Learning from Synthetic Human Group Activities

Che-Jui Chang†
Rutgers University
chejui.chang@rutgers.edu

Danrui Li
Rutgers University
danrui.li@rutgers.edu

Deep Patel
NEC Laboratories
dpatel@nec-labs.com

Parth Goel
Rutgers University
goel.parth210@gmail.com

Honglu Zhou
NEC Laboratories
hozhou@nec-labs.com

Seonghyeon Moon†
Rutgers University
sm206@cs.rutgers.edu

Samuel S. Sohn
Rutgers University
samuel.sohn@rutgers.edu

Sejong Yoon
The College of New Jersey
yoons@tcnj.edu

Vladimir Pavlovic
Rutgers University
vladimir@rutgers.edu

Mubbasir Kapadia
Roblox
mkapadia@roblox.com

Abstract

The study of complex human interactions and group activities has become a focal point in human-centric computer vision. However, progress in related tasks is often hindered by the challenges of obtaining large-scale labeled datasets from real-world scenarios. To address the limitation, we introduce M²Act, a synthetic data generator for multi-view multi-group multi-person human atomic actions and group activities. Powered by Unity Engine, M²Act features multiple semantic groups, highly diverse and photorealistic images, and a comprehensive set of annotations, which facilitates the learning of human-centered tasks across single-person, multi-person, and multi-group conditions. We demonstrate the advantages of M²Act across three core experiments. The results suggest our synthetic dataset can significantly improve the performance of several downstream methods and replace real-world datasets to reduce cost. Notably, M²Act improves the state-of-the-art MOTRv2 on DanceTrack dataset, leading to a hop on the leaderboard from 10th to 2nd place. Moreover, M²Act opens new research for controllable 3D group activity generation. We define multiple metrics and propose a competitive baseline for the novel task. Our code and data are available at our project page: http://cjerry1243.github.io/M²Act.

1. Introduction

Understanding collective human activities and social groups carries significant implications across diverse domains, as it contributes to bolstering public safety within surveillance systems, ensuring safe navigation for autonomous robots and vehicles amidst human crowds, and enriching social awareness in human-robot interactions [8, 9, 11, 12, 21, 37, 49, 51]. However, the advancement in related tasks is often impeded by the challenges of obtaining large-scale human group activity datasets in real-world scenarios with fine-grained multifaceted annotations.

Generating synthetic data is an emerging alternative to collecting real-world data due to its capability of producing large-scale datasets with perfect annotations. Nonetheless, most synthetic datasets [4, 20, 40, 48, 53] are primarily designed to facilitate human pose and shape estimation. They can only provide data with independently-animated persons, which is unsuitable for tasks in single-group and multi-group conditions [51]. To address the limitation, we propose M²Act, a synthetic data generator, with multi-view multi-group multi-person human actions and group activities. As presented in Tab. 1, M²Act stands out by offering comprehensive annotations including both 2D and 3D annotations as well as fine-grained person-level and group-level labels, thereby making it an ideal synthetic dataset generator to support tasks such as human activity recognition and multi-person tracking across all listed real-world datasets.

Illustrated in Fig. 1, our synthetic data generator features multiple semantic groups, highly diverse and photorealistic images, and a rich set of annotations. It encompasses 25 photometric 3D scenes, 104 HDRIs (High Dynamic Range Images), 5 lighting volumes, 2200 human models, 384 animations (categorized into 14 atomic action classes), and 6 group activities. For our experiments, We generated two

†Work done during internship at Roblox
datasets, $M^3$ActRGB and $M^3$Act3D. $M^3$ActRGB contains both single-group and multi-group data with a total of 6M frames of RGB images and 48M bounding boxes, rendered in 20 FPS. $M^3$Act3D is a 3D-only and single-group dataset, which contains 3D motions of all persons within a group. It has large group sizes (max 27 people) and an average of 6.7 persons per group. In total, $M^3$Act3D contains a duration of 87.6 hours of group activities, captured in 30 FPS.

We first demonstrate the merit of $M^3$Act via synthetic data pre-training and mixed training on multi-person tracking and group activity recognition. For multi-person tracking, training with our synthetic data yields significant performance gain on several downstream methods [23, 52, 56, 57]. We also demonstrate notable improvements in the state-of-the-art MOTRv2 method [57] and observe that our synthetic data can substitute for 62.5% more real-world data, without compromising performance. In terms of group activity recognition, results indicate that pre-training with $M^3$ActRGB greatly improves both person-level and group-level accuracy for Composer [58] and ActorTransformer [25] methods. Based on our generated data, we then introduce a novel task, controllable 3D group activity generation, which aims to synthesize a group of 3D human motions, given control signals such as activity labels and group sizes. We systematically approach the new task by introducing both learning-based and heuristics-based metrics, along with a competitive baseline to generate meaningful human activities.

