Motion Diversification Networks

Hee Jae Kim Eshed Ohn-Bar
Boston University
{hjkim37, eohnbar}@bu.edu

Abstract

We introduce Motion Diversification Networks, a novel framework for learning to generate realistic and diverse 3D human motion. Despite recent advances in deep generative motion modeling, existing models often fail to produce samples that capture the full range of plausible and natural 3D human motion within a given context. The lack of diversity becomes even more apparent in applications where subtle and multi-modal 3D human forecasting is crucial for safety, such as robotics and autonomous driving. Towards more realistic and functional 3D motion models, we highlight limitations in existing generative modeling techniques, particularly in overly simplistic latent code sampling strategies. We then introduce a transformer-based diversification mechanism that learns to effectively guide sampling in the latent space. Our proposed attention-based module queries multiple stochastic samples to flexibly predict a diverse set of latent codes which can be subsequently decoded into motion samples. The proposed framework achieves state-of-the-art diversity and accuracy prediction performance across a range of benchmarks and settings, particularly when used to forecast intricate in-the-wild 3D human motion within complex urban environments. Our models, datasets, and code are available at https://mdncvpr.github.io/.

1. Introduction

Humans can seamlessly navigate intricate environments by anticipating and planning over a diverse range of possible futures, i.e., potential future behaviors of those around them [6, 52, 79]. Particularly for complex, long-term scenarios that are filled with ambiguity and uncertainty, such as autonomous driving in dense social settings, modeling a diverse set of subtle future motions could mean the difference between a cautiously safe maneuver and a catastrophe. For instance, drivers may slow down for the *mere possibility* of a child abruptly crossing the street, e.g., due to an upcoming narrow sidewalk or a construction zone [19]. Within such safety-critical scenarios, even slight hand gestures or visual scanning could signal intent and thus provide crucial cues

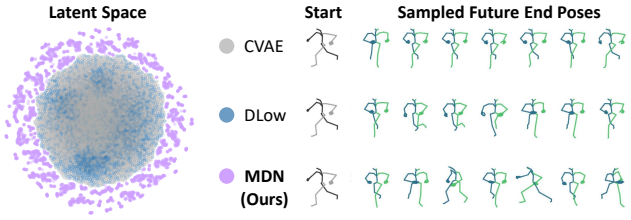

Figure 1. **Our Motion Diversification Network (MDN) Captures Diverse Data Modes.** The proposed framework produces highly diverse 3D human motion through a novel diversification module, as shown for an example starting pose on the right. We also visualize the t-SNE-based embedding [58] derived from the latent space of 3D motion data samples, showing higher coverage of data modes compared to baseline generative models (CVAE [32] and DLow [79]).

for predicting future pose and trajectory [30, 49, 52, 66, 76]. In this work, we are interested in designing expressive generative models that are capable of fully characterizing and predicting such natural and nuanced multi-modal 3D motion in realistic contexts.

Even within the controlled motion captured settings in which articulated 3D human motion synthesis is generally studied, i.e., with minimal surrounding context and restrictive social dynamics [7, 25, 26, 38, 39, 41, 45, 55, 63, 70], generating 3D human motion that is *both diverse and realistic* remains a difficult task and often results in a trade-off [3, 65, 72, 78, 79]. For instance, diffusion-based techniques have recently been shown to improve the realism of generated motion samples [3, 49], yet lead to reduced overall sample diversity [3, 53, 57]. Introducing stochastic sampling mechanisms, such as in Variational Autoencoding (VAE) [2, 10, 12, 21, 22, 32, 36, 42, 56, 71, 84], similarly fails to cover the full spectrum of human motion as the underlying likelihood-based sampling tends to emphasize the major modes in the data [72, 79]. While prior methods have attempted to mitigate the issue of mode collapse by introducing affine or simple learnable mechanisms in the latent space [72, 79], these can be difficult to optimize and only provide limited diversity within more rare modes, as these may be poorly captured by the latent representa-

tion. Critically, prior methods often assume a single isolated human actor and thus do not easily generalize to more complex settings with diverse human-scene (e.g., layouts, curbs, crosswalks), human-object (e.g., vehicles, benches, buildings, carried objects), and human-human (e.g., jogging in dense crowds, group dynamics) scenarios.

Contribution: We introduce Motion Diversification Networks (MDN), a high-capacity model for generating diverse and realistic 3D human motion across contexts. Our key insight is twofold: First, we introduce a transformer-based mechanism (referred to as *z-transformer*) for diversifying a set of sampled random variables prior to decoding into 3D motion samples. Second, due to the difficulty of the learning task, we demonstrate the importance of incorporating *deterministic motion primitives* that can guide sample diversity, promote controllability, and reduce modeling complexity in diverse scenarios. Our proposed module provides a simple yet highly effective strategy for enhancing sample diversity without compromising on realism, showing state-of-the-art performance across a range of existing 3D pose forecasting benchmarks. We analyze the benefits of the proposed approach within a concrete use case using an introduced urban navigation benchmark (based on dense pedestrian simulation in CARLA [14]). However, as there is a current lack of suitable real-world datasets that can be used to evaluate the broad spectrum of 3D motion in diverse scenes and social contexts, i.e., beyond simple collision avoidance, we analyze robustness and modeling capacity using a diverse but noisy real-world data extracted from YouTube videos. Through the comprehensive analysis, our study takes a step towards closing the current gap between simulated, generated, and realistic 3D human motion, particularly in cross-dataset, scene-aware, and in-the-wild settings.

2. Related Work

Due to its subtle and complex nature, long-term modeling and forecasting of 3D human pose beyond simplified and controlled settings can easily confound existing models, as we discuss next.

Learning to Synthesize Diverse 3D Motion: Stochastic approaches in human motion prediction are widely studied, focusing on circumventing the well-known mode collapse problem in generative models. Yuan and Kitani [79] learn an affine transform mapping parameters to explicitly map a single sampled random variable into a set of latent codes. However, the simplistic affine mapping cannot effectively account for uncertainty *across the various modes* in the data. In modes where data is rare, the generated latent codes will also not effectively account for motion, rendering the affine transform ineffective. Another related study is by Xu et al. [72], which proposes a learnable deterministic anchor to simplify the prediction of diverse modes. However, this proposed approach can only provide limited guidance toward diverse samples, and the learned parameters can similarly fail to capture rare data modes. In contrast, we demonstrate the need for incorporating explicit motion primitives that can reliably ease the challenging learning task, which significantly outperforms in terms of diversity [72]. We also note that the aforementioned approaches emphasize a single actor that is performing short-term and contrived behaviors [4, 18, 20, 38, 45, 55, 70, 79], and cannot easily generalize to more complex settings, e.g., with scene context. More recently, researchers have incorporated elements of scene-awareness when forecasting plausible future poses [26, 63–65], e.g., training a model to predict how a human may walk over and sit on a couch given 3D information of the surrounding room, yet diversity is often sacrificed for prediction accuracy [3, 49, 70]. In this work, we demonstrate a generalizable approach that can easily incorporate scene context and diverse human motion, even when leveraging challenging in-the-wild and noisy monocular data. More broadly, our study is complementary to the aforementioned methods (e.g., [46, 49]), as we explore an orthogonal approach to guiding human motion generation models.

