this transition, the purple character previously interacts with the blue one and then is coordinated to interact with the red character. (Right Upper) We generate and display the starting key frame of this transition where the blue and purple characters are interacting with each other. (Right Low) The ending key frame of this transition where the purple and red characters are interacting with each other.

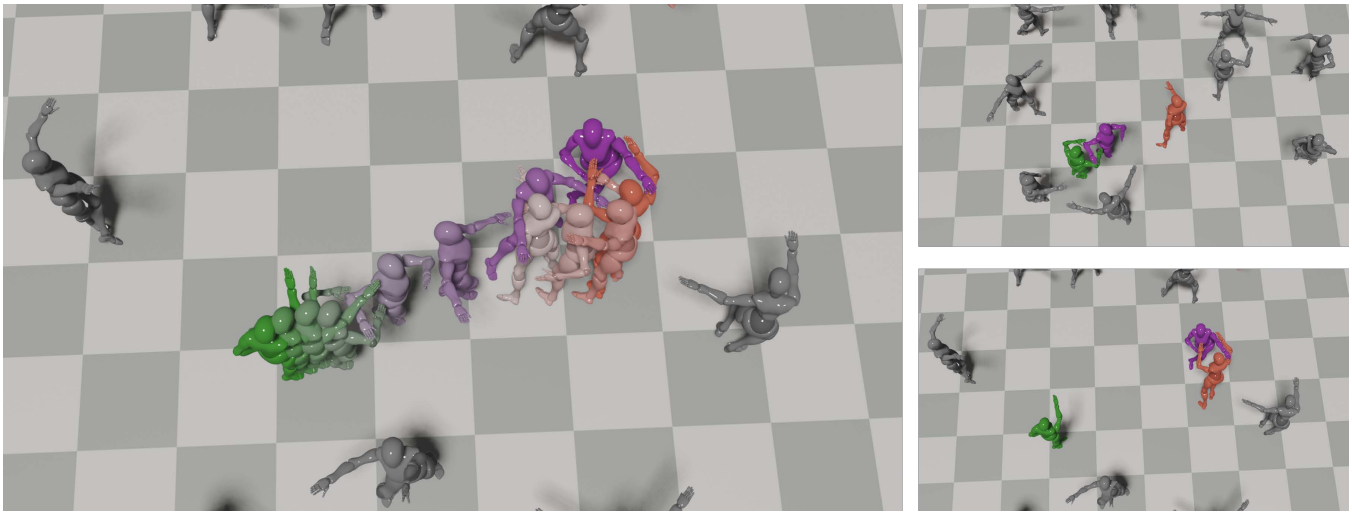

Figure 10: Snapshot of key frames during a transition period in the experiment of generating a large number of characters. The more saturated the color, the more recent the frame is. We highlight the three characters with red, purple, and green color while others are grey because this transition only involves the three colored characters. (Left) We show their movements during the transition using key frames. In this transition, the purple character previously interacts with the green one and then is coordinated to interact with the red character. (Right Upper) We generate and display the starting key frame of this transition where the green and purple characters are interacting with each other. (Right Low) The ending key frame of this transition where the purple and red characters are interacting with each other.