Towards Open Domain Text-Driven Synthesis of Multi-Person Motions

Anonymous ECCV 2024 Submission

Paper ID #8171

Abstract. This work aims to generate natural and diverse group motions of multiple humans from textual descriptions. While single-person text-to-motion generation is extensively studied, it remains challenging to synthesize motions for more than one or two subjects from in-the-wild prompts, mainly due to the lack of available datasets. In this work, we curate human pose and motion datasets by estimating pose information from large-scale image and video datasets. Our models use a transformer-based diffusion framework that accommodates multiple datasets with any number of subjects or frames. Experiments explore both generation of multi-person static poses and generation of multi-person motion sequences. To our knowledge, our method is the first to generate multi-subject motion sequences with high diversity and fidelity from a large variety of textual prompts.

Keywords: Human Motion Generation · Human Pose Dataset · Text-to-Motion Generation · Multi-Person Motion Generation

1 Introduction

This work presents a framework for tackling the challenge of generating multi-person motions from a natural language text prompt. Human motion modeling is a widely studied topic with applications spanning areas such as robotics, games, and VR/AR. Conventional approaches for creating computational models of human motion require artists to animate 3D assets [36] or an elaborate motion capture process [48]. Recently, human motion synthesis with generative models has seen significant progress. Text-driven motion generation [13,23,39,53,65,68,69], greatly increases the efficiency, flexibility and accessibility of motion animation. Nonetheless, prior works are limited to single-person or two-person motions, and often are not compatible with prompts that extend beyond a restricted distribution. These limitations are primarily due to available motion data. Prior works are confined to training models with single-person [13] or two-person [39] motion datasets with moderately diverse prompt distributions.

We address the challenge of multi-person motion modeling by introducing novel datasets which provide multi-person poses and motions along with text descriptions. Given the difficulties of multi-person motion capture, our strategy is instead to build upon recent advances for estimating pose and motions from image and video sources. In particular, we leverage human pose estimation

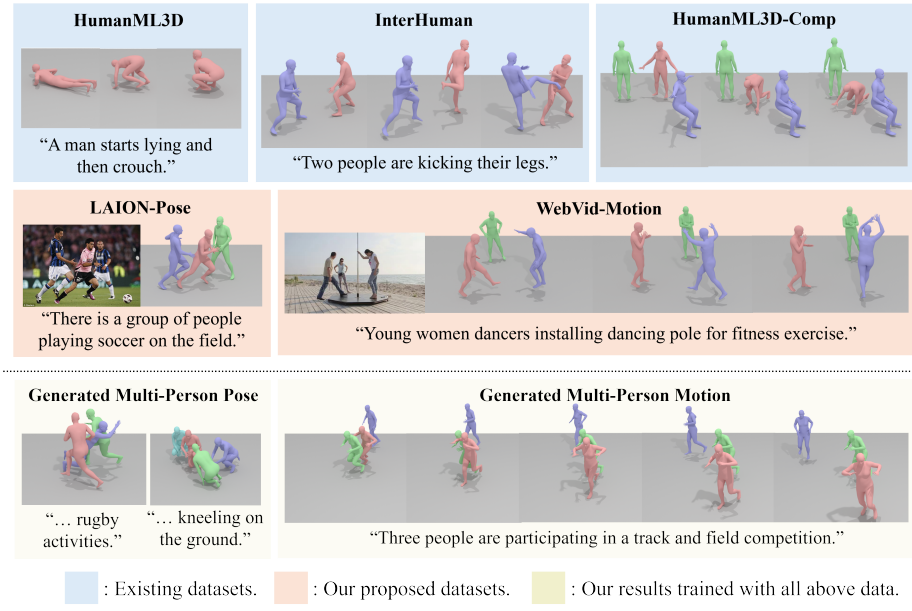

Fig. 1: We jointly train with multiple data sources including motion capture data and pose/motion extracted from image/video datasets. The model generates motion sequences from text for an arbitrary number of subjects.

methods BEV [62, 63] and TRACE [62] to extract multi-person static pose and dynamic motion from large-scale image dataset LAION-400M [59] and video dataset WebVid-10M [2]. This results in our LAION-Pose dataset with 8 million `(image, pose, text)` tuples and our WebVid-Motion dataset with 3,500 `(video, motion, text)` tuples. Sample text descriptions span a diverse variety of poses and motions that occur in open-domain images and videos.

To accommodate multiple data sources (single/multi subject, single/multi frame), we represent all samples in a common format given by the SMPL [42] parameters of each frame. Inspired by video generation architectures [21], our multi-person motion network uses interleaved pose and motion transformer encoder layers [70]. Before each pose/motion layer, the temporal/subject dimension of the sample is reshaped into the batch dimension. After each layer, the sample is reshaped to its original format. Consequently, the pose and motion layers focus on generating plausible group poses for each frame, and temporal connection among frames for each subject, respectively.

We adopt a denoising diffusion framework for motion modeling [20, 69]. Our proposed method uses a two-stage pipeline. The first stage model produces a single frame containing the poses of multiple people. This sample is used as a condition for the second stage model, which will generate a motion sequence that has the first-stage pose sample as the middle frame. Both stages use the same text prompt for generation. This achieves the effect of generating a pose and animating it over time. During sampling time, we also use pose and single-

person motion models as guidance terms to further boost the quality of the multi-person motion results.

Models are evaluated in a decomposed manner by measuring the quality of each multi-person frame and each single-person motion sequence. We train feature extractors for pose and motion in the SMPL format with LAION-Pose and HumanML3D [13] data following CLIP training [57].

Our main contributions can be summarized as follows:

- We present the first model to generate multi-person motion sequences given open-domain textual prompts with an arbitrary number of subjects.
- We introduce LAION-Pose and WebVid-Motion, the first large-scale 3D datasets of text-annotated multi-person poses/motions collected from in-the-wild images and videos.
- We design a factorized method to evaluate our results in the absence of ground-truth multi-person motion data by building contrastive encoder backbones for text and pose/motion, and demonstrate that our method outperforms existing methods qualitatively and quantitatively.

2 Related Works

Human Mesh Recovery. Human mesh recovery aims to jointly estimate 3D human body poses and shapes from images or videos. Recent advancements in body parametric models, such as SMPL [42], facilitate such development in diverse human-centric tasks including 3D human body reconstruction [25], animation [80], and pose estimation [30,82]. A variety of works explore single-person human body recovery from single images [12, 26, 30–33, 37, 40, 77, 78, 83, 86] and videos [5, 8, 27, 29, 38, 64, 75, 82], as well as multi-person mesh recovery scenarios [6, 10, 28, 56, 58, 61–63, 79, 84]. In particular, ROMP [61] regresses multiple 3D meshes along with the horizontal/vertical locations and a rough depth estimate from a single image. BEV [63] improves upon ROMP by enabling more precise multi-person depth estimation. TRACE [62] extends the setup of BEV and supports motion regression with frame-wise consistency from a dynamic camera video. Our paper builds upon 3D human body models and estimation tools to recover multi-person mesh poses from large-scale image and video datasets.