Datasets and In-the-Wild Synthesis: With advances in reliable motion capture technologies, several large and accurate temporal 3D motion datasets have been released by the research community, including Human3.6M [27], AMASS [41], PROX [25], COUCH [85], HPS [24], EgoBody [83], and 3DPW [60]. However, with the exception of HPS and 3DPW, such datasets generally involve short-term static indoor settings with minimal social and layout interactions. This begs the question: do methods generalize beyond such benchmarks to more diverse settings? While 3DPW is an ambitious outdoor benchmark (enabled by the use of on-body Inertial Measurement Units in conjunction with 3rd person video), only a handful of the scenes capture natural goal-driven pedestrian behavior without urban scene elements such as dynamic intersections with intricate scene-induced constraints. Hence, to study such settings, researchers often tend to resort to simulation environments (e.g., [1, 7, 14, 49, 54, 56, 67]) or only collect trajectory-level annotations, e.g., in the Bird's Eye View [2, 22, 50, 71]. Yet, the former generally employ naive simulations that lack in motion realism (e.g., Habitat [47] and Gibson [1]) while the latter only provides a coarse representation of nuanced real-world human motion (i.e., motion is reduced to a single global root joint). Given the difficulty in obtaining clean 3D data for outdoor urban settings, in our experiments, we sought to explore both a novel simulation benchmark and scalable solutions using large and diverse publicly available data sources. We also perform a user study in order to better address the gap between sim-

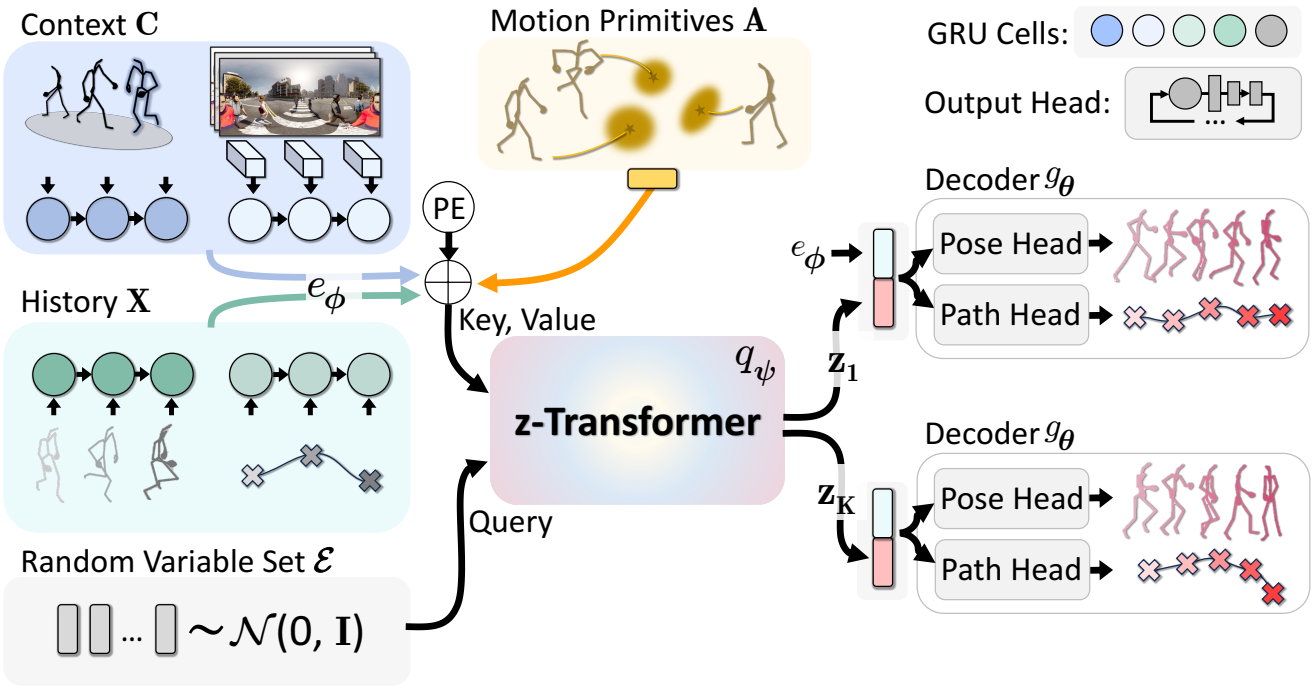

Figure 2. **Our Motion Diversification Network.** The model employs an encoder e_ϕ, which takes as input 3D pose and 2D path data. Given a sampled set of random variables, a transformer-based module q_ψ produces a diverse set of latent codes obtained by fusing information from the encoder and a set of motion primitives $\mathbf{A}$ (encoded via one fully connected layer). Finally, a generator function g_θ decodes the latent codes into motion samples. The proposed diversification module can be used to incorporate additional context, i.e., scene and social context information $\mathbf{C}$, if available. The model is trained in two stages, first as a Conditional Variational Autoencoder (CVAE), and then we train a transformer module [59] to facilitate the diversifying latent codes. We denote GRU [8] cells with circles. PE denotes positional encoding.

ulated and realistic motion, which can facilitate future research, e.g., in robotics and autonomous driving. While studies in computer vision generally focus on in-the-wild pose estimation (e.g., [11, 17, 28, 29, 44, 51, 68, 69, 75, 87]) and not on motion synthesis [80], we naturally ask whether the output of such methods may be used to supervise and facilitate increased sample diversity, i.e., of novel realistic behaviors, when paired with our proposed context-aware motion generation network.

3. Motion Diversification Networks

Our overarching goal is to predict multiple future 3D motion sequences that are realistic, i.e., human-like and close to one of the ground-truth motion sequences as possible, and diverse, i.e., covering the full range of plausible motions. We build on prior work in deep generative models (outlined in Sec. 3.1) through an effective *diversification module* (outlined in Fig. 2). At its core, the module leverages a learnable *transformer-based mechanism* for diversifying an initial set of independently sampled random variables (used as query, discussed in Sec. 3.2). Due to the difficulty in learning to balance diversity and realism, our key insight is to also utilize the attention mechanism to effectively inject additional context constraints and motion generation guidance, specifically using data-driven *deterministic motion primitives* (Sec. 3.3).