This paper makes the following contributions:

- We propose a novel synthetic data generator, $M^3$Act, and provide two large-scale synthetic datasets with highly diverse human activities, photorealistic multi-view videos, and comprehensive annotations.

- We demonstrate that $M^3$Act can significantly improve benchmark performances for multi-person tracking and group activity recognition and replace a large portion of real-world training data to reduce cost.

- $M^3$Act promotes new research initiatives for controllable 3D group activity generation, suggesting that synthetic data can not only support existing tasks but also create datasets for novel research.

2. Related Works

Human Centered Synthetic Datasets. The use of synthetic datasets for human-centered tasks has become increasingly prominent due to their diversity, scalability, and perfect annotations, with proven merits connected to various fields in machine learning, including domain adaptation [34, 46], heterogeneous multitask learning [54], and sim2real [24] or task2sim [39] transfer. Most previous synthetic datasets are constructed to support human pose estimation. For example, SURREAL [48] contains renderings of human motions from 145 avatars composited to a background image. Subsequent works [2, 20, 40] managed to improve the image quality by leveraging realistic 3D scenes, high-quality renderings, and HDRI images. Recently, synthetic datasets have been proposed to tackle human shape and mesh estimation. SynBody [53] constructs layered human assets to increase character diversity. BEDLAM [4] adds physically simulated hair and clothes to achieve state-of-the-art performances on shape and mesh estimation. Nonetheless, data with collective human motions and group activities cannot be obtained from them. Our work, $M^3$Act, is constructed with animated human groups tailored to multi-person and multi-group research.

Real Multi-Person Datasets. Real-world datasets [15, 17,
Table 1. A comparison of synthetic datasets as well as commonly-used real datasets for activity understanding and person tracking. We refer to JRDB as a union set of JRDB, JRDB-Act, and JRDB-Pose datasets. Note that it offers 3D bounding boxes, but not poses.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Image Type</th>
<th>Avatar Num.</th>
<th>Video</th>
<th>Multi-View</th>
<th>Multi-Person</th>
<th>Multi-Group</th>
<th>2D</th>
<th>3D</th>
<th>Annotations</th>
<th>Group Act.</th>
</tr>
</thead>
<tbody>
<tr>
<td>SURREAL, 2017 [48]</td>
<td>Composite</td>
<td>145</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>AGORA, 2021 [40]</td>
<td>HDRI</td>
<td>350</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>HSPACE, 2021 [2]</td>
<td>3D Scene</td>
<td>1600</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>GTA-Humans, 2021 [6]</td>
<td>3D Scene</td>
<td>600</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>PSP-HDRI+, 2022 [20]</td>
<td>HDRI</td>
<td>28</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>SynBody, 2023 [53]</td>
<td>3D Scene</td>
<td>10k</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>BEDLAM, 2023 [4]</td>
<td>3D Scene</td>
<td>271</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

|M³Act (Ours)                     | Photometric 3D + HDRI | 2200 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| HiEve, 2020 [32]                | Real          | -         | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| PoseTrack, 2021 [18]           | Real          | -         | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| MOT, 2020 [17]                 | Real          | -         | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| DanceTrack, 2022 [45]          | Real          | -         | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| JRDB, 2023 [21, 37, 49]        | Real          | -         | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

18, 29, 32, 33, 37, 45] with multiple persons are usually collected for tasks such as group activity understanding, multi-person tracking, and human trajectory prediction. Recognizing and parsing collective human activities [25, 58] rely primarily on multiple modalities (RGB, bounding box, pose) and hierarchical action and activity labels. These fine-grained labels are provided by datasets like CAD [15] and Volleyball Dataset [29]. On the other hand, datasets for person tracking, such as HiEve [32], MOT [17, 38], and DanceTrack [45], require not only 2D annotations (e.g., bounding box) for individual frames, but also the association of the objects between them. Specifically, DanceTrack provides multiple persons in a group with the same clothing, making it difficult for the association of the individuals. MOT datasets target tracking for human crowds and contain mostly outdoor scenes from a bird-eye view. Recently, JRDB [21, 37, 49] with rich annotations is released. The images are captured by a social robot, navigating around daily scenes. It provides fine-grained annotations that support various tasks, including person detection, pose estimation, tracking, collective action detection, and understanding. M³Act not only offers the same modalities and annotations for supporting the aforementioned tasks, but it also provides full 3D annotations, making it suitable for a wide range of applications beyond the 2D domain.