Conditional Human Motion Generation. Generating realistic 3D human motions with conditional signals is an active research area. Reference motions can be used as conditions to predict future motions [3, 11, 19, 46, 74], intermediate motions [9, 17, 18], and motion variations [35]. Motion generation can also be conditioned by action class [52], audio [49, 50], music [34, 43, 90], and natural language, which is the focus of this work. Text-conditioned motion generation can be accomplished by a joint embedding of language and pose [68], VAE structure [13, 53, 85], or VQ-VAE [14, 23, 87]. Several recent works apply diffusion models [20] for text-to-motion generation [4, 69, 81, 88]. We build upon the MDM [69] framework that uses transformer encoders and extend the architecture to accommodate the joint training scenario with more than one subject.

Multi-person Motion Generation. Despite the success of single-person motion models, generating group motions of more than one person remains challenging. One direction of research explores multi-subject motion prediction based on past 3D skeletons [15,44,67,72]. In the area of text-driven multi-person motion, ComMDM [60] proposed to use a generative prior coupled with a streamlined communication block to facilitate coordinated interaction. RIG [66] attempts to recover 3D interaction motions from a noisy depth dataset [41] and translates the motion labels into sentences. InterGen [39] introduces two collaborative transformer-based denoisers with a mutual attention mechanism. InterControl [73] enables flexible spatial control over each joint to generate plausible interactions. These works focus predominantly on interactions involving only two individuals and are often incompatible with open domain text prompts. Our model generates multi-person motion directly from textual descriptions for one, two, and more than two people, while also exhibiting flexible text generalization abilities learned from in-the-wild datasets.

Human Motion Datasets. Datasets play a crucial role in propelling motion generation research forward. Action label datasets [41, 55] and text annotation datasets [13,54] support the advancement of single-person motion generation [4]. However, models trained with single-motion datasets will have difficulty generalizing to group motions. Meanwhile, various multi-person motion datasets have been developed including 4DAssociation [89], UMPM [1], 3DPW [71], MuPoTS-3D [47], ExPI [15], CMU-Panoptic [24], You2Me [51]. Among these datasets, only 3DPW is compatible with the SMPL format which is a common starting point of current motion generation models. ComMDM [60] contributed textual annotations to 3DPW, but the resulting dataset contains only 27 two-person motion sequences. InterHuman [39] is a recently introduced dataset with diverse multi-person motions with textual annotations. However, it remains constrained to interactions between two individuals. To the best of our knowledge, we contribute the first large-scale datasets of text-annotated 3D poses and motions with more than two people.

3 Dataset

To address the data scarcity problem in multi-person domain, we introduce two datasets: LAION-Pose and WebVid-Motion, containing `(image, pose, text)` and `(video, motion, text)` tuples extracted from in-the-wild images and videos. We present our data processing techniques of LAION-Pose in Sec. 3.1 and of WebVid-Motion in Sec. 3.2. See Fig. 2 for selected examples.

3.1 LAION-Pose

We apply BEV [63] to each image in LAION-400M. BEV estimates human global translation, SMPL joint rotation, and SMPL shape information for an arbitrary number of people in a monocular image. If BEV cannot detect a human occurrence in an image, the image sample is ignored. BEV often provides reasonable

Fig. 2: Dataset visualizations. Top 2 rows: LAION-Pose dataset. Left is original image from LAION-400M [59], right is BEV [63] detection. Bottom 2 rows: Webvid-Motion dataset. Left is original video first frame from WebVid-10M [2], right is the motion sequence estimated by TRACE [62] visualized from a different camera angle.

mesh estimates for images with clearly visible poses. Nonetheless there are also many circumstances where it does not produce satisfactory results. Subsequent filtering steps are performed to retain predominantly BEV meshes that accurately predict ground truth poses in an image. The filtering stages start with coarse criteria and become progressively more fine.

- **Person Detection.** Detectron-2 [76] is used to identify whether people are in an image. Samples without any positive detections are removed.
- **Mesh Completion.** We retain all meshes such that over 85% of the mesh falls in the image frame and discard those predictions where human poses are only partially contained within the image frame.
- **Mesh De-duplication.** BEV sometimes predict duplicate, overlaid meshes from the monocular image perspective. Duplicates are detected by projecting the mesh vertices to the image plane and measuring the overlap. One sample in each pair with over 25% IoU overlap is discarded.
- **Hand-Crafted Prompt Filter.** We hand-craft a set of filtering prompts such as *"a DVD cover"*. We discard all image samples whose CLIP [57] similarity with a filter prompt is higher than a threshold
- **Few-Shot Filter.** After the previous filtering steps, we manually annotate 1000 random data samples with binary labels indicating whether the BEV prediction were a visually acceptable match to the reference image. A logistic regression classifier is then trained using the quality annotations and the CLIP image features of the corresponding images. We set a threshold that such that 90% of model selections in a validation set are annotated as high-quality to ensure that mostly good samples are kept.

After data filtering, we are left with around 8 million `(image, pose, text)` triplets which are the raw form of our dataset. The final dataset is created by refining pose samples and generating more informative text captions.

- **Vertical Height Adjustment.** BEV faces inherently limitations when estimating spatial relationships among multiple objects due to monocular ambiguity. In most cases, editing group poses so that all individuals have a consistent vertical height preserves the essence of the pose description while making group pose appearance more natural. A smaller dataset without height adjustment is also kept to learn cases where relative heights are meaningful.
- **Mesh Separation.** We correct mesh overlap by optimizing each pair of SMPL parameters to minimize a mesh collision loss. We compute the degree of overlap between each pair as the sum of SDF (signed distance function) from all vertices in the first mesh to the second mesh, and then run gradient descent to back-propagate through the SMPL layers to optimize for a fixed number of 25 steps. The resulting meshes have no or very little overlap.
- **InstructBLIP Captioning.** LAION-400M [59] captions are not informative for pose and motion modeling because they almost always focus on appearance, style, or factual metadata of the image instead of human actions. We use InstructBLIP [7] instead as an image captioning tool. We prompt InstructBLIP with the request *"describe the person or group of people's action and body poses in the image"*. This simple instruction yields robust captions that are almost always more suitable pose and action descriptions than the LAION-400M text captions.

3.2 WebVid-Motion

Person Detection. We first filter WebVid-10M [59] by applying Detectron-2 [76] on the middle frame of each video, and only keep videos with at least two detected humans. This works in most cases since WebVid-10M videos tend to have limited camera movement and scene changes.

TRACE Estimation. Next, we apply TRACE [62] to filtered videos to predict per frame global translation, joint rotation, and shape information for an arbitrary number of people in a single-view video with dynamic camera. We apply a Gaussian filter with scale $s = 1$ to smooth out the motion jittering effects.