3.1. Background

Generative Modeling Formulation: Future 3D motion can be modeled using a stochastic model, such as a conditional variational autoencoder (CVAE), which can be conditioned on context information (e.g., historical observations) [32, 79]. We denote $\mathbf{X} \in \mathbb{R}^{T_h \times N_h}$ as the input sequence of context history (i.e., condition) of T_h time steps. We then leverage the input to predict K future motion sequences of length T_f, denoted as $\{\widehat{\mathbf{Y}}_k\}_{k=1}^K$, where $\widehat{\mathbf{Y}}_k \in \mathbb{R}^{T_f \times N_f}$. In general, prior methods assume data in the form of 3D coordinates representing root-normalized joint positions for J joints, such that $\mathbf{X}[t], \widehat{\mathbf{Y}}_k[t] \in \mathbb{R}^{J \times 3}$. However, we keep our formulation generic so that additional inputs and outputs can be readily incorporated, e.g., 2D image-based path motion [7, 21, 34, 71, 77]. We also denote optional additional context $\mathbf{C} \in \mathbb{R}^{T_h \times N_c}$, e.g., image or social context [7], and note that our formulation does not

require such input (optional in Fig. 2) while the proposed diversification module can operate on arbitrary contexts (as further discussed in Sec. 3.4). During training, we assume to be given motion demonstrations $\mathbf{Y}$, and follow [32] to reparameterize the data distribution using a latent variable $\mathbf{z} \in \mathbb{R}^{N_z}$ which can account for *underlying intent and uncertainty* regarding future motion, i.e.,

$$p(\mathbf{Y}|\mathbf{X}, \mathbf{C}) = \int p(\mathbf{Y}|\mathbf{z}, \mathbf{X}, \mathbf{C}) p(\mathbf{z}) \mathrm{d}\mathbf{z} \tag{1}$$

where the latent code is sampled from a prior multivariate Gaussian distribution $\mathbf{z} \sim p_0(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$. We note that the inputs are first processed using an *encoder neural network*, $e_{\boldsymbol{\phi}}(\mathbf{X}, \mathbf{C})$, which learns to map raw observations to an N_z-dimensional embedding. During inference, motion generation can be implemented as a deterministic mapping function, a neural network $g_{\boldsymbol{\theta}}$ with parameters $\boldsymbol{\theta}$, which maps a sampled latent code z and historical observations to a prediction,

$$\widehat{\mathbf{Y}} = g_{\boldsymbol{\theta}}(\mathbf{z}, \mathbf{X}, \mathbf{C}) \tag{2}$$

Random Latent Code Sampling: Once $g_{\boldsymbol{\theta}}$ is trained, obtaining multiple future motions $\{\widehat{\mathbf{Y}}_k\}_{k=1}^K$ can be done through repeated sampling from latent prior, i.e., to obtain $\mathbf{Z} = \{\mathbf{z}_k\}_{k=1}^K$. However, such independent sampling and subsequent decoding can only provide limited sample diversity as the data likelihood emphasizes the most common modes in the data (i.e., mode collapse [79]) and similar generated motion. To address the lack of coverage, recent work alters the sampling process in order to promote sample diversity, as discussed next.

Affine Latent Mapping Functions: To improve over the independent latent code sampling process, DLow [79] introduces a diversity-inducing neural network $\{\mathbf{a}_k, \mathbf{b}_k\}_{k=1}^K = q_{\boldsymbol{\psi}}(\mathbf{X})$ which learns to output a K set of *affine mapping parameters* based on context input. The parameters are then used to produce a diverse set of latent codes $\mathbf{Z}$, i.e., by applying the parameters to a single sampled unit Gaussian variable, $\boldsymbol{\epsilon} \sim p_0(\boldsymbol{\epsilon}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$. We note that training takes place in two stages, first via CVAE-based pre-training [79] of $g_{\boldsymbol{\theta}}$ followed by the diversification network training while $g_{\boldsymbol{\theta}}$ is kept frozen. Given the diversified set of latent codes $\mathbf{Z}$, each code $\mathbf{z}_k$ is decoded separately, as in Eq. 2, to produce the final K samples. This sampling process has been previously shown to facilitate greater predicted sample diversity, i.e., by leveraging a common random variable $\boldsymbol{\epsilon}$ and modeling repulsion among the samples. However, purely relying on a single random variable alone to capture future uncertainty *among all modes* can be difficult to learn. Moreover, we argue that the learnable affine mapping functions can only provide a limited diversification and correlation modeling mechanism, as discussed next.

3.2. Latent Variable Transformer

This section introduces our z-transformer, an attention-based module that effectively captures correlations among samples and modes in the data. Instead of relying on an affine mapping, our key insight is to leverage transformer [59] in order to obtain a diverse set of latent codes. Specifically, we sample a set of K random variables from $p_0(\boldsymbol{\epsilon})$, i.e., $\mathcal{E} \in \mathbb{R}^{K \times N_z}$, and leverage it as a *query* matrix,

$$\mathbf{Z} = \mathrm{Attn}(\mathcal{E}, \mathbf{K}, \mathbf{V}) = \mathrm{softmax}\left(\frac{\mathcal{E}\mathbf{K}^\intercal}{\sqrt{N_z}}\right)\mathbf{V} \tag{3}$$

where keys and values $\mathbf{K}, \mathbf{V} \in \mathbb{R}^{K \times N_z}$ are computed based on the encoded input $\mathbf{X}$. We leverage self-attention to the key and value matrices first, followed by cross-attention with the query matrix. In our implementation, we also leverage a positional encoding [15, 73] of dimension $K \times N_z$.

Intuitively, leveraging the set of random variables as a query enables us to better capture uncertainty *within each mode*. We note that Eq. 3 details one transformer layer, and these are stacked (we leverage eight blocks) to incrementally update the latent code with multi-head cross-attention and an MLP. We find the proposed latent variable transform to result in significantly higher sample diversity compared to more simplistic affine-based approaches [79] in Sec. 4, yet still lack in coverage, particularly for rare modes of the data. Thus, to further guide the model to learn correct cross-mode trends as well as within-mode uncertainty, we incorporate additional guidance in the form of motion primitives.

