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 M^{3}Act*, a synthetic data generator for* ***m****ulti-view* ***m****ulti-group* ***m****ulti-person human atomic* ***act****ions and group* ***act****ivities. Powered by Unity Engine,* M^{3}Act *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* M^{3}Act *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,* M^{3}Act *improves the state-of-the-art MOTRv2 on DanceTrack dataset, leading to a hop on the leaderboard from* 10^{th} *to* 2^{nd} *place. Moreover,* M^{3}Act *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/M3Act.*

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 M^{3}Act, a synthetic data generator, with **m**ulti-view **m**ulti-group **m**ulti-person human **act**ions and group **act**ivities. As presented in Tab. 1, M^{3}Act 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

Figure 1. **M^{3}Act is a large-scale synthetic data generator designed to support multi-person and multi-group research topics.** M^{3}Act features multiple semantic groups and produces highly diverse and photorealistic videos with a rich set of annotations suitable for human-centered tasks including multi-person tracking, group activity recognition, and controllable human group activity generation.

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,

Dataset	Image Type	Avatar Num.	Video	Multi-View	Multi-Person	Multi-Group	Annotations 2D	3D	Atomic Atn.	Group Act.
SURREAL, 2017 [48]	Composite	145	✓				✓	✓		
AGORA, 2021 [40]	HDRI	350			✓		✓	✓		
HSPACE, 2021 [2]	3D Scene	1600	✓	✓	✓		✓	✓		
GTA-Humans, 2021 [6]	3D Scene	600	✓		✓		✓	✓		
PSP-HDRI+, 2022 [20]	HDRI	28			✓		✓		✓	
SynBody, 2023 [53]	3D Scene	10k	✓	✓	✓		✓	✓		
BEDLAM, 2023 [4]	3D Scene	271	✓		✓		✓	✓		
M^{3}Act (Ours)	Photometric 3D + HDRI	2200	✓	✓	✓	✓	✓	✓	✓	✓
CAD, 2011 [15]	Real	-	✓		✓	✓	✓		✓	✓
Volleyball Dataset, 2016 [29]	Real	-	✓		✓	✓	✓		✓	✓
NTU-RGBD 120, 2019 [33]	Real	-	✓	✓	✓		✓	✓	✓	
HiEve, 2020 [32]	Real	-	✓		✓	✓	✓		✓	
PoseTrack, 2021 [18]	Real	-	✓		✓	✓	✓			
MOT, 2020 [17]	Real	-	✓		✓	✓	✓			
DanceTrack, 2022 [45]	Real	-	✓		✓		✓			
JRDB, 2023 [21, 37, 49]	Real	-	✓		✓	✓	✓		✓	✓

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.

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 activity detection, and understanding. M^{3}Act 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. M^{3}Act

M^{3}Act is a **m**ulti-view **m**ulti-group **m**ulti-person human atomic **act**ion and group **act**ivity 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, M^{3}Act 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 initiated 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. M^{3}Act 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

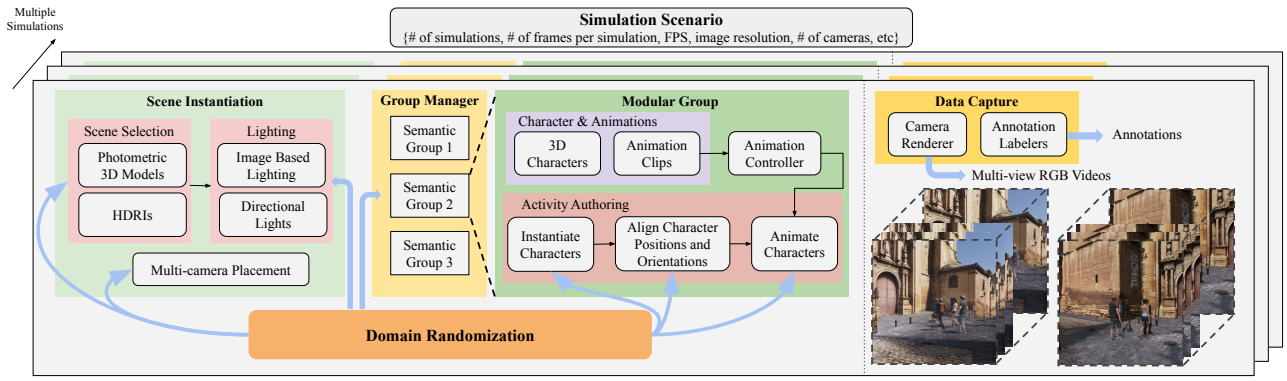

Figure 2. The data generation process of M^{3}Act. It consists of multiple data simulations with scene instantiation, group activity authoring, and a data capture module. A high degree of randomization is involved in all aspects of the process to ensure diverse data.