Motion Grouping. According to the temporal nature of video, there might be multiple subjects appearing and disappearing through time, and there might be multiple useful motion clips in one video sample. To obtain multi-person motion samples from the raw TRACE outputs, we select all clips where at least two people appear in the video simultaneously for at least 30 frames.

Manual Selection. The data is manually inspected to select 3,500 samples with the best fidelity and the most interesting group or interactive behavior. Manual selection filters out motions with highly unrealistic translations due to the ambiguity of monocular video, and removes duplicated and static motions.

InstructBLIP Captioning. Similar to the image case, we use InstructBLIP [7] as a video captioning tool by instructing it to *"describe the person or group of people's action and body poses in the image"* for the middle frame of the video.

4 Model

Our goal is to generate high-quality multi-person motion sequences given a textual description. We leverage a transformer architecture made of interleaved pose and motion layers that facilitate plausible frame-wise poses and temporal movements, respectively. Sec. 4.1 introduces the representation we use for pose and motion. Sec. 4.2 describe the design of our motion and pose layers as well as the two-stage joint training architecture. Sec. 4.3 presents our sampling strategy with specific guidance term. The model overview is shown in Fig. 3.

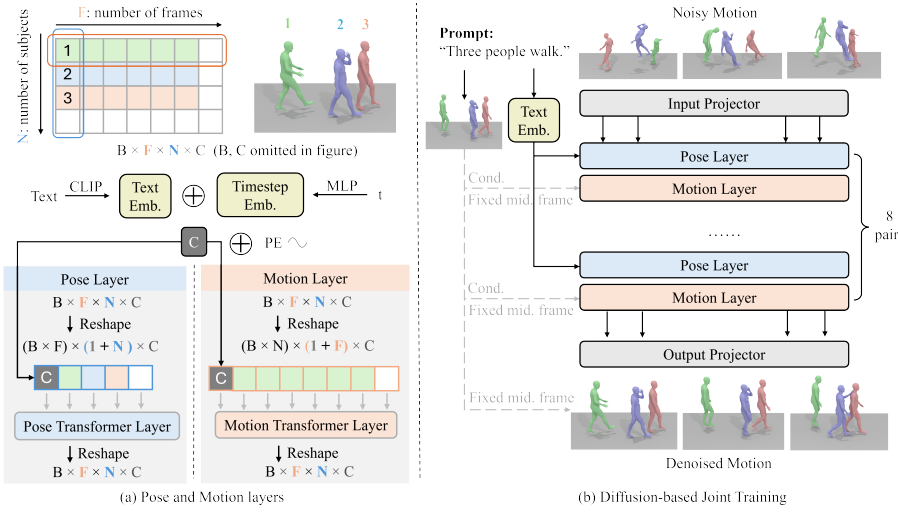

Fig. 3: Our model is a diffusion framework consisting of interleaving pose and motion layers. At each pose/motion layer, we reshape the temporal/subject dimension into the batch dimension so that the layer focuses on generating per frame subject interaction and per-subject temporal movements respectively. Each layer is implemented as a transformer encoder. Diffusion time steps and text or pose conditions are encoded and summed up as a condition token concatenated to the beginning of the sequence.

4.1 Motion Representation

We base our motion representation on the widely used SMPL [42] format, which describe the human body in 24×3 SMPL θ parameters for pose and 11 SMPL β parameters for motion. For training stability, we follow common practice and turn the axis angle rotation vectors into 6D representation, and add in a global translation $t \in \mathbb{R}^3$ to model movements and spatial relationship among multiple subjects. A single person pose $y \in \mathbb{R}^{158}$ is defined to be the concatenation of $\beta \in \mathbb{R}^{11}$, $\theta \in \mathbb{R}^{24\times6}$ and $t \in \mathbb{R}^3$. A multi-person motion is defined as a set of poses $x = \{y^{n,f}\}_{n=1}^{N}{}_{f=1}^{F}$ where $F \geq 1$ and $N \geq 1$ are the number of frames and people that comprise the motion. In practice, differences in motion lengths/number of poses are handled by padding to a maximum number of frames/poses and masking attention for pad tokens. Special cases include $F = 1$, which corresponds

to a multiple poses over a single frame, and $N = 1$, which corresponds to single-person motion.

Note that we refrain from using the popular HumanML3D [13] format for two reasons. First, it contains velocity which is meaningless for pose data and thus not adaptable to our joint training strategy. Second, it canonicalizes joint positions and velocities to the root frame, and thus lose spatial interaction information for multi-person scenarios (as also mentioned in [39]).

4.2 Training

Diffusion Model Objective. Denoising diffusion probabilistic [20] models are a class of generative models that aim to approximate a data distribution through a progressive denoising process. Our forward diffusion procedure can be modeled as a Markov noising process, $\{x_t^{1:N,1:F}\}_{t=0}^{T}$ where t is the diffusion time step, $x_0^{1:N,1:F}$ is drawn from the training data distribution, and

$$q\left(x_t^{1:N,1:F} \mid x_{t-1}^{1:N,1:F}\right) = \mathcal{N}\left(\sqrt{\alpha_t} x_{t-1}^{1:N.1:F}, (1-\alpha_t)\, I\right)$$

where $\alpha_t \in (0, 1)$ are constants. Our motion generation model is defined as learning the reverse process of the Markov Chain, gradually cleaning up x_T to get $p(x_0 \mid c)$ with condition c. Following MDM [69], we design our model to predict the original signal x_0 instead of noise ϵ_0 with the simple loss in [20]:

$$\mathcal{L} = E_{x_0 \sim q(x_0|c), t\sim[1,T], x_t \sim q_t(x_t|x_0,c)} \left[\|x_0 - G(x_t, t, c)\|_2^2\right]$$

where G is the motion generation model described in the next section and q_t give the forward process condition distributions. During training we randomly mask out the textual condition by setting the textual embedding to zero for 10% of samples to enable classifier-free guidance at inference time.

Pose and Motion Layers. The components of our model follow designs proposed by MDM [69]. Each layer is implemented in a straightforward transformer [70] encoder architecture. The transformer layer can used masked attention to remove the influence of padded tokens and thus can handle motions of arbitrary lengths or with arbitrary number of subjects. Pose layers are applied frame-wise, learning relations between poses and locations of multiple subjects for each single frame. Motion layers are applied subject-wise, learning plausible temporal movement of each person. We project the noise time-step t and the CLIP [57] embedding of text to the transformer dimension by separate feed-forward networks, then sum them up to yield the condition token. The token is then appended to the start of each sequence.

Joint Architecture. Our joint training architecture (Fig. 3 right) uses pairs of pose and motion layers. We start with data padded to the shape $B \times F \times N \times C$ where B is the batch size, F is the maximum number of frames, N is the maximum number of subjects, and $C = 158$ is the pose channel. We apply a linear layer that projects the data into a high dimension ($C' = 512$). The batch

data then goes through each of the pose and motion layers in order. Note that we only apply text condition to pose layers so that single-frame and multi-frame samples have a consistent text conditioning mechanism.