3.3. Deterministic Motion Primitives

Predicting diverse outputs can be challenging for most machine-learning models. Our second key insight lies in the flexibility of the z-transformer for incorporating additional constraints, i.e., when computing the key and value matrices. To prevent the stochastic diversification process from resulting in high diversity yet unrealistic samples and ease inference of highly diverse motions, we leverage the z-transformer to incorporate additional constraints and context. Specifically, we introduce a set of data-driven exemplars $\mathbf{A} \in \mathbb{R}^{K \times N_f}$ that can aid in predicting rare modes. While we have experimented with various forms of primitives and constraints, we found simple pose-space primitives to work based. We leverage simple k-means [40] clustering of the inputs in the training dataset $\mathbf{X}$ as primitives, which are then encoded via a single fully connected layer, replicated, and summed with encoded context when computing $\mathbf{K}$ and $\mathbf{V}$, i.e., prior to computing the attention in Eq. 3. Thus, our complete diversification network can now be written explicitly as

$$\mathbf{Z} = q_{\boldsymbol{\psi}}(\mathbf{X}, \mathbf{C}, \mathcal{E}, \mathbf{A}) \tag{4}$$

Here, we hypothesize that the clustering-based primitives can help in scenarios where the latent space representations

(and generator network) may be less effective, such as in the highly diverse or rare modes in the data. We have experimented with various elaborate (e.g., learnable [72]) primitives, yet found our clustering-based strategy to outperform. Intuitively, the primitives can provide the generator with a strong initial bias towards plausible modes of human motion. In our implementation, we do not find the random initialization in the clustering algorithm to have a significant impact on the resulting sample diversity, most likely due to the effective combination with the transformer-based module in Sec. 3.2. In summary, in our motion generation experiments, we leverage the diversification module as a *unified mechanism for incorporating various contextual information*, e.g., additional social and image context, as outlined below.

3.4. Training and Network Architecture

Loss: Our framework follows two-stage training, where the encoder e_{ϕ} and generator g_{θ} are first trained using a variational lower bound loss [32, 79],

$$\mathcal{L}_{\text{CVAE}} = \text{KL}\left(e_{\phi}(\mathbf{z}|\mathbf{X}, \mathbf{C}) \parallel p_0(\mathbf{z})\right) + \|g_{\boldsymbol{\theta}}(\mathbf{z}) - \mathbf{Y}\|_2^2 \quad (5)$$

where KL is the Kullback-Leibler divergence. When training the encoder, the output of the GRUs from different sources of data are summed. To ensure meaningful comparison, we follow DLow [79] where the training loss for the second stage comprises a weighted sum of two terms, a reconstruction loss,

$$\mathcal{L}_r = \min_k \|\widehat{\mathbf{Y}}_k - \mathbf{Y}\|^2 \quad (6)$$

and a diversity-promoting loss that promotes pairwise distances among the K generated predictions, i.e.,

$$\mathcal{L}_d = \frac{2}{K(K-1)} \sum_{j=1}^{K} \sum_{k=j+1}^{K} \exp\left(-\frac{\|\widehat{\mathbf{Y}}_j - \widehat{\mathbf{Y}}_k\|_1}{\alpha}\right) \quad (7)$$

where α is a hyperparameter. We set the scaling parameter to be $\alpha = 100$ and the weight for the diversity loss term 25.

Basic Architecture Details: To ensure meaningful comparison to prior work, our baseline architecture only incorporates 3D pose data, i.e., $N_z = 128, N_h = J \times 3$, and the encoder is implemented as a GRU [8, 9] with 128 hidden units, followed by an MLP.

2D Path, Social and Scene Context: Our diversification module flexibly supports other types of inputs and can be used to effectively integrate scene and social context. While rarely available in 3D human motion datasets in our complete unified architecture the encoder also incorporates 2D (in the image space) pose history (as shown in Fig. 2), image context, and explicit social context. Each of these inputs leverages a dedicated GRU with 128 hidden units, followed by a final summation to obtain a context-aware $\mathbf{K}, \mathbf{V}$ in Eq. 3. In the case of visual context, the encoder first processes a padded person-centered window to extract a 1000-dimensional feature using a Swin Transformer [37] prior to input to the dedicated GRU. Our social context is computed over neighborhoods in the image, where surrounding agents' 3D poses are similarly encoded into a 128-dimensional social context feature with a GRU (with a maximum of neighboring agents set at eight). Given 2D path information in our image-aware experiments, the generator network has two output branches, one for image coordinates 2D path and one for the 3D pose. We found our multi-task structure to improve over sequential prediction of the 2D path followed by 3D pose (in contrast to Cao et al. [7] and Rempe et al. [49]). We optimize the networks with Adam [31]. Additional details regarding the implementation can be found in the supplementary.

4. Experiments

To fully evaluate diverse and nuanced social motion generation capabilities, in this section, we comprehensively analyze our complete framework across a set of settings. In addition to standard benchmarks, we also incorporate a simulation benchmark captured in crowded social settings (in CARLA [14, 81]). We further analyze diverse motion generation by utilizing realistic motion sequences from publicly available internet videos.

Real-World Benchmarks: To validate our proposed diversification strategy, we evaluate MDN on Human3.6M [27], a large-scale motion capture dataset. Human3.6M contains 3.6 million video frames at 50 Hz. 15 daily actions in indoor settings performed by 11 subjects are recorded in a 17-joint representation. We use the standard training and (S1, S5, S6, S7, and S8) and validation splits (S9 and S11). Nonetheless, we note that Human3.6M is largely saturated at this stage and only contains isolated indoor scenes. To analyze the context-aware diverse motion generation, we also evaluate our proposed model on two commonly employed real-world datasets, 3DPW and HPS. 3DPW contains outdoor motion sequences in SMPL (24 joints) joint representation. To standardize evaluation, we adopt a common set of joints between CARLA and SMPL joint representation for training and testing, resulting in 19 body joints. We use a standard test split for evaluation. HPS comprises indoor and outdoor motion sequences based on a 23-joint representation from an inertial capture suit. As there is a lack of real-world naturalistic human motion datasets with 3D ground truth, this experimental setup allows for holistic evaluation of diverse human motion generation, i.e., beyond currently restrictive benchmarks and simplified simulations.

Simulation Benchmark: Prior benchmarks for context-aware 3D human motion modeling (e.g., GTA-IM [7], JTA [16]) only incorporate limited urban conditions (e.g.,

without crosswalks, groups, and intersections) and motion diversity. In this work, we employ CARLA (version 0.9.13) [14] to evaluate the utility of the proposed approach. While CARLA is heavily used for developing autonomous vehicle policies, the implemented pedestrians are currently controlled using naive kinematics, exhibiting plausible but generally basic motion with limited diversity. We can, therefore, directly quantify the benefits of our proposed approach to enhancing motion modeling and realism while impacting future research in interactive robot learning. To collect an initial dataset with dense scenes, we spawn the maximum possible pedestrians (250) in constrained areas in Town 10. For consistency, six cameras are used to create a similar 360° view to the YouTube scenes. The collected synthetic dataset (referred to as **DenseCity**) comprises $30K$ frames with 82 walking sequences. On average, DenseCity has 21.3 pedestrians per image. Our chosen experimental setup of enhancing and diversifying from an initially limited dataset involves a common use case in robotics. We further filter the benchmark in order to remove samples with implausible physics and unnatural motion (which do occasionally occur in CARLA, as discussed further in our supplementary).