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.

Dataset	FPS	# of Acty.	# of Persons Avg	# of Persons Max	# of Actn.	Duration
SBU [55]	15	8	2.0	2	-	7.6 mins
Duet Dance [30]	25	5	2.0	2	-	2.3 hrs
CHI3D [22]	50	8	2.0	2	-	2.7 hrs
NTU RGBD 120 [33]	30	**26**	2.0	2	-	15.0 hrs
GTA-Combat [44]	15	1	3.2	5	-	39.0 hrs
M^{3}Act3D	30	6	**6.7**	**27**	**8**	**87.6 hrs**

Table 2. A comparison of datasets for 3D multi-person human activities. M^{3}Act3D 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 M^{3}Act 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 M^{3}ActRGB 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.

Training Data	HOTA↑	DetA↑	AssA↑	IDF1↑	MOTA↑
DT$^{\circledast}$	69.8	83.0	58.9	71.6	89.3
DT	68.8 (10)	82.5	57.4	70.3	90.8
DT + Syn (D)	59.0	75.5	46.1	59.0	82.6
DT + Syn (WR)	70.1	83.1	59.4	72.5	92.0
DT + Syn (WR+D)	**71.9** (2)	**83.6**	**62.0**	**74.7**	**92.6**
DT + Syn† (WR+D)	72.2	83.4	62.6	75.5	**92.7**
DT (MOTRv2*)	**73.4**	**83.7**	**64.4**	**76.0**	92.1
DT + BEDLAM [4]	55.9	68.7	44.5	53.8	79.1
DT + GTA-Humans [6]	54.1	66.8	44.2	52.1	78.8

Table 3. MPT results on DanceTrack with MOTRv2. "D" means synthetic dance group. "WR" means walk and run groups. "WR+D" refers to "D" and "WR" combined. The symbol $\circledast$ represents the author-provided checkpoint. The symbol † marks the same model with additional association at inference. Numbers in parentheses represent the rank in the DanceTrack leaderboard.

Model	Syn. Data	HOTA	DetA	AssA	IDF1	MOTA
MOTR [56]		54.2	73.5	40.2	51.5	79.7
	✓	**60.0**	**76.4**	**48.1**	**56.0**	**83.8**
MeMOTR [23]		68.5	80.5	58.4	71.2	89.9
	✓	**71.1**	**81.8**	**62.3**	**74.1**	**92.2**
CO-MOT [52]		69.4	82.1	58.9	71.9	91.2
	✓	**72.5**	**83.6**	**63.3**	**75.9**	**92.8**
MOTRv2 [57]		68.8	82.5	57.4	70.3	90.8
	✓	**71.9**	**83.6**	**62.0**	**74.7**	**92.6**
MOTRv2* [57]		73.4	83.7	64.4	76.0	92.1
	✓	**74.6**	**84.1**	**64.9**	**76.4**	**93.1**

Table 4. MPT results on DanceTrack using different methods trained with our synthetic data.