Before entering a pose layer, we reshape the temporal dimension into the batch dimension so that the data shape becomes $(B \times F) \times N \times C'$. We apply pose layer to the subject (N) dimension so that it is able to learn the per-frame interaction among multiple subjects. Similarly, before entering a motion layer, we reshape the subject dimension into the batch dimension so that the data shape becomes $(B \times N) \times F \times C'$. We apply motion layer to the temporal (F) dimension so that it learns the per-subject movements through time. After each transformer encoder layer, we discard the first output token which corresponds to the conditional signal, and feed the rest into the next layer with corresponding reshaping. After the last layer, we reshape back to $B \times F \times N \times C'$ and project back to the pose dimension C with a linear layer. For simplicity, we omit the additional condition token in the above discussion of reshaping.

Our joint architecture takes both text and a reference pose frame as conditions. The reference pose represents the middle frame of the sequence being generated and the joint model is intended to extend the reference frame forward and backward in time. The pose conditioning is implemented in an analogous way to the text conditioning of pose layers. The reference pose is concatenated with the positional embedding of the middle frame, fed through an MLP, and added to a separate copy of the timestep embedding vector. This embedding is appended as the first token before motion layers.

Two Stage Training. We train the model in a two-stage manner. First, we train a text-to-pose model with only single-frame multi-person pose data. This model follows the structure of MDM, except for lack of positional encoding. After training, the model can generate single frame pose samples which will be animated over time. The multi-person motion model is initialized by inserting a motion layer after each pose layer and freezing the pose layers. It takes both text and the the middle frame of data motions as conditions during training. Single-frame inputs are assigned a null pose condition, so that the new motion layers can still learn from single-frame samples. Our approach is inspired by similar techniques in the video generation literature such as AnimateDiff [16] which insert new temporal layers between frozen spatial layers to preserve certain behaviors of the spatial layers over time. In our case, the we want to preserve the flexible text conditioning learned from our large-scale pose dataset.

4.3 Sampling

Diffusion Model Sampling. We sample from our model in an iterative manner following DDPM [20]. Given a time step t at sampling time, we predict the starting data $\hat{x}_0 = G(x_t, t, c)$ and noise it back to x_{t-1}. We repeat this process until we reach $t = 0$ from $t = T$. Classifier-free guidance is used to encourage better text alignment by modifying model predictions with a guidance scale $s > 1$ according to $G_s(x_t, t, c) = G(x_t, t, \emptyset) + s \cdot (G(x_t, t, c) - G(x_t, t, \emptyset))$.

Two Stage Sampling. At inference time, we draw samples from our model in two stages. A single-frame multi-person sample is drawn from our pose-only model given a text prompt. The text prompt and the single-frame pose are then used to generate a multi-person motion where the single-frame pose condition condition serves as the center motion frame.

Pose/Motion Guidance. We additionally leverage separate text-to-pose and text-to-single-person motion models as guidance terms for our multi-person generation. It is relatively straightforward to train single-frame or single-person models with high-quality results, and we may push our multi-person motion generations towards them frame-wise and/or subject-wise. With scale $s_p \geq 0$ and $s_m \geq 0$ for pose and motion models, the guided model can be expressed as

$$G_{s_p,s_m}(x_t, t, c) = (1 - s_p - s_m) \cdot G(x_t, t, c) + s_p \cdot G_p(x_t, t, c) + \mathrm{s}_m \cdot G_m(x_t, c)$$

where $s_p + s_m \leq 1$ and G_p, G_m are separate text-to-pose and unconditional single-person motion generators, with frame/subject dimensions put along the batch dimension for pose/motion models respectively. We can easily control the scales to make the results better in motion quality (higher s_m), or better in subject interaction (higher s_p).

5 Experiments

5.1 Dataset

We use five datasets in our joint training: one pose and four motion datasets. All samples are shifted by a global translation such that the average global coordinate of each person in the horizontal plane is the origin for the center motion frame. Samples are augmented by randomly rotating the group motions around the vertical axis passing through the origin.

LAION-Pose (Sec. 3.1) has 8 million pairs of multi-person poses and texts.

WebVid-Motion (Sec. 3.2) has 3,500 pairs of multi-person motions and texts.

HumanML3D [13] is the most widely used text-motion dataset containing 14,616 single-person motions. We discard the HumanAct12 subset and use it in the original AMASS [45] SMPL [42] format as explained in Sec. 4.1.

HumanML3D-Comp is a synthetic dataset that we compose by arbitrarily selecting 2 to 6 motion sequences from HumanML3D and putting them together in a 3D space with randomly initialized starting translation. Generated samples are checked to ensure there is no geometric collision in meshes between subjects. Since there is no text describing the resulting group motion, they are paired with empty text as unconditional multi-person motion samples.

InterHuman [39] is a two-subject interactive motion dataset with a diverse range of around 8,000 text-annotated motions.

5.2 Implementation Details

Our model is trained with maximum 61 frames and 10 subjects. The dataset is an amalgamation of the data sources with a split ratio (LP 50%, WVM 10%,

HML 15%, HML-C 10%, IH 15%). First stage training uses a random data frame while second stage training uses up to 61 frames from a motion sequence. Longer sequences are randomly truncated. Each pose and motion layer is implemented as a transformer encoder layer with latent dimension of 512. The first stage model consists of 8 pose layers and the second stage model consists of 8 pairs of pose and motion layers. The input and output projectors are linear layers. The first stage model is trained on 4 GPUs for 500k steps over 1 day and the second stage model is trained on 8 A100 GPUs for 250K steps over 2 days.

5.3 Evaluation Setting

Number of Poses Determination. We use language model Mistral [22] to decide the number of subjects from a given textual description by asking it *"how many subjects appear in this description"*.

Decomposed Evaluation. Existing motion generation models typically evaluate their results with the text-motion shared embedding and metrics proposed by [13]. See the supplementary material for a description of each metric. In the shortage of ground truth data for training a multi-person motion encoder, we choose to implement a decomposed evaluation mechanism. Specifically, we evaluate our generative metrics on each single frame's multi-person pose result as well as each single subject's single-person motion result. We train two feature encoders with `(text, pose)` pairs from LAION-Pose and `(text, motion)` pairs from HumanML3D in the original SMPL format following CLIP [57] training. In practice, to speed up the evaluation we only calculate pose metrics on sparse frames that are separated by 14 frames, yielding 5 evaluation points that are the first, last, and middle quarters of the frame sequence.

We use a withheld validation set of LAION-Pose samples for pose metric and as evaluation prompts covering a wide range of natural descriptions. R-Precision and Similarity are not applicable for motions generated from LAION-Pose prompts because the motion encoder trained with HumanML3D is not compatible with LAION-Pose prompts. Instead, we primarily focus on motion FID calculated using HumanML3D reference motions. This serves as a rough measure of motion realism that describes how well the distribution of generated motions match the distribution of ground truth single-person motions.