3. M³Act

M³Act is a multi-view multi-group multi-person human atomic action and group activity data generator built with Unity Engine and the Perception [5] library. Inspired by PeopleSansPeople [19] that populates randomly posed human avatars in a scene and renders static images, M³Act not only offers the same functionalities for human poses but also extends it to the spatio-temporal domain. It generates RGB videos for dynamic human motions and produces a rich set of annotations simultaneously, including (a) 2D and 3D joints/meshes, (b) 2D and 3D bounding boxes for individual persons, (c) atomic action and group activity categories, (d) tracking information such as individual and group IDs, (e) segmentation, depth, and normal images, and (f) scene description.

3.1. Data Generation

The process of our data generation is illustrated in Fig. 2. First, the generation process is configured by the simulation scenario that manages multiple independent simulations of human activities. Then for each simulation, a 3D scene with background objects, lights, and cameras is set up, and groups of human characters are instantiated to be animated. Lastly, the multi-view RGB image frames are rendered and the annotations are exported at the end of the simulation.

Scene Instantiation. We represent the environment through 25 photometric 3D scenes and 104 panoramic HDRIs. Each scene is initialized with randomized lighting and camera configuration. To attain a balance between realistic environmental illumination and pronounced shadow detail, our lighting schema integrates HDRI Sky lighting with a directional light. This directional light is subject to random variations in its direction, color temperature, and intensity. Regarding camera placement, we always point cameras towards the center of avatar groups and introduce variability by randomizing both the field of view and the camera’s distance to these groups.

Human Models and Motion Assets. M³Act leverages 2000 human models generated by Synthetic Humans [42], ranging across all ages (1 ~ 100), genders, ethnicities (such as Caucasian, Asian, Latin American, African, Middle Eastern), diverse body shapes, hair, and clothing. We also incorporated 200 widely-used human characters from RenderPeople [1]. For the human motions, we collected 384 animation clips from AMASS [36], categorized into 14 atomic action classes. We created a universal animation controller
that blends styles, including arm space and stride size, to create diverse motions from the collected clips.

**Modular Group Activities.** Each group activity is structured as a parameterized module, allowing for the customization of numerous variables. These variables include the number of individuals in the group and the specific atomic actions permitted within the group activity. This modularization ensures easy duplication, repositioning, and reuse of the group activities, enabling simulations of multiple groups at the same time. To procedurally animate a group of humans within a modular group, we establish the positions and orientations of the selected characters while choosing the appropriate animation clip for each character through the activity script. It’s important to note that, despite drawing characters from the same set of avatars, the configurations and animations of these characters can vary significantly from one group to another. For example, animating a queueing activity may require all characters to be aligned in a straight line, while those in a walking group may form various shapes. The atomic actions that a person can perform also depend on the specific group activity. We carefully consider all these factors when authoring activities and provide a summary in the Sup. Mat.

**Domain Randomization.** $M^3$Act provides domain randomization for almost all aspects of the data generation process to ensure the simulation data is highly diverse. These aspects include the number of groups in a scene, the number of persons in each group, the positions of groups, the alignment of persons in a group, the positions of individuals, the textures for the instantiated characters, and the selection of scenes, lighting conditions, camera positions, characters, group activities, atomic actions, and animation clips. Despite the fact that animating group activities inherently limits the degree of freedom in the placement of characters, by altering the shapes in which characters align (e.g., either in a cluster, a straight line, or a curve), $M^3$Act nonetheless generates diverse activities and achieves sufficient randomization for downstream model generalization. More details regarding the randomization variables and their distributions are provided in Sup. Mat.

**Rendering and Annotations.** $M^3$Act utilizes the Unity high-definition render pipeline for the creation of photorealistic RGB images and leverages the Perception library for capturing annotations. On average, data is generated at a rate of 4.2 FPS using one NVIDIA RTX 3070 Ti graphics card, with a resolution of 1920x1080, and all annotations are enabled. Similar to PeopleSansPeople, the 2D skeleton follows COCO [31] format, with additional labelers for exporting 3D joints, meshes, group IDs, and activity classes. After the data generation, the 3D joints and meshes are fitted with SMPL parameters [35, 41].