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 10^{th} to 2^{nd} 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, M^{3}Act 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

Model	Pretrained Syn. Data	Group Activity Top 1 Acc (%) ↑	Person Action Top 1 Acc (%) ↑
Composer [58]	N/A	$84.87^{\pm 2.3}$ (88.20)	$81.31^{\pm 2.4}$ (83.13)
	10%	$86.12^{\pm 1.8}$ (87.87)	$84.16^{\pm 1.8}$ (86.03)
	25%	$87.65^{\pm 1.2}$ (89.01)	$86.36^{\pm 1.3}$ (86.81)
	50%	$89.39^{\pm 0.4}$ (90.14)	$86.68^{\pm 1.5}$ (87.99)
	100%	$\mathbf{89.74}^{\pm 1.0}$ (**91.51**)	$\mathbf{88.74}^{\pm 1.7}$ (**89.05**)
	Gains	**+4.87 (+3.31)**	**+7.43 (+5.92)**
Actor Transformer [25]	N/A	$78.08^{\pm 1.0}$ (79.47)	$76.22^{\pm 2.2}$ (78.07)
	10%	$77.59^{\pm 2.4}$ (81.00)	$76.01^{\pm 3.2}$ (79.76)
	25%	$81.36^{\pm 2.1}$ (83.19)	$78.86^{\pm 2.4}$ (80.05)
	50%	$82.72^{\pm 1.3}$ (84.56)	$79.95^{\pm 1.6}$ (81.47)
	100%	$\mathbf{83.67}^{\pm 1.2}$ (**84.88**)	$\mathbf{81.65}^{\pm 1.2}$ (**82.22**)
	Gains	**+5.59 (+5.41)**	**+5.43 (+4.15)**

Table 5. Results of 2D keypoint-based **group activity and person action recognition** on CAD2 dataset. The best results are shown in parentheses. The results suggest that pre-training with our synthetic data largely increases the accuracy for both group activity and person actions. Note that group accuracy saturates at 93.4% and 86.2% for Composer and Actor Transformer respectively.

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 9 player action labels and 8 team activity labels.

Synthetic Dataset. We use a subset of all single-group data from M^{3}ActRGB. 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 pre-training 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.

Model	2D	RGB	Flow	CAD2	Syn+CAD2	VD	Syn+VD
Composer	✓			88.2	91.5	94.6	95.1
Actor Transformer	✓			79.5	84.9	92.3	93.7
		✓		78.2	80.7	91.4	92.5
	✓	✓		81.0	**85.2**	93.5	94.3
		✓	✓	81.3	85.0	94.4	**95.0**
	✓		✓	79.5	81.9	93.0	94.1

Table 6. The group activity recognition accuracy on CAD2 and Volleyball Dataset using different input modalities.

4.3. Controllable 3D Group Activity Generation

While the procedural generation of our group activities in M^{3}Act 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 $G_t^p = \{m_i^n\}_{i=1\sim t, n=1\sim p}$ be a group of human motions with t frames and p persons. The individual pose is denoted as $m_i^n \in R^{j\times d}$, 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_t^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^{3}Act3D

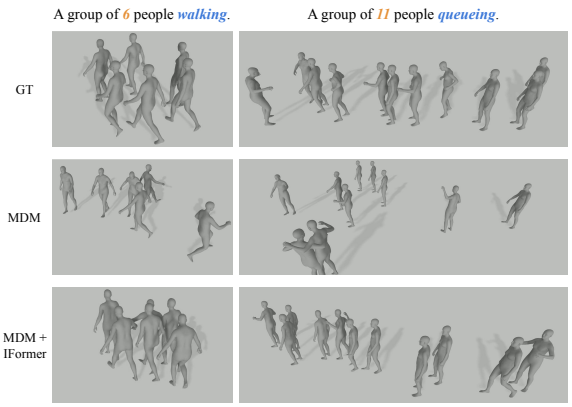

Figure 3. The qualitative comparison of two group activities from ground truth (GT), MDM, and MDM+IFormer. The distribution of the persons from MDM+IFormer is closer to GT.

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 individual persons 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 $\times$ #frames $\times$ 26 $\times$ 6). For a fair comparison, both baseline models were trained on an NVIDIA RTX 3090 graphics card with 90% data from $\text{M}^3\text{Act3D}$ for 320K iterations and then tested on the other 10%.

4.3.1 Evaluation

Metrics. Due to the probabilistic nature of the task, we consider the following learning-based metrics, recognition accuracy, Frechet Inception Distance (FID), diversity, and multimodality, defined in [26]. These metrics, however, were originally designed for single-person motion generation. To evaluate the generated group activities, we report them at both group and person levels because they account for the fidelity and variations for the groups and the individuals. We train a multi-scale group activity recognition model using the Composer [58] architecture for the metrics. See Sup. Mat. for detailed explanations of how to construct the learning-based metrics, including the recognition model as well as the latent representations at both levels.