Baselines. To our knowledge, there is no existing work that generates motion with an arbitrary number of subjects to compare with. As naive baselines, we use *Pose-Only* models which lack temporal connections or *Motion-Only* model which lack pose connections. The baselines both provide evidence of importance of joint modeling as well as provide rough upper bounds for the quality metrics that are achievable for frame-wise and person-wise generation in our setup. In partciular, our *Pose-Only* baseline takes a sample from our first stage model and keeps the pose static throughout the whole sequence. As a *Motion-Only* baseline, we fix the middle frame as the pose result from the first stage, and animate each single subject independently with a model containing only motion layers. We additionally experiment with sampling two subjects motion to compare with two-person motion baselines RIG [65], InterGen [39] and ComMDM [60].

5.4 Qualitative Evaluation

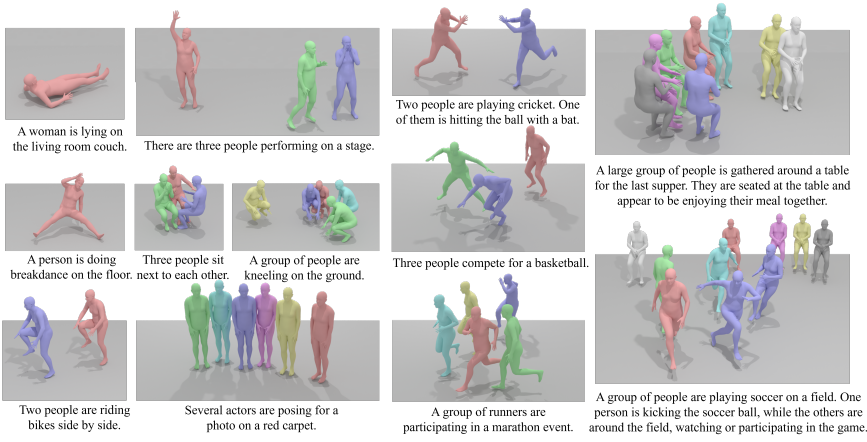

Fig. 4: Qualitative result for text-to-pose generation.

Pose Generation. Our first stage text-to-pose model can generate diverse, realistic human poses given text prompts shown in Fig. 4. Note that it works well on group prompts out of the distribution of existing motion datasets. Our high quality pose generation module paves the first step for a successful motion generation and pose animation model by providing the fixed middle frame, condition for motion layers, and optionally additional sampling guidance.

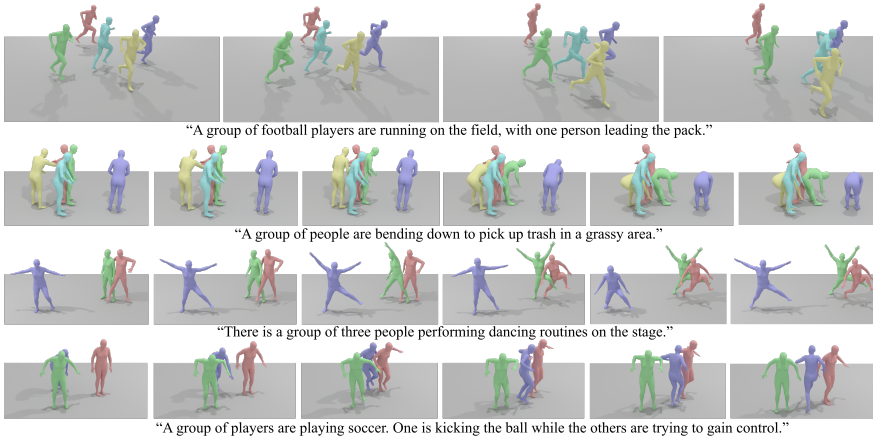

Fig. 5: Qualitative results for text-to-motion generation.

Motion Generation. In Fig. 5, we visualize a few motion sequences our model generates with the given text prompt. Results show that our model is able to generate high quality motion and interaction for various number of subjects given a wide variety of prompts. In Fig. 6 we compare our model with baseline results by restricting the sampling subjects to two. For two-person motion, we are

able to generate much more realistic motion than RIG [65] and ComMDM [60], as we take advantage of multi-person pose and motion data in our training instead relying heavily on single-person priors and adding interactive terms. Our model also works with more diverse prompts than InterGen [39], benefiting from our joint training strategy that enables training with both high-quality studio captured data as well as in-the-wild motion regression data. Readers are encouraged to view the supplementary videos for the animated results.

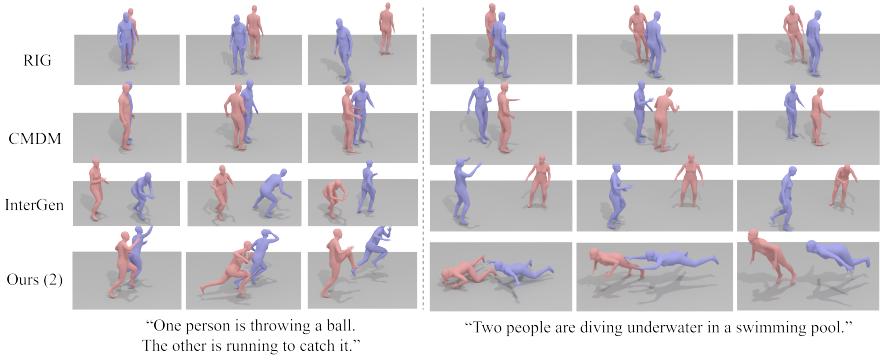

Fig. 6: Qualitative comparison with 2-person motion generation baselines.

5.5 Quantitative Evaluation

Tab. 1 shows our quantitative results comparing with the naive *Pose-Only* and *Motion-Only* baselines for single-person motion and multi-person pose. Results present the *Pose-Only* and *Motion-Only* results as the upper bounds of performance when solely focusing on per-frame pose quality and per-subject motion quality. Our joint training solution makes a balance between these two extremes, which is able to maintain per frame pose interaction plausibility while adding temporal smoothness throughout the sequences.

Table 1: Quantitative metrics comparing our multi-person results with naive baselines. Metrics for real data are evaluated with LAION-Pose and HumanML3D.

Methods	P-R-Precision ↑			P-FID ↓	P-Sim ↑	P-Div →	P-MM ↑	M-FID ↓	M-Div →	M-MM ↑
	Top-1	Top-2	Top-3							
Data	0.621	0.737	0.819	0.000	0.378	1.366	-	0.000	1.342	-
Pose-Only	0.678	0.822	0.885	0.077	0.371	1.368	0.903	0.976	1.006	0.667
Motion-Only	0.202	0.307	0.380	0.317	0.157	1.200	0.837	0.613	1.274	0.894
Ours	0.539	0.704	0.776	0.229	0.304	1.329	0.894	0.684	1.220	0.833

Tab. 2 illustrates how our model, when restricted to only two-person motion generation, is able to produce higher-quality 2-person motion results than baselines when conditioned with open-domain text prompts. This is mainly because of the limited distribution of prompts baselines are trained on, and their poor ability to generalize to more in-the-wild textual descriptions that are not constrained by the motion capture studio context.