### 3.2. Dataset Statistics

$M^3$Act comprises 25 photometric 3D scenes, 104 HDRIs, 5 lighting volumes, 2200 human models, 384 animations (categorized into 14 atomic action classes), and 6 group activities. Using the generator, we first generated our synthetic dataset, $M^3$ActRGB. It contains 6K simulations of every single-group activity and 9K simulations of multi-group configuration, with 4 camera views. Each simulation produces a 5-second video clip, captured in 20 FPS and FHD resolution. In total, $M^3$ActRGB contains 6M RGB images and 48M bounding boxes. The average track length is 4.65 seconds.

Additionally, we generated $M^3$Act3D, a large-scale 3D-only dataset. It consists of more than 65K simulations of single-group activity. Each contains 150 frames of multi-person collective interactions in 30 FPS, resulting in a total duration of 87.6 hours. As shown in Tab. 2, both the group size and the interaction complexity are significantly higher than those in previous multi-person motion datasets. Specifically, when compared with GTA-Combat [44], $M^3$Act3D contains more persons in a group, has more group activities, and provides a variable number of persons in a group. To the best of our knowledge, $M^3$Act3D is the first large-scale 3D dataset for human group activities with large group sizes as well as per-frame individual action labels. See Sup. Mat.
Table 2. A comparison of datasets for 3D multi-person human activities. m³Act3D is the largest dataset with labels for atomic actions and more persons in a group.

for detailed statistics of both datasets.

4. Experiments
We showcase the practical utilities of m³Act through three core experiments: Multi-Person Tracking (MPT), Group Activity Recognition (GAR), and controllable Group Activity Generation (GAG). The experiments are carefully designed to cover the following three perspectives:

- **Multi-modality**: Our experiments cover various modalities contained within our dataset, including RGB videos, 2D keypoints, and 3D joints. We leverage the rich annotations including bounding boxes, tracklets, group activities, and person action labels.

- **Performance**: We conduct the ablation study by altering the amount and the type of synthetic data used for training to see its effect on the model performance.

- **Novel task**: We introduce a novel generative task (GAG), showing that synthetic data can not only support existing CV tasks but also create datasets for new research.

4.1. Multi-Person Tracking
The objective of MPT is to predict the trajectories of all persons from a dynamic video stream. Typically, person tracking involves two separate processes, person detection and association. While the tracking task is approached in some prior works with the tracking-by-detection method [3, 7, 50], we consider end-to-end approaches [23, 52, 56, 57] to evaluate the use of synthetic data on the performance of MPT as a whole, in lieu of an improved performance caused only by refined detection.

**Real-world Dataset: DanceTrack [45]** (DT) is a challenging MPT dataset characterized by dynamic movements with human subjects in uniform appearances. It has a total of 100 videos with over 105K frames.

**Synthetic Dataset.** Given the motion categories in the real-world dataset, we select a subset of m³ActRGB with groups of people dancing, walking, and running. We use 1000 video clips with a single “dance” group as well as 1500 videos with a “walk” group and a “run” group simulated at the same time (denoted as WalkRun). We alter the use of the synthetic group activities (Dance, WalkRun, and Dance+WalkRun) in our experiments.

Results. We mix together both synthetic and real data during training and present the results in Tab. 3. First, adding our synthetic data yields significant improvement in all 5 tracking metrics as well as a hop in ranking on HOTA from 10th to 2nd place. The model trained with our synthetic data plus the extra association, marked as DT+Syn (WR+D), achieves similar performance to MOTRv2*, the same model that is trained with additional validation data with an ensemble of 4 models [57]. This suggests that the synthetic data used in our experiment is equivalent to at least 62.5% more real data. Second, Compared with other synthetic data sources, such as BEDLAM and GTA-Humans, m³Act demonstrates superior performance, indicating its better suitability for multi-person dynamic conditions. Third, we observe that the type of synthetic groups affects the model performance on real data. Adding the “WalkRun” groups to the training data is more effective than adding the “Dance” group. This is because while the DanceTrack dataset contains dynamic dance movements, the real challenge lies in detection and tracking when the subjects switch positions. By design, the positions of the characters in our dance group are well-staged and the movements are nearly synchronous. (See Sup. Mat. for the design.) Contrarily, having a walk and a run group together
in a scene leads to frequent position switches relative to the camera view and thus improves the model performance. Lastly, Tab. 4 presents the tracking results using different methods. Results indicate that our synthetic data is effective across various models.