YouTube Dataset: Given the restricted diversity in motion samples of the aforementioned datasets, we additionally analyze the role of incorporating in-the-wild extracted motion sequences from diverse and dense YouTube videos [33, 82]. This allows us to understand whether our introduced model can fully accommodate natural diversity in human motion while also closing the current gap between simulated and realistic motion. We download a diverse dataset of publicly available internet videos collected in various urban settings and locations (Boston, Quebec, Orlando, Seoul, Tokyo, Amsterdam, and Naples). We leverage 360° clips with a larger view of pedestrians and context (cropped equirectangular projection imagery to the resolution of 1920 × 1080) and longer-term trajectories. After filtering small person instances which often result in highly noisy trajectories, we obtain a motion dataset totaling 288 motion sequences (average length of 5.8s). On average, the YouTube dataset has 19.2 pedestrians per image (based on detected skeletons).

Metrics and Baselines: We leverage standard diversity (Average Pairwise Distance, APD) and quality (Average Displacement Error, ADE, and Final Displacement Error, FDE) metrics over root-aligned joints [70, 79]. Our main baselines include CVAE [32], which is commonly employed for motion modeling, as well as DLow [79] and PoseGPT [38], which are designed to better capture diverse motion. However, both baselines have only modeled 3D human motion in isolation and without surrounding context. When training PoseGPT, we remove the action conditioning and allow the network to learn to predict diverse actions. We follow Cao et al. [7] and predict two-second futures of a motion given 0.5s context (the supplementary contains additional experiments with a longer horizon).

Table 1. **Evaluation on Human3.6M**. Our method achieves the best-known performance on the Human3.6M benchmark in terms of sample diversity without sacrificing prediction accuracy.

Method	APD ↑	ADE ↓	FDE ↓
Pose-Knows [62]	6.723	0.461	0.560
MT-VAE [74]	0.403	0.457	0.595
HP-GAN [4]	7.214	0.858	0.867
BoM [5]	6.265	0.448	0.533
GMVAE [13]	6.769	0.461	0.555
DeLiGAN [23]	6.509	0.483	0.534
DSF [78]	9.330	0.493	0.592
DLow [79]	11.741	0.425	0.518
MOJO [86]	12.579	0.412	0.514
GSPS [43]	14.757	0.389	0.496
BeLFusion [3]	7.602	0.372	0.474
STARS [72]	15.884	0.358	0.445
MDN (Ours)	**17.450**	**0.355**	**0.442**

Table 2. **Evaluation on 3DPW and HPS.** Our model shows state-of-the-art results, in terms of diversity and accuracy, for two additional real-world benchmarks. Moreover, our model benefits from diverse data. Specifically, the addition of human motion trajectories extracted from in-the-wild videos (full details in our supplementary) is shown to further improve realism and diversity.

	3DPW [61]			HPS [24]		
Method	APD ↑	ADE ↓	FDE ↓	APD ↑	ADE ↓	FDE ↓
CVAE [32]	3.068	1.032	1.096	3.592	0.970	1.019
DLow [79]	4.111	1.010	1.069	4.835	0.958	1.005
MDN	6.355	0.982	1.032	**6.943**	0.916	0.971
CVAE [32]+YouTube	3.100	0.991	1.060	3.634	0.913	0.958
DLow [79]+YouTube	4.160	0.969	1.034	5.010	0.898	0.938
MDN+YouTube	**8.266**	**0.918**	**0.986**	6.925	**0.886**	**0.930**

4.1. Results

Benefits of Latent Variable Transformer: We first analyze the role of the introduced z-transformer on the standard Human3.6M dataset. In this experiment, we only include 3D pose history as the input and output of the model without the additional path and image context. Even within the simplified and saturated performance on this dataset, Table 1 demonstrates our proposed z-transformer to improve sample diversity significantly, achieving the best results to date. Remarkably, we also achieve high accuracy and do not observe the diversity-realism trade-off that plagues prior models, e.g., [3]. In particular, our model achieves 17.450 in diversity score, improving over prior state-of-the-art by over 9.8% (note that even small improvements in diversity score are considered significant [70, 79]). Nonethe-

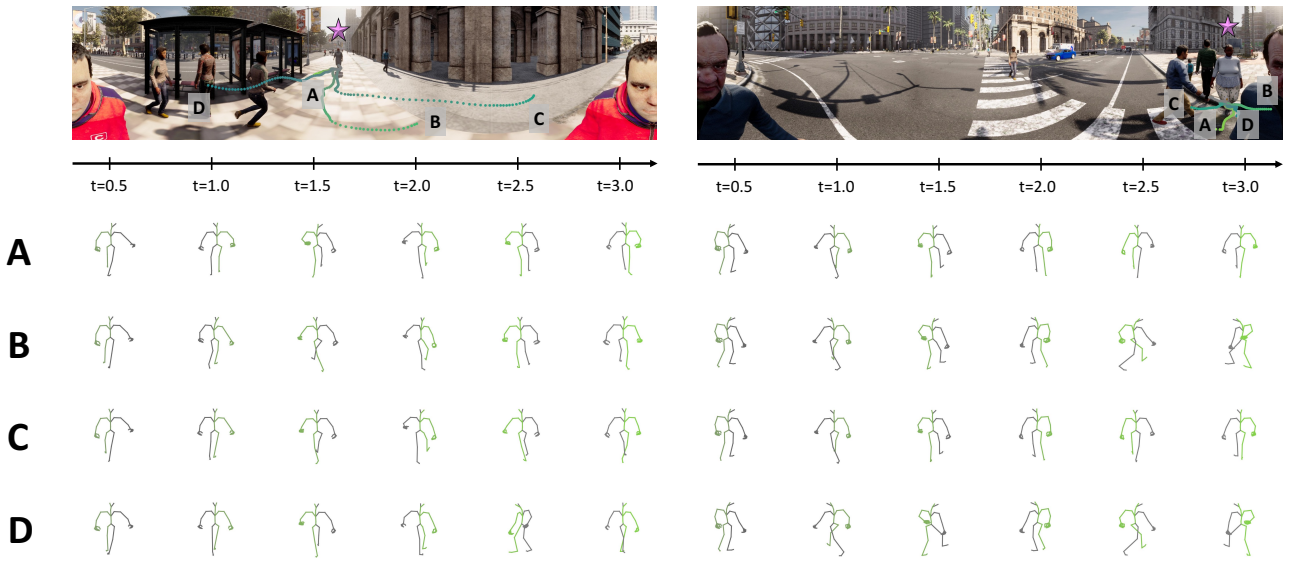

Figure 3. **Nuanced and Diverse Motion Predicted by MDN.** We show diverse 2D paths predicted in two scenarios with intricate social settings. Corresponding 3D poses to the annotated 2D paths show correct turning around social interactions (e.g., sequence D).

less, we highlight that Human3.6M itself lacks diversity, which motivates our subsequent experiments on additional benchmarks. We also evaluate our model on two real-world benchmarks, 3DPW and HPS. We train our model using our CARLA benchmark using the joint representation of SMPL and Xsens (an IMU-based motion capture suit) for evaluation on 3DPW and HPS, respectively. CVAE [32] and DLow [79] are selected as baselines. Table 2 demonstrates consistent benefits of the proposed sampling diversification in z-transformer compared to DLow and CVAE.