Table 2: Quantitative metrics comparing our 2-person results with baselines. The pose and motion metrics for real data are evaluated with LAION-Pose and InterHuman [39].

Methods	P-R-Precision ↑			P-FID ↓	P-Sim ↑	P-Div →	P-MM ↑	M-FID ↓	M-Div →	M-MM ↑
	Top-1	Top-2	Top-3							
Data (2)	0.599	0.722	0.783	0.000	0.378	1.308	-	0.204	1.218	-
Pose-only (2)	0.567	0.754	0.831	0.106	0.382	1.300	0.894	0.953	1.149	0.680
Motion-Only (2)	0.162	0.260	0.339	0.520	0.191	1.164	0.602	0.558	1.270	0.811
InterGen [39]	0.073	0.126	0.176	1.038	0.113	0.880	0.643	0.734	1.190	0.778
RIG [65]	0.037	0.072	0.103	1.376	0.061	0.639	0.501	0.925	1.078	0.596
ComMDM [60]	0.043	0.085	0.124	1.160	0.080	0.819	0.714	0.821	1.109	0.765
Ours (2)	0.323	0.480	0.591	0.435	0.271	1.329	0.856	0.667	1.213	0.803

5.6 Ablation Study

We present ablation studies in Tab. 3 to validate our multiple design choices.

Table 3: Ablation results for different design choices. Bold is the best score.

Experiments	P-R-Precision ↑			P-FID ↓	P-Sim ↑	P-Div →	P-MM ↑	M-FID ↓	M-Div →	M-MM ↑
	Top-1	Top-2	Top-3							
Data	0.621	0.737	0.819	0.000	0.378	1.366	-	0.002	1.342	-
A: One stage	0.386	0.536	0.638	0.407	0.238	1.298	0.802	0.671	1.182	0.745
B: w/o freeze pose	0.202	0.307	0.380	0.503	0.157	1.199	0.734	0.613	1.274	0.830
C: w/o WebVid	0.383	0.534	0.630	0.399	0.246	1.254	0.711	**0.405**	**1.315**	0.698
D: w/o motion text	0.313	0.451	0.542	0.368	0.213	1.264	0.749	0.632	1.289	0.804
Ours	**0.539**	**0.704**	**0.776**	**0.229**	**0.304**	**1.329**	**0.894**	0.684	1.220	**0.833**

Architecture. A one-stage model without pre-trained pose results as a condition (A) does not work as well as the two-stage counterpart, because a high-quality pose model can provide strong condition signal to make each frame more plausible. Not freezing the pre-trained pose layers in the joint training stage (B) also does not perform as well, showing that motion data might interfere with the high-quality pose representation learned by pose layers during the first stage.
Data. Training without WebVid-Motion data (C) shows a downgrade in text-pose alignments, indicating that multi-person motion data helps improving interaction qualities. Removing all motion texts (D), essentially training the temporal layers unconditionally, also decreases the alignment between pose and text.

6 Conclusion

In this work, we propose the first open-domain text-driven multi-person motion generation algorithm. We design a diffusion-based joint training mechanism with interleaved pose and motions layers, which can utilize multiple data sources simultaneously. We demonstrate diverse and realistic motion generation results qualitatively and quantitatively superior to baseline methods. We additionally contribute two datasets LAION-Pose and WebVid-Motion, opening more possibilities for future investigation in the field of multi-person motion generation.

References

1. Van der Aa, N., Luo, X., Giezeman, G.J., Tan, R.T., Veltkamp, R.C.: Umpm benchmark: A multi-person dataset with synchronized video and motion capture data for evaluation of articulated human motion and interaction. In: 2011 IEEE international conference on computer vision workshops (ICCV Workshops). pp. 1264–1269. IEEE (2011) 4
2. Bain, M., Nagrani, A., Varol, G., Zisserman, A.: Frozen in time: A joint video and image encoder for end-to-end retrieval. In: IEEE International Conference on Computer Vision (2021) 2, 5
3. Barsoum, E., Kender, J., Liu, Z.: Hp-gan: Probabilistic 3d human motion prediction via gan. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 1418–1427 (2018) 3
4. Chen, X., Jiang, B., Liu, W., Huang, Z., Fu, B., Chen, T., Yu, G.: Executing your commands via motion diffusion in latent space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18000–18010 (2023) 3, 4
5. Choi, H., Moon, G., Chang, J.Y., Lee, K.M.: Beyond static features for temporally consistent 3d human pose and shape from a video. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1964–1973 (2021) 3
6. Choi, H., Moon, G., Park, J., Lee, K.M.: Learning to estimate robust 3d human mesh from in-the-wild crowded scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1475–1484 (2022) 3
7. Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning (2023) 6
8. Doersch, C., Zisserman, A.: Sim2real transfer learning for 3d human pose estimation: motion to the rescue. Advances in Neural Information Processing Systems **32** (2019) 3
9. Duan, Y., Shi, T., Zou, Z., Lin, Y., Qian, Z., Zhang, B., Yuan, Y.: Single-shot motion completion with transformer (2021) 3
10. Fieraru, M., Zanfir, M., Szente, T., Bazavan, E., Olaru, V., Sminchisescu, C.: Remips: Physically consistent 3d reconstruction of multiple interacting people under weak supervision. Advances in Neural Information Processing Systems **34**, 19385–19397 (2021) 3
11. Fragkiadaki, K., Levine, S., Felsen, P., Malik, J.: Recurrent network models for human dynamics. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4346–4354 (2015). https://doi.org/10.1109/ICCV.2015.494 3
12. Goel, S., Pavlakos, G., Rajasegaran, J., Kanazawa, A., Malik, J.: Humans in 4d: Reconstructing and tracking humans with transformers. arXiv preprint arXiv:2305.20091 (2023) 3
13. Guo, C., Zou, S., Zuo, X., Wang, S., Ji, W., Li, X., Cheng, L.: Generating diverse and natural 3d human motions from text. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5152–5161 (June 2022) 1, 3, 4, 8, 10, 11
14. Guo, C., Zuo, X., Wang, S., Cheng, L.: Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In: European Conference on Computer Vision. pp. 580–597. Springer (2022) 3