4.2. Group Activity Recognition

The goal of GAR is to determine the class of the group activity performed by the dominant group as well as the action class of each person. We consider Composer [58] and Actor Transformer [25] as the benchmark models. The former is a multi-scale transformer-based model and accepts only 2D keypoints as input. The latter can take combinations of multiple input modalities.

Real-world Dataset: CAD2 [15] and Volleyball Dataset [29]. CAD2 is an extended version of the Collective Activity Dataset [14] that records human group activities and is widely used for GAR benchmarks [51]. Volleyball Dataset (VD) is an action recognition dataset. It has 55 videos with 84 player action labels and 8 team activity labels.

Synthetic Dataset. We use a subset of all single-group data from M^3Act.RGB. It contains a total of 10K videos and over 600K frames. It contains all group activity and individual action classes that CAD2 provides and further includes 7 more action types.

Results. We experimentally study how the size of our pretraining synthetic dataset and the capacity of a GAR model affect generalization from the synthetic to real domains. We first train the models on different amounts of synthetic data. Then we fine-tune them on CAD2 and report the top 1 accuracy of both group activity and person action recognition on the test set of CAD2. Tab. 5 presents the results using only 2D keypoints as input. We see a common trend for both GAR models that the recognition accuracy increases as more synthetic data is used for pre-training. With 100% of synthetic data, the accuracy of Composer increases, with an average of 4.87% at the group level and 7.43% at the person level, whereas Actor Transformer sees a 5.59% increase at group level and 5.43% increase at person level. Moreover, Tab. 6 shows the group recognition accuracy using different input modalities on CAD2 and VD. The performance gains in the experiment indicate that our synthetic data can effectively benefit the downstream GAR task across different methods, input modalities, and datasets.

4.3. Controllable 3D Group Activity Generation

While the procedural generation of our group activities in M^3Act yields realistic and diverse human activities, the implementation requires considerable effort and involves the design and application of specific heuristics. Learning a generative model for human group activities, instead, encodes the heuristics to the architecture inherently and encompasses the capabilities of probabilistically generating diverse activities, with control over the entire group of human motions from various signals. To this end, we introduce controllable 3D group activity generation (GAG).

Definition. Let \( G^p \) be a group of human motions with \( t \) frames and \( p \) persons. The individual pose is denoted as \( m^i_j \in R^{j \times d} \), where \( j \) is the number of joints in the skeleton and \( d \) is the joint dimension. The goal of GAG is to synthesize a group of 3D human motions \( G^p \) from Gaussian noise, given an activity label and an arbitrary group size as input conditions. It requires a model capable of learning the temporal and spatial motion dependencies among persons within the same group and generating human motions with any group size simultaneously. GAG is related to dyadic motion generation [10, 16] and partner-conditioned reaction generation [43], but involves the motion interactions of more than two persons.

Baselines. Although previous works [16, 43, 44] can generate motions for multiple persons, they are limited to dyadic scenarios or groups with a fixed number of persons. Therefore, we present two baselines. The first one is the vanilla motion diffusion model, MDM [47]. It was proposed for probabilistic single-person motion generation from an input condition. We adopt their action-to-motion architecture for conditional synthesis and train the model on M^3Act.RGB.
for generating an individual person’s motion from an input group activity class label. In order to generate a group of human motions from a given group size, we repeat the single-person inference several times. In other words, the individual motions are generated independently by MDM. Our second baseline, MDM+IFormer is extended from MDM and includes an additional interaction transformer (IFormer) that works along the dimension of the persons. The interaction transformer encourages the model to learn the inter-person motion dependencies. At inference, MDM+IFormer is capable of producing coordinated group activities in one forward pass, due to its modeling of human interactions.

**Implementation.** We utilize a common skeleton for all individuals with 25 joints. We process the data so each motion is represented as both the 6D joint rotations [59] and the root positions. The final representation of a collective group activity with multiple persons is a tensor with shape (#persons × #frames × 26 × 6). For a fair comparison, both baseline models were trained on an NVIDIA RTX 3090 graphics card with 90% data from the baseline, MDM+IFormer is capable of generating group activities with low collision frequency and similar social force values to the ground truth. For the evaluation, we generated 500 samples for each group activity using the two well-trained baseline models. Each generated sequence contains 60 frames. We use the test split as the ground truth and randomly extract the group activity of the same length for evaluation. To ensure the distributions of group sizes are similar to the ground truth, we calculate the minimum and maximum group sizes from the training split and uniformly sample a group size from that range to generate group activities. Please refer to Sup. Mat. for more details regarding the baseline architectures, metrics formulas, and evaluation.