In addition to the learning-based metrics, we tailor four position-based metrics, *collision frequency*, *repulsive interaction force*, *contact repulsive force*, and *total repulsive force*, to the evaluation of human groups. The latter three are based on the social force model [27, 28], which explains crowd behaviors using socio-psychological and physical forces. Here we describe the four metrics:

Collision frequency indicates how often a collision (or an invalid interaction) would occur within a group. The collision count is calculated based on a distance threshold between any two persons in a group. It is then normalized by the total number of interactions to obtain the frequency. In other words, if N persons are in a group, the normalization denominator is $N \cdot (N-1)/2$.

Repulsive interaction force describes the psychological tendency of two persons to stay away from each other. As the distance between two persons decreases, the repulsive force increases exponentially.

Contact repulsive force represents the compression body force when two persons collide with each other. The contact force is nonzero only when two persons collide. A larger contact force means the interaction is less likely to occur.

Total repulsive force is the accumulation of interaction and contact forces.

All four position-based metrics are calculated using the Euclidean distances of the persons' positions, with the shoulder width as the collision threshold. The social forces are calculated by averaging the magnitude of each individual's force accumulated through all its interactions with other persons. A well-performing model should generate 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-

	Group Level				Person Level		
	Acc ↑	FID ↓	Diversity →	Multimodality →	FID ↓	Diversity →	Multimodality →
GT	99.937	0.001 ± 0.000	17.752 ± 0.025	3.491 ± 0.012	0.001 ± 0.000	14.506 ± 0.013	7.546 ± 0.010
MDM	97.367	3.909 ± 0.019	17.683 ± 0.037	$\mathbf{4.155} \pm 0.019$	4.434 ± 0.010	14.158 ± 0.035	$\mathbf{7.588} \pm 0.013$
MDM+IFormer	**98.100**	$\mathbf{3.242} \pm 0.016$	17.855 ± 0.040	4.198 ± 0.021	$\mathbf{3.066} \pm 0.007$	14.827 ± 0.031	6.945 ± 0.011

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).

	Collision Freq. ↓	Interaction Force →	Contact Force →	Total Force →
GT	0.037	65.79	46.55	112.33
MDM	3.643	7,121.50	3,822.47	10,903.57
MDM+IFormer	**1.157**	**1,796.25**	**1,373.40**	**3167.80**

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.

teraction transformer. The persons are better aligned in a group and they have coordinated motions.

Both baselines are capable of generating diverse activities that match the input condition, but MDM+IFormer obtains better FID scores. Tab. 7 shows the results for the learning-based metrics. When compared with ground truth, both baselines obtain similar scores on recognition accuracy, diversity, and multimodality at both levels. The results indicate that both models successfully learn to generate distinguishable individual motions and group activities. The generated motions are also as diverse as the ground truth. The observations align with the results on action-to-motion generation in MDM [47]. MDM+IFormer receives a lower FID score than MDM, suggesting that MDM+IFormer generates group activities with higher quality.

The interaction transformer in MDM+IFormer greatly lowers the collision frequency within the generated group activities. As shown in Tab. 8, the collision frequency of the group activities generated by MDM+IFormer is much lower than the vanilla MDM. It suggests that the interaction transformer better learns the inter-person dependencies and generates more valid person interactions. In fact, we observe that the group activities generated by the vanilla MDM sometimes contain overlapping person positions. The high collision frequency of the MDM baseline also affects the repulsive forces, which makes social forces within the group activities of MDM implausible.

5. Discussion and Conclusion

We show the merit of M^3Act 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^3Act 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[1] under Grant Award Number 22STESE00001 01 01.

[1]Disclaimer. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security.