15. Guo, W., Bie, X., Alameda-Pineda, X., Moreno-Noguer, F.: Multi-person extreme motion prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13053–13064 (2022) 4
16. Guo, Y., Yang, C., Rao, A., Liang, Z., Wang, Y., Qiao, Y., Agrawala, M., Lin, D., Dai, B.: Animatediff: Animate your personalized text-to-image diffusion models without specific tuning (2023) 9
17. Harvey, F.G., Pal, C.: Recurrent transition networks for character locomotion. In: ACM SIGGRAPH Asia 2018 Technical Briefs. Association for Computing Machinery, New York, NY, USA (2018). https://doi.org/10.1145/3283254.3283277, https://doi.org/10.1145/3283254.3283277 3
18. Harvey, F.G., Yurick, M., Nowrouzezahrai, D., Pal, C.: Robust motion inbetweening. ACM Transactions on Graphics (TOG)) **39**(4) (2020). https://doi.org/10.1145/3386569.3392480 3
19. Hernandez, A., Gall, J., Moreno, F.: Human motion prediction via spatio-temporal inpainting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 7133–7142 (2019). https://doi.org/10.1109/ICCV.2019.00723 3
20. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. arXiv preprint arxiv:2006.11239 (2020) 2, 3, 8, 9
21. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. arXiv:2204.03458 (2022) 2
22. Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L.R., Lachaux, M.A., Stock, P., Scao, T.L., Lavril, T., Wang, T., Lacroix, T., Sayed, W.E.: Mistral 7b (2023) 11
23. Jiang, B., Chen, X., Liu, W., Yu, J., Yu, G., Chen, T.: Motiongpt: Human motion as a foreign language. arXiv preprint arXiv:2306.14795 (2023) 1, 3
24. Joo, H., Simon, T., Cikara, M., Sheikh, Y.: Towards social artificial intelligence: Nonverbal social signal prediction in a triadic interaction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10873–10883 (2019) 4
25. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: Computer Vision and Pattern Recognition (CVPR) (2018) 3
26. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7122–7131 (2018) 3
27. Kanazawa, A., Zhang, J.Y., Felsen, P., Malik, J.: Learning 3d human dynamics from video. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5614–5623 (2019) 3
28. Khirodkar, R., Tripathi, S., Kitani, K.: Occluded human mesh recovery. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1715–1725 (2022) 3
29. Kocabas, M., Athanasiou, N., Black, M.J.: Vibe: Video inference for human body pose and shape estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5253–5263 (2020) 3
30. Kocabas, M., Huang, C.H.P., Hilliges, O., Black, M.J.: Pare: Part attention regressor for 3d human body estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11127–11137 (2021) 3
31. Kolotouros, N., Pavlakos, G., Black, M.J., Daniilidis, K.: Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2252–2261 (2019) 3

32. Kolotouros, N., Pavlakos, G., Daniilidis, K.: Convolutional mesh regression for single-image human shape reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4501–4510 (2019) 3
33. Kolotouros, N., Pavlakos, G., Jayaraman, D., Daniilidis, K.: Probabilistic modeling for human mesh recovery. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11605–11614 (2021) 3
34. Le, N., Pham, T., Do, T., Tjiputra, E., Tran, Q.D., Nguyen, A.: Music-driven group choreography. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8673–8682 (2023) 3
35. Li, P., Aberman, K., Zhang, Z., Hanocka, R., Sorkine-Hornung, O.: Ganimator: Neural motion synthesis from a single sequence. ACM Transactions on Graphics (TOG) **41**(4), 1–12 (2022). https://doi.org/10.1145/3528223.3530157 3
36. Li, Y., Takehara, H., Taketomi, T., Zheng, B., Nießner, M.: 4dcomplete: Non-rigid motion estimation beyond the observable surface. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12706–12716 (2021) 1
37. Li, Z., Liu, J., Zhang, Z., Xu, S., Yan, Y.: Cliff: Carrying location information in full frames into human pose and shape estimation. In: European Conference on Computer Vision. pp. 590–606. Springer (2022) 3
38. Li, Z., Xu, B., Huang, H., Lu, C., Guo, Y.: Deep two-stream video inference for human body pose and shape estimation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 430–439 (2022) 3
39. Liang, H., Zhang, W., Li, W., Yu, J., Xu, L.: Intergen: Diffusion-based multi-human motion generation under complex interactions. arXiv preprint arXiv:2304.05684 (2023) 1, 4, 8, 10, 11, 13, 14
40. Lin, K., Wang, L., Liu, Z.: Mesh graphormer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 12939–12948 (2021) 3
41. Liu, J., Shahroudy, A., Perez, M., Wang, G., Duan, L.Y., Kot, A.C.: 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) 4
42. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A Skinned Multi-Person Linear Model. Association for Computing Machinery, New York, NY, USA, 1 edn. (2023), https://doi.org/10.1145/3596711.3596800 2, 3, 7, 10
43. Ma, J., Bai, S., Zhou, C.: Pretrained diffusion models for unified human motion synthesis. arXiv preprint arXiv:2212.02837 (2022) 3
44. Maheshwari, S., Gupta, D., Sarvadevabhatla, R.K.: Mugl: Large scale multi person conditional action generation with locomotion. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 257–265 (2022) 4
45. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: AMASS: Archive of motion capture as surface shapes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5442–5451 (2019) 10
46. Martinez, J., Black, M.J., Romero, J.: On human motion prediction using recurrent neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2891–2900 (2017) 3
47. Mehta, D., Sotnychenko, O., Mueller, F., Xu, W., Sridhar, S., Pons-Moll, G., Theobalt, C.: Single-shot multi-person 3d pose estimation from monocular rgb. In: 2018 International Conference on 3D Vision (3DV). pp. 120–130. IEEE (2018) 4

48. Moeslund, T.B., Granum, E.: A survey of computer vision-based human motion capture. Computer Vision and Image Understanding **81**(3), 231–268 (2001). https://doi.org/https://doi.org/10.1006/cviu.2000.0897 1
49. Ng, E., Joo, H., Hu, L., Li, H., Darrell, T., Kanazawa, A., Ginosar, S.: Learning to listen: Modeling non-deterministic dyadic facial motion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20395–20405 (2022) 3
50. Ng, E., Romero, J., Bagautdinov, T., Bai, S., Darrell, T., Kanazawa, A., Richard, A.: From audio to photoreal embodiment: Synthesizing humans in conversations. arXiv preprint arXiv:2401.01885 (2024) 3
51. Ng, E., Xiang, D., Joo, H., Grauman, K.: You2me: Inferring body pose in egocentric video via first and second person interactions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9890–9900 (2020) 4
52. Petrovich, M., Black, M.J., Varol, G.: Action-conditioned 3d human motion synthesis with transformer vae. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10985–10995 3
53. Petrovich, M., Black, M.J., Varol, G.: TEMOS: Generating diverse human motions from textual descriptions. In: Proceedings of the European Conference on Computer Vision (2022) 1, 3
54. Plappert, M., Mandery, C., Asfour, T.: The kit motion-language dataset. Big Data **4**(4), 236–252 (dec 2016). https://doi.org/10.1089/big.2016.0028, http://dx.doi.org/10.1089/big.2016.0028 4
55. Punnakkal, A.R., Chandrasekaran, A., Athanasiou, N., Quiros-Ramirez, A., Black, M.J.: Babel: Bodies, action and behavior with english labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 722–731 (2021) 4
56. Qiu, Z., Yang, Q., Wang, J., Feng, H., Han, J., Ding, E., Xu, C., Fu, D., Wang, J.: Psvt: End-to-end multi-person 3d pose and shape estimation with progressive video transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21254–21263 (2023) 3
57. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision (2021) 3, 5, 8, 11
58. Rajasegaran, J., Pavlakos, G., Kanazawa, A., Malik, J.: Tracking people with 3d representations. arXiv preprint arXiv:2111.07868 (2021) 3
59. Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., Komatsuzaki, A.: LAION-400M: open dataset of clip-filtered 400 million image-text pairs **abs/2111.02114** (2021), https://arxiv.org/abs/2111.02114 2, 5, 6
60. Shafir, Y., Tevet, G., Kapon, R., Bermano, A.H.: Human motion diffusion as a generative prior. arXiv preprint arXiv:2303.01418 (2023) 4, 11, 13, 14
61. Sun, Y., Bao, Q., Liu, W., Fu, Y., Michael J., B., Mei, T.: Monocular, One-stage, Regression of Multiple 3D People. In: ICCV (2021) 3
62. Sun, Y., Bao, Q., Liu, W., Mei, T., Black, M.J.: TRACE: 5D Temporal Regression of Avatars with Dynamic Cameras in 3D Environments. In: CVPR (2023) 2, 3, 5, 6
63. Sun, Y., Liu, W., Bao, Q., Fu, Y., Mei, T., Black, M.J.: Putting People in their Place: Monocular Regression of 3D People in Depth. In: CVPR (2022) 2, 3, 4, 5
64. Sun, Y., Ye, Y., Liu, W., Gao, W., Fu, Y., Mei, T.: Human mesh recovery from monocular images via a skeleton-disentangled representation. In: Proceedings of