Full Model Human Motion Generation Results: In this experiment, we investigate the performance of our full context-aware generation model. We evaluate our model on the introduced DenseCity benchmark, which includes frequent social scenarios. We further incorporate a 2D image-based path modeling task, which is a commonly studied task in autonomous driving. We also utilize social and pedestrian-oriented image context in z-transformer along with 3D pose history. Noticeably, our proposed MDN also achieves the highest diversity score in complex urban environments as well as the best prediction accuracy, as shown in Table 3. In particular, we outperform DLow in all metrics, which also builds upon CVAE while adopting affine transform for diversity-promoting strategy. This finding demonstrates the utility of the proposed pipeline of z-transformer for learning diverse human motion patterns. In the context of joint future prediction of pose and path, our results demonstrate that MDN improves over the prior scene-aware method Cao et al. [7]. Moreover, our model is more efficient due to the joint 2D path and 3D pose reasoning, as opposed to sequentially inferring the path followed by the 3D pose. Additional data from YouTube mixed with the training data demonstrate our high-capacity model can easily absorb general naturalistic motion sequences. Additional ablations can be found in the supplementary.

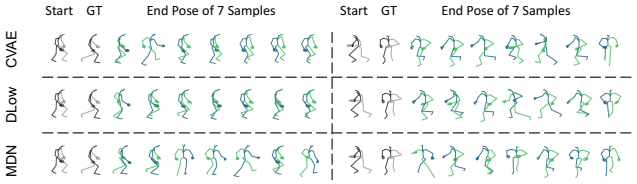

Figure 4. **MDN Predicts Diverse Motion Samples.** Compared to DLow and CVAE-based predictions on DenseCity, MDN better captures nuanced walking patterns and anticipates realistic potential future scenarios.

Improving Diversity and Generalization with In-the-Wild YouTube Data: While DenseCity provides extensive social and context-based 3D human motion, it can be limited in diversity due to limitations in current simulation environments. For instance, we can leverage YouTube-extracted 3D motion trajectories to learn a generalized human motion model better. As shown in Table 2, our model is able to benefit from such data and produce samples with increased diversity and realism. We note that models are trained over simulated and YouTube data in these experiments, and evaluated on 3DPW and HPS. Additional details and analysis can be found in the supplementary. When incorporating YouTube data and evaluating on DenseCity, as

Table 3. **Evaluation on DenseCity.** Results are shown in terms of diversity (APD) and accuracy (ADE, FDE) against several baselines for both 3D pose and 2D trajectory error over a **two seconds future prediction** task.

	3D Pose			2D Path		
Method	APD ↑	ADE ↓	FDE ↓	APD ↑	ADE ↓	FDE ↓
CVAE [32]	7.451	0.610	0.932	-	-	-
PoseGPT [38]	9.099	0.913	0.980	-	-	-
HuMoR [48]	11.134	0.705	1.030	-	-	-
DLow [79]	11.980	0.596	0.899	-	-	-
Cao et al. [7]	5.810	0.858	1.285	0.978	0.701	0.675
MDN	16.799	0.584	**0.879**	1.065	0.666	0.621
Cao et al. [7]+YouTube	5.593	0.805	1.096	0.782	0.697	0.632
MDN+YouTube	**16.812**	**0.578**	0.921	**1.331**	**0.646**	**0.589**

shown in Table 3, we observe mixed results with higher diversity but also lower accuracy. While APD remains high, we do find that incorporating YouTube data hurts 3D pose performance on DenseCity, which was collected using the built-in pedestrian controller in CARLA. However, upon inspection of the trajectories, we found highly realistic samples due to the added YouTube data. We hypothesize that while additional realistic motions are added, CARLA leverages a fixed and highly tuned controller, which sometimes lacks realism. To further study this insight, we performed a perceptual user study with closed-loop simulation.

Perceptual User Study on Realism: We further conduct a perceptual user study to understand the realism of human motion prediction under human-human and human-scene interaction. We leverage YouTube data during the training of a generative model for realistic motion generation. Paired with an inverse kinematics model [35], we generated human animations of the model in closed-loop on CARLA. 13 Participants compared two sequences side-by-side, one generated by CARLA and another by MDN, without knowledge of which is which. Each participant then provided a (1) preference selection and (2) absolute realism feedback for ranking each sequence on a 5-point Likert scale. In our experiments, we emphasized social and object interaction scenarios. As shown in Fig. 5, we find that users overwhelmingly preferred the proposed MDN over the built-in CARLA kinematics 72.5% of the time. The absolute average realism rank was also higher for our proposed approach.

Qualitative Results: Visualization of the nuanced and diverse motion predicted by MDN can be seen in Fig. 3. Specifically, we visualize how 2D paths predicted by our model are context-aware, effectively reasoning over scene layout (e.g., sidewalk or intersection) as well as subtle social interaction (e.g., at close proximity to surrounding pedestrians along the path). We demonstrate the MDN generates plausible and often smoother social interactions than the privileged and highly-tuned baseline CARLA controller.

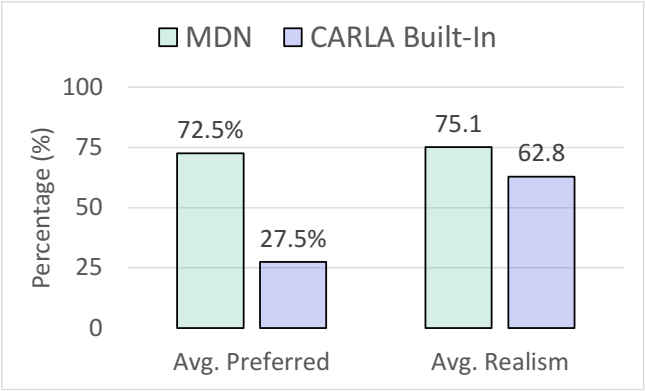

Figure 5. **Perceptual User Study.** MDN outperforms the CARLA built-in controller both in terms of average preference (side-by-side comparison) and absolute realism scores. The scores of absolute average realism are normalized on a scale of 0 to 100.