References

[1] Renderpeople. https://renderpeople.com/, 2023. Accessed: 2023-02-10. 3

[2] Eduard Gabriel Bazavan, Andrei Zanfir, Mihai Zanfir, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Hspace: Synthetic parametric humans animated in complex environments. *arXiv preprint arXiv:2112.12867*, 2021. 2, 3

[3] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *2016 IEEE international conference on image processing (ICIP)*, pages 3464–3468. IEEE, 2016. 5

[4] 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. 1, 2, 3, 5

[5] Steve Borkman, Adam Crespi, Saurav Dhakad, Sujoy Ganguly, Jonathan Hogins, You-Cyuan Jhang, Mohsen Kamalzadeh, Bowen Li, Steven Leal, Pete Parisi, et al. Unity perception: Generate synthetic data for computer vision. *arXiv preprint arXiv:2107.04259*, 2021. 3

[6] Zhongang Cai, Mingyuan Zhang, Jiawei Ren, Chen Wei, Daxuan Ren, Zhengyu Lin, Haiyu Zhao, Lei Yang, Chen Change Loy, and Ziwei Liu. Playing for 3d human recovery. *arXiv preprint arXiv:2110.07588*, 2021. 3, 5

[7] Jinkun Cao, Jiangmiao Pang, Xinshuo Weng, Rawal Khirodkar, and Kris Kitani. Observation-centric sort: Rethinking sort for robust multi-object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9686–9696, 2023. 5

[8] Che-Jui Chang. Transfer learning from monolingual asr to transcription-free cross-lingual voice conversion. *arXiv preprint arXiv:2009.14668*, 2020. 1

[9] Che-Jui Chang and Shyh-Kang Jeng. Acoustic anomaly detection using multilayer neural networks and semantic pointers. *Journal of Information Science & Engineering*, 37(1), 2021. 1

[10] Che-Jui Chang, Sen Zhang, and Mubbasir Kapadia. The ivi lab entry to the genea challenge 2022–a tacotron2 based method for co-speech gesture generation with locality-constraint attention mechanism. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, pages 784–789, 2022. 6

[11] Che-Jui Chang, Long Zhao, Sen Zhang, and Mubbasir Kapadia. Disentangling audio content and emotion with adaptive instance normalization for expressive facial animation synthesis. *Computer Animation and Virtual Worlds*, 33(3-4): e2076, 2022. 1

[12] Che-Jui Chang, Samuel S Sohn, Sen Zhang, Rajath Jayashankar, Muhammad Usman, and Mubbasir Kapadia. The importance of multimodal emotion conditioning and affect consistency for embodied conversational agents. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, pages 790–801, 2023. 1

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

[14] Wongun Choi, Khuram Shahid, and Silvio Savarese. What are they doing?: Collective activity classification using spatio-temporal relationship among people. In *2009 IEEE 12th international conference on computer vision workshops, ICCV Workshops*, pages 1282–1289. IEEE, 2009. 6

[15] Wongun Choi, Khuram Shahid, and Silvio Savarese. Learning context for collective activity recognition. In *CVPR 2011*, pages 3273–3280. IEEE, 2011. 2, 3, 6

[16] Baptiste Chopin, Hao Tang, Naima Otberdout, Mohamed Daoudi, and Nicu Sebe. Interaction transformer for human reaction generation. *IEEE Transactions on Multimedia*, 2023. 6

[17] Patrick Dendorfer, Hamid Rezatofighi, Anton Milan, Javen Shi, Daniel Cremers, Ian Reid, Stefan Roth, Konrad Schindler, and Laura Leal-Taixé. Mot20: A benchmark for multi object tracking in crowded scenes. *arXiv preprint arXiv:2003.09003*, 2020. 2, 3

[18] Andreas Doering, Di Chen, Shanshan Zhang, Bernt Schiele, and Juergen Gall. Posetrack21: A dataset for person search, multi-object tracking and multi-person pose tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20963–20972, 2022. 3

[19] Salehe Erfanian Ebadi, You-Cyuan Jhang, Alex Zook, Saurav Dhakad, Adam Crespi, Pete Parisi, Steven Borkman, Jonathan Hogins, and Sujoy Ganguly. Peoplesanspeople: a synthetic data generator for human-centric computer vision. *arXiv preprint arXiv:2112.09290*, 2021. 3

[20] Salehe Erfanian Ebadi, Saurav Dhakad, Sanjay Vishwakarma, Chunpu Wang, You-Cyuan Jhang, Maciek Chociej, Adam Crespi, Alex Thaman, and Sujoy Ganguly. Psp-hdri +: A synthetic dataset generator for pre-training of human-centric computer vision models. *arXiv preprint arXiv:2207.05025*, 2022. 1, 2, 3