### 4.3.2 Results

MDM+IFormer is capable of generating group activities with well-aligned character positions. As shown in Fig. 3, MDM generates human groups that are poorly positioned. For example, the persons in a walking group do not walk in the same direction and the persons are poorly placed in a queueing group. This is because MDM generates the group activities by inferring the individual motions independently. The placement of the individuals simply follows the probabilistic distribution of all persons’ positions in the dataset. On the other hand, MDM+IFormer successfully learns the probabilistic distribution for the entire group due to its in-
We show the merit of $M^3$Act by conducting three core experiments with multiple modalities and enhanced performances, as well as introducing a novel generative task. In both MPT and GAR experiments, we observe positive correlations between the volume of synthetic data used for training and model performance, indicating an improved model generalizability to unseen test cases with more synthetic data. Moreover, our comparison between DT+Syn and MOTRv2 reveals that synthetic data can replace certain real-world data from the target domain without sacrificing performance [13]. Essentially, our synthetic data reduces the need for extensive real data during training, thereby effectively lowering the costs associated with data collection and annotation. This discovery represents a promising step towards achieving few-shot and potentially zero-shot sim2real transfer. In our 3D Group Activity Generation experiment, we observe that MDM+IFormer, despite being a baseline for the novel task, learns to embed the heuristics for person interactions and produces well-aligned groups given the controls. It’s important to highlight that the generative approach, though currently underperforms the procedural method (GT), demonstrates the unique potential of controlling the group motions directly from various signals, including activity class, group size, trajectory, density, speed, and text inputs. With the anticipation of more data availability and increased model capacity for generative models in the future, we expect the generative method to eventually prevail, leading to broader applications for social interactions and human collective activities.

While the complexity of group behaviors in our dataset may be constrained by the heuristics used for activity authoring, $M^3$Act offers notable flexibility for incorporating new group dynamics tailored to any specific downstream tasks. These new groups could be derived from expert-guided heuristics, rules generated by large language models, or outputs from our 3D GAG model. Furthermore, we recognize the existing domain gaps between synthetic and real-world data. With more assets included in our data generator in future iterations, we can enhance model generalizability and alleviate the disparities.

### 6. Acknowledgement

The research was supported in part by NSF awards: IIS-1703883, IIS-1955404, IIS-1955365, RETTL-2119265, and EAGER-2122119. This work was also partially supported by the Center for Smart Streetscapes, an NSF Engineering Research Center, under cooperative agreement EEC-2133516. This material is based upon work supported by the U.S. Department of Homeland Security under Grant Award Number 22STESE00001 01 01.

---

#### Table 7. The results of the generated group activities with the learning-based metrics at both levels. An up arrow means the result is better when the metric score is higher. A right arrow means the metric score should be close to ground truth (GT).

<table>
<thead>
<tr>
<th></th>
<th>Group Level</th>
<th>Person Level</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Acc ↑</td>
<td>FID ↓</td>
</tr>
<tr>
<td>GT</td>
<td>99.937</td>
<td>0.001 ± 0.000</td>
</tr>
<tr>
<td>MDM</td>
<td>97.367</td>
<td>3.909 ± 0.019</td>
</tr>
<tr>
<td>MDM+IFormer</td>
<td>98.100</td>
<td>3.242 ± 0.016</td>
</tr>
</tbody>
</table>

#### Table 8. Results of the generated human activities with position-based metrics. The collision frequency is calculated on a 60-frame group activity and normalized by the total number of interactions in a group.

<table>
<thead>
<tr>
<th></th>
<th>Collision</th>
<th>Interaction</th>
<th>Contact</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Freq ↓</td>
<td>Force →</td>
<td>Force →</td>
<td>Force →</td>
</tr>
<tr>
<td>GT</td>
<td>0.037</td>
<td>65.79</td>
<td>46.55</td>
<td>112.33</td>
</tr>
<tr>
<td>MDM</td>
<td>3.643</td>
<td>7,121.50</td>
<td>3,822.47</td>
<td>10,903.57</td>
</tr>
<tr>
<td>MDM+IFormer</td>
<td>1.157</td>
<td>1,796.25</td>
<td>1,373.40</td>
<td>3,167.80</td>
</tr>
</tbody>
</table>
References


[12] Che-Jui Chang, Danrui Li, Seonghyeon Moon, and Mubbasir Kapadia. On the equivalency, substitutability, and flexibility of synthetic data, 2024. 8