5. Conclusion

Diverse and realistic generation of human motion can facilitate more seamless and safe real-world intelligent systems. As a step towards this overarching goal, we propose a framework for learning nuanced multi-modal human motion in realistic contexts. In particular, we introduced an efficient transformer-based latent variable diversification mechanism. Our analysis demonstrates improved sample diversity without a trade-off in prediction accuracy, thus effectively generalizing models from more restricted (e.g., indoor, single-agent, minimal context) conditions to more diverse and complex real-world settings. While we take a step towards closing the gap between generated, simulated, and realistic human motion in realistic settings, there are many fundamental challenges for modeling the rich diversity in 3D human motion. In the future, we plan to further investigate the challenging problem of learning motion priors from large online real-world video resources to step towards scalable human motion generation. However, while we employed publicly available internet videos for a potentially societally beneficial application—more robust and generalized computer vision and robotics models, there could be ethical considerations with leveraging such publicly available data, e.g., it may contain private and sensitive information. Moreover, the observed individuals may not wish their data to be used. Nonetheless, by leveraging a simulation environment, our study can potentially facilitate privacy while providing more realistic behaviors for future researchers developing human-interactive systems.

Acknowledgments: We thank the Rafik B. Hariri Institute for Computing and Computational Science and Engineering (Focused Research Program award #2023-07-001) and the National Science Foundation (IIS-2152077) for supporting this research.

References

[1] iGibson social navigation challenge. https://svl.stanford.edu/igibson/challenge.html, 2022. 2

[2] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social LSTM: Human trajectory prediction in crowded spaces. In *CVPR*, 2016. 1, 2

[3] German Barquero, Sergio Escalera, and Cristina Palmero. Belfusion: Latent diffusion for behavior-driven human motion prediction. In *ICCV*, 2023. 1, 2, 6

[4] Emad Barsoum, John Kender, and Zicheng Liu. Hp-gan: Probabilistic 3d human motion prediction via gan. In *CVPR*, 2018. 2, 6

[5] Apratim Bhattacharyya, Bernt Schiele, and Mario Fritz. Accurate and diverse sampling of sequences based on a "best of many" sample objective. In *CVPR*, 2018. 6

[6] Elliot C Brown and Martin Brüne. The role of prediction in social neuroscience. *Frontiers in Human Neuroscience*, 2012. 1

[7] Zhe Cao, Hang Gao, Karttikeya Mangalam, Qi-Zhi Cai, Minh Vo, and Jitendra Malik. Long-term human motion prediction with scene context. In *ECCV*, 2020. 1, 2, 3, 5, 6, 7, 8

[8] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *NeurIPS Deep Learning and Representation Learning Workshop*, 2014. 3, 5

[9] Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. Gated feedback recurrent neural networks. In *ICML*, 2015. 5

[10] Alexander Cui, Sergio Casas, Abbas Sadat, Renjie Liao, and Raquel Urtasun. Lookout: Diverse multi-future prediction and planning for self-driving. In *ICCV*, 2021. 1

[11] Qi Dang, Jianqin Yin, Bin Wang, and Wenqing Zheng. Deep learning based 2D human pose estimation: A survey. *Tsinghua Sci. Technol*, 2019. 3

[12] Nachiket Deo, Eric Wolff, and Oscar Beijbom. Multimodal trajectory prediction conditioned on lane-graph traversals. In *CoRL*, 2022. 1

[13] Nat Dilokthanakul, Pedro AM Mediano, Marta Garnelo, Matthew CH Lee, Hugh Salimbeni, Kai Arulkumaran, and Murray Shanahan. Deep unsupervised clustering with gaussian mixture variational autoencoders. *ICLR*, 2016. 6

[14] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An open urban driving simulator. In *CoRL*, 2017. 2, 5, 6

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 4

[16] Matteo Fabbri, Fabio Lanzi, Simone Calderara, Andrea Palazzi, Roberto Vezzani, and Rita Cucchiara. Learning to detect and track visible and occluded body joints in a virtual world. In *ECCV*, 2018. 5

[17] Matteo Fabbri, Fabio Lanzi, Simone Calderara, Stefano Alletto, and Rita Cucchiara. Compressed volumetric heatmaps for multi-person 3d pose estimation. In *ICCV*, 2020. 3

[18] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent network models for human dynamics. In *ICCV*, 2015. 2

[19] Ray Fuller. Towards a general theory of driver behaviour. *Accident Analysis & Prevention*, 2005. 1

[20] Erik Gärtner, Mykhaylo Andriluka, Erwin Coumans, and Cristian Sminchisescu. Differentiable dynamics for articulated 3d human motion reconstruction. In *ICCV*, 2022. 2

[21] Tianpei Gu, Guangyi Chen, Junlong Li, Chunze Lin, Yongming Rao, Jie Zhou, and Jiwen Lu. Stochastic trajectory prediction via motion indeterminacy diffusion. In *ICCV*, 2022. 1, 3

[22] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social GAN: Socially acceptable trajectories with generative adversarial networks. In *CVPR*, 2018. 1, 2

[23] Swaminathan Gurumurthy, Ravi Kiran Sarvadevabhatla, and R Venkatesh Babu. Deligan: Generative adversarial networks for diverse and limited data. In *CVPR*, 2017. 6

[24] Vladimir Guzov, Aymen Mir, Torsten Sattler, and Gerard Pons-Moll. Human poseitioning system (HPS): 3D human pose estimation and self-localization in large scenes from body-mounted sensors. In *CVPR*, pages 4318–4329, 2021. 2, 6

[25] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *ICCV*, 2019. 1, 2

[26] Mohamed Hassan, Yunrong Guo, Tingwu Wang, Michael Black, Sanja Fidler, and Xue Bin Peng. Synthesizing physical character-scene interactions. *SIGGRAPH*, 2023. 1, 2

[27] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *PAMI*, 2013. 2, 5

[28] Umar Iqbal, Pavlo Molchanov, and Jan Kautz. Weakly-supervised 3d human pose learning via multi-view images in the wild. In *CVPR*, 2020. 3

[29] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3d human model fitting towards in-the-wild 3d human pose estimation. In *3DV*, 2021. 3

[30] Christoph G Keller and Dariu M Gavrila. Will the pedestrian cross? a study on pedestrian path prediction. *T-ITS*, 2013. 1

[31] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv.org, 1412.6980*, 2014. 5

[32] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv.org, 1312.6114*, 2013. 1, 3, 4, 5, 6, 7, 8

[33] Lei Lai, Zhongkai Shangguan, Jimuyang Zhang, and Eshed Ohn-Bar. XVO: Generalized visual odometry via cross-modal self-training. In *ICCV*, 2023. 6

[34] Namhoon Lee, Wongun Choi, Paul Vernaza, Christopher B Choy, Philip HS Torr, and Manmohan Chandraker. Desire: Distant future prediction in dynamic scenes with interacting agents. In *CVPR*, 2017. 3

[35] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. HybrIK: A hybrid analytical-neural inverse kinematics solution for 3D human pose and shape estimation. In *ICCV*, 2021. 8

[36] Yuejiang Liu, Qi Yan, and Alexandre Alahi. Social NCE: Contrastive learning of socially-aware motion representations. In *ICCV*, 2021. 1