[21] Mahsa Ehsanpour, Fatemeh Saleh, Silvio Savarese, Ian Reid, and Hamid Rezatofighi. Jrdb-act: A large-scale dataset for spatio-temporal action, social group and activity detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20983–20992, 2022. 1, 3

[22] Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Three-dimensional reconstruction of human interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7214–7223, 2020. 5

[23] Ruopeng Gao and Limin Wang. MeMOTR: Long-term memory-augmented transformer for multi-object tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9901–9910, 2023. 2, 5

[24] Ruohan Gao, Zilin Si, Yen-Yu Chang, Samuel Clarke, Jeannette Bohg, Li Fei-Fei, Wenzhen Yuan, and Jiajun Wu. Objectfolder 2.0: A multisensory object dataset for sim2real transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10598–10608, 2022. 2

[25] Kirill Gavrilyuk, Ryan Sanford, Mehrsan Javan, and Cees GM Snoek. Actor-transformers for group activity

recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 839–848, 2020. 2, 3, 6

[26] 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. 7

[27] Dirk Helbing and Peter Molnar. Social force model for pedestrian dynamics. *Physical review E*, 51(5):4282, 1995. 7

[28] Dirk Helbing, Illés Farkas, and Tamas Vicsek. Simulating dynamical features of escape panic. *Nature*, 407(6803):487–490, 2000. 7

[29] Mostafa S Ibrahim, Srikanth Muralidharan, Zhiwei Deng, Arash Vahdat, and Greg Mori. A hierarchical deep temporal model for group activity recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1971–1980, 2016. 3, 6

[30] Jogendra Nath Kundu, Himanshu Buckchash, Priyanka Mandikal, Anirudh Jamkhandi, Venkatesh Babu Radhakrishnan, et al. Cross-conditioned recurrent networks for long-term synthesis of inter-person human motion interactions. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2724–2733, 2020. 5

[31] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 4

[32] Weiyao Lin, Huabin Liu, Shizhan Liu, Yuxi Li, Rui Qian, Tao Wang, Ning Xu, Hongkai Xiong, Guo-Jun Qi, and Nicu Sebe. Human in events: A large-scale benchmark for human-centric video analysis in complex events. *arXiv preprint arXiv:2005.04490*, 2020. 3

[33] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(10):2684–2701, 2019. 3, 5

[34] Qing Liu, Adam Kortylewski, Zhishuai Zhang, Zizhang Li, Mengqi Guo, Qihao Liu, Xiaoding Yuan, Jiteng Mu, Weichao Qiu, and Alan Yuille. Learning part segmentation through unsupervised domain adaptation from synthetic vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19140–19151, 2022. 2

[35] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015. 4

[36] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision*, pages 5442–5451, 2019. 3

[37] Roberto Martin-Martin, Mihir Patel, Hamid Rezatofighi, Abhijeet Shenoi, JunYoung Gwak, Eric Frankel, Amir Sadeghian, and Silvio Savarese. Jrdb: A dataset and benchmark of egocentric robot visual perception of humans in built environments. *IEEE transactions on pattern analysis and machine intelligence*, 2021. 1, 3

[38] Anton Milan, Laura Leal-Taixé, Ian Reid, Stefan Roth, and Konrad Schindler. Mot16: A benchmark for multi-object tracking. *arXiv preprint arXiv:1603.00831*, 2016. 3

[39] Samarth Mishra, Rameswar Panda, Cheng Perng Phoo, Chun-Fu Richard Chen, Leonid Karlinsky, Kate Saenko, Venkatesh Saligrama, and Rogerio S Feris. Task2sim: Towards effective pre-training and transfer from synthetic data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9194–9204, 2022. 2

[40] Priyanka Patel, Chun-Hao P Huang, Joachim Tesch, David T Hoffmann, Shashank Tripathi, and Michael J Black. Agora: Avatars in geography optimized for regression analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13468–13478, 2021. 1, 2, 3