the IEEE/CVF international conference on computer vision. pp. 5349–5358 (2019) 3
65. Tanaka, M., Fujiwara, K.: Role-aware interaction generation from textual description. In: ICCV (2023) 1, 11, 13, 14
66. Tanaka, M., Fujiwara, K.: Role-aware interaction generation from textual description. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15999–16009 (2023) 4
67. Tanke, J., Zhang, L., Zhao, A., Tang, C., Cai, Y., Wang, L., Wu, P.C., Gall, J., Keskin, C.: Social diffusion: Long-term multiple human motion anticipation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9601–9611 (2023) 4
68. Tevet, G., Gordon, B., Hertz, A., Bermano, A.H., Cohen-Or, D.: Motionclip: Exposing human motion generation to clip space. In: Proceedings of the 17th European Conference on Computer Vision. pp. 358–374 (2022) 1, 3
69. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-or, D., Bermano, A.H.: Human motion diffusion model. In: Proceedings of the 11th International Conference on Learning Representations (2023) 1, 2, 3, 8
70. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30 (2017) 2, 8
71. Von Marcard, T., Henschel, R., Black, M.J., Rosenhahn, B., Pons-Moll, G.: Recovering accurate 3d human pose in the wild using imus and a moving camera. In: Proceedings of the European conference on computer vision (ECCV). pp. 601–617 (2018) 4
72. Wang, J., Xu, H., Narasimhan, M., Wang, X.: Multi-person 3d motion prediction with multi-range transformers. Advances in Neural Information Processing Systems **34** (2021) 4
73. Wang, Z., Wang, J., Lin, D., Dai, B.: Intercontrol: Generate human motion interactions by controlling every joint. arXiv preprint arXiv:2311.15864 (2023) 4
74. Wei, M., Miaomiao, L., Mathieu, S.: History repeats itself: Human motion prediction via motion attention. In: Proceedings of the European Conference on Computer Vision (2020) 3
75. Wei, W.L., Lin, J.C., Liu, T.L., Liao, H.Y.M.: Capturing humans in motion: Temporal-attentive 3d human pose and shape estimation from monocular video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13211–13220 (2022) 3
76. Wu, Y., Kirillov, A., Massa, F., Lo, W.Y., Girshick, R.: Detectron2 (2019) 5, 6
77. Yao, P., Fang, Z., Wu, F., Feng, Y., Li, J.: Densebody: Directly regressing dense 3d human pose and shape from a single color image. arXiv preprint arXiv:1903.10153 (2019) 3
78. Yu, Z., Wang, J., Xu, J., Ni, B., Zhao, C., Wang, M., Zhang, W.: Skeleton2mesh: Kinematics prior injected unsupervised human mesh recovery. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8619–8629 (2021) 3
79. Yuan, Y., Iqbal, U., Molchanov, P., Kitani, K., Kautz, J.: Glamr: Global occlusion-aware human mesh recovery with dynamic cameras. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11038–11049 (2022) 3

80. Yuan, Y., Li, X., Huang, Y., De Mello, S., Nagano, K., Kautz, J., Iqbal, U.: Gavatar: Animatable 3d gaussian avatars with implicit mesh learning. arXiv preprint arXiv:2312.11461 (2023) 3
81. Yuan, Y., Song, J., Iqbal, U., Vahdat, A., Kautz, J.: PhysDiff: Physics-guided human motion diffusion model. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (October 2023) 3
82. Yuan, Y., Wei, S.E., Simon, T., Kitani, K., Saragih, J.: Simpoe: Simulated character control for 3d human pose estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7159–7169 (2021) 3
83. Zanfir, A., Bazavan, E.G., Zanfir, M., Freeman, W.T., Sukthankar, R., Sminchisescu, C.: Neural descent for visual 3d human pose and shape. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14484–14493 (2021) 3
84. Zanfir, A., Marinoiu, E., Zanfir, M., Popa, A.I., Sminchisescu, C.: Deep network for the integrated 3d sensing of multiple people in natural images. Advances in neural information processing systems **31** (2018) 3
85. Zhai, Y., Huang, M., Luan, T., Dong, L., Nwogu, I., Lyu, S., Doermann, D., Yuan, J.: Language-guided human motion synthesis with atomic actions. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 5262–5271 (2023) 3
86. Zhang, H., Tian, Y., Zhou, X., Ouyang, W., Liu, Y., Wang, L., Sun, Z.: Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11446–11456 (2021) 3
87. Zhang, J., Zhang, Y., Cun, X., Huang, S., Zhang, Y., Zhao, H., Lu, H., Shen, X.: T2m-gpt: Generating human motion from textual descriptions with discrete representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023) 3
88. Zhang, M., Guo, X., Pan, L., Cai, Z., Hong, F., Li, H., Yang, L., Liu, Z.: Remodiffuse: Retrieval-augmented motion diffusion model. arXiv preprint arXiv:2304.01116 (2023) 3
89. Zhang, Y., An, L., Yu, T., Li, X., Li, K., Liu, Y.: 4d association graph for realtime multi-person motion capture using multiple video cameras. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1324–1333 (2020) 4
90. Zhao, Z., Bai, J., Chen, D., Wang, D., Pan, Y.: Taming diffusion models for music-driven conducting motion generation. arXiv preprint arXiv:2306.10065 (2023) 3