[41] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. 4

[42] Francesco Picetti, Shrinath Deshpande, Jonathan Leban, Soroosh Shahtalebi, Jay Patel, Peifeng Jing, Chunpu Wang, Charles Metze III au2, Cameron Sun, Cera Laidlaw, James Warren, Kathy Huynh, River Page, Jonathan Hogins, Adam Crespi, Sujoy Ganguly, and Salehe Erfanian Ebadi. Anthronet: Conditional generation of humans via anthropometrics. 2023. 3

[43] Md Ashiqur Rahman, Jasorsi Ghosh, Hrishikesh Viswanath, Kamyar Azizzadenesheli, and Aniket Bera. Pacmo: Partner dependent human motion generation in dyadic human activity using neural operators. *arXiv preprint arXiv:2211.16210*, 2022. 6

[44] Ziyang Song, Dongliang Wang, Nan Jiang, Zhicheng Fang, Chenjing Ding, Weihao Gan, and Wei Wu. Actformer: A gan transformer framework towards general action-conditioned 3d human motion generation. *arXiv preprint arXiv:2203.07706*, 2022. 4, 5, 6

[45] Peize Sun, Jinkun Cao, Yi Jiang, Zehuan Yuan, Song Bai, Kris Kitani, and Ping Luo. Dancetrack: Multi-object tracking in uniform appearance and diverse motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20993–21002, 2022. 3, 5

[46] Tao Sun, Mattia Segu, Janis Postels, Yuxuan Wang, Luc Van Gool, Bernt Schiele, Federico Tombari, and Fisher Yu. Shift: a synthetic driving dataset for continuous multi-task domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21371–21382, 2022. 2

[47] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022. 6, 8

[48] Gul Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 109–117, 2017. 1, 2, 3

[49] Edward Vendrow, Duy Tho Le, Jianfei Cai, and Hamid Rezatofighi. Jrdb-pose: A large-scale dataset for multi-person pose estimation and tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4811–4820, 2023. 1, 3

[50] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE international conference on image processing (ICIP)*, pages 3645–3649. IEEE, 2017. 5

[51] Li-Fang Wu, Qi Wang, Meng Jian, Yu Qiao, and Bo-Xuan Zhao. A comprehensive review of group activity recognition in videos. *International Journal of Automation and Computing*, 18:334–350, 2021. 1, 6

[52] Feng Yan, Weixin Luo, Yujie Zhong, Yiyang Gan, and Lin Ma. Bridging the gap between end-to-end and non-end-to-end multi-object tracking, 2023. 2, 5

[53] Zhitao Yang, Zhongang Cai, Haiyi Mei, Shuai Liu, Zhaoxi Chen, Weiye Xiao, Yukun Wei, Zhongfei Qing, Chen Wei, Bo Dai, et al. Synbody: Synthetic dataset with layered human models for 3d human perception and modeling. *arXiv preprint arXiv:2303.17368*, 2023. 1, 2, 3

[54] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020. 2

[55] Kiwon Yun, Jean Honorio, Debaleena Chattopadhyay, Tamara L Berg, and Dimitris Samaras. Two-person interaction detection using body-pose features and multiple instance learning. In *2012 IEEE computer society conference on computer vision and pattern recognition workshops*, pages 28–35. IEEE, 2012. 5

[56] Fangao Zeng, Bin Dong, Yuang Zhang, Tiancai Wang, Xiangyu Zhang, and Yichen Wei. Motr: End-to-end multiple-object tracking with transformer. In *European Conference on Computer Vision*, pages 659–675. Springer, 2022. 2, 5

[57] Yuang Zhang, Tiancai Wang, and Xiangyu Zhang. Motrv2: Bootstrapping end-to-end multi-object tracking by pre-trained object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22056–22065, 2023. 2, 5

[58] Honglu Zhou, Asim Kadav, Aviv Shamsian, Shijie Geng, Farley Lai, Long Zhao, Ting Liu, Mubbasir Kapadia, and Hans Peter Graf. Composer: Compositional reasoning of group activity in videos with keypoint-only modality. *Proceedings of the 17th European Conference on Computer Vision (ECCV 2022)*, 2022. 2, 3, 6, 7

[59] 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. 7