[37] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 5

[38] Thomas Lucas, Fabien Baradel, Philippe Weinzaepfel, and Grégory Rogez. PoseGPT: quantization-based 3d human motion generation and forecasting. In *ECCV*, 2022. 1, 2, 6, 8

[39] Diogo Luvizon, Marc Habermann, Vladislav Golyanik, Adam Kortylewski, and Christian Theobalt. Scene-aware 3d multi-human motion capture from a single camera. *arXiv.org, 2301.05175*, 2023. 1

[40] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1967. 4

[41] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *ICCV*, 2019. 1, 2

[42] Karttikeya Mangalam, Harshayu Girase, Shreyas Agarwal, Kuan-Hui Lee, Ehsan Adeli, Jitendra Malik, and Adrien Gaidon. It is not the journey but the destination: Endpoint conditioned trajectory prediction. In *ECCV*, 2020. 1

[43] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. Generating smooth pose sequences for diverse human motion prediction. In *ICCV*, 2021. 6

[44] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *ICCV*, 2017. 3

[45] Behnam Parsaeifard, Saeed Saadatnejad, Yuejiang Liu, Taylor Mordan, and Alexandre Alahi. Learning decoupled representations for human pose forecasting. In *ICCV*, 2021. 1, 2

[46] Xue Bin Peng, Pieter Abbeel, Sergey Levine, and Michiel Van de Panne. DeepMimic: Example-guided deep reinforcement learning of physics-based character skills. *TOG*, 2018. 2

[47] Xavier Puig, Eric Undersander, Andrew Szot, Mikael Dallaire Cote, Tsung-Yen Yang, Ruslan Partsey, Ruta Desai, Alexander William Clegg, Michal Hlavac, So Yeon Min, et al. Habitat 3.0: A co-habitat for humans, avatars and robots. *arXiv.org, 2310.13724*, 2023. 2

[48] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J Guibas. HuMoR: 3d human motion model for robust pose estimation. In *ICCV*, 2021. 8

[49] Davis Rempe, Zhengyi Luo, Xue Bin Peng, Ye Yuan, Kris Kitani, Karsten Kreis, Sanja Fidler, and Or Litany. Trace and pace: Controllable pedestrian animation via guided trajectory diffusion. In *CVPR*, 2023. 1, 2, 5

[50] Alexandre Robicquet, Alexandre Alahi, Amir Sadeghian, Bryan Anenberg, John Doherty, Eli Wu, and Silvio Savarese. Forecasting social navigation in crowded complex scenes. *arXiv.org, 1601.00998*, 2016. 2

[51] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration. In *ICCV*, 2021. 3

[52] Andrey Rudenko, Luigi Palmieri, Michael Herman, Kris M Kitani, Dariu M Gavrila, and Kai O Arras. Human motion trajectory prediction: A survey. *IJRR*, 2020. 1

[53] Seyedmorteza Sadat, Jakob Buhmann, Derek Bradely, Otmar Hilliges, and Romann M Weber. CADS: Unleashing the diversity of diffusion models through condition-annealed sampling. *arXiv.org, 2310.17347*, 2023. 1

[54] Bokui Shen, Fei Xia, Chengshu Li, Roberto Martín-Martín, Linxi Fan, Guanzhi Wang, Claudia Pérez-D'Arpino, Shyamal Buch, Sanjana Srivastava, Lyne Tchapmi, et al. iGibson 1.0: A simulation environment for interactive tasks in large realistic scenes. In *IROS*, 2021. 2

[55] Theodoros Sofianos, Alessio Sampieri, Luca Franco, and Fabio Galasso. Space-time-separable graph convolutional network for pose forecasting. In *ICCV*, 2021. 1, 2

[56] Simon Suo, Sebastian Regalado, Sergio Casas, and Raquel Urtasun. Trafficsim: Learning to simulate realistic multi-agent behaviors. In *ICCV*, 2021. 1, 2

[57] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. *ICLR*, 2023. 1

[58] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *JMLR*, 2008. 1

[59] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017. 3, 4

[60] Timo Von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *ECCV*, 2018. 2

[61] Timo von Marcard, Roberto Henschel, Michael J. Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3D human pose in the wild using imus and a moving camera. In *ECCV*, 2018. 6

[62] Jacob Walker, Kenneth Marino, Abhinav Gupta, and Martial Hebert. The pose knows: Video forecasting by generating pose futures. In *ICCV*, 2017. 6

[63] Jiashun Wang, Huazhe Xu, Jingwei Xu, Sifei Liu, and Xiaolong Wang. Synthesizing long-term 3d human motion and interaction in 3d scenes. In *ICCV*, 2021. 1, 2

[64] Jingbo Wang, Sijie Yan, Bo Dai, and Dahua Lin. Scene-aware generative network for human motion synthesis. In *ICCV*, 2021.

[65] Jingbo Wang, Yu Rong, Jingyuan Liu, Sijie Yan, Dahua Lin, and Bo Dai. Towards diverse and natural scene-aware 3D human motion synthesis. In *ICCV*, 2022. 1, 2

[66] Jingbo Wang, Ye Yuan, Zhengyi Luo, Kevin Xie, Dahua Lin, Umar Iqbal, Sanja Fidler, and Sameh Khamis. Learning human dynamics in autonomous driving scenarios. In *ICCV*, 2023. 1

[67] Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. HUMANISE: Language-conditioned hu-

man motion generation in 3D scenes. *arXiv.org, 2210.09729*, 2022. 2

[68] Chenyan Wu, Yukun Chen, Jiajia Luo, Che-Chun Su, Anuja Dawane, Bikramjot Hanzra, Zhuo Deng, Bilan Liu, James Z Wang, and Cheng-hao Kuo. Mebow: Monocular estimation of body orientation in the wild. In *CVPR*, 2020. 3

[69] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. Monocular total capture: Posing face, body, and hands in the wild. In *CVPR*, 2019. 3

[70] Kevin Xie, Tingwu Wang, Umar Iqbal, Yunrong Guo, Sanja Fidler, and Florian Shkurti. Physics-based human motion estimation and synthesis from videos. In *ICCV*, 2021. 1, 2, 6

[71] Pei Xu, Jean-Bernard Hayet, and Ioannis Karamouzas. Socialvae: Human trajectory prediction using timewise latents. In *ECCV*, 2022. 1, 2, 3

[72] Sirui Xu, Yu-Xiong Wang, and Liang-Yan Gui. Diverse human motion prediction guided by multi-level spatial-temporal anchors. In *ECCV*, 2022. 1, 2, 5, 6

[73] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *CVPR*, 2022. 4

[74] Xinchen Yan, Akash Rastogi, Ruben Villegas, Kalyan Sunkavalli, Eli Shechtman, Sunil Hadap, Ersin Yumer, and Honglak Lee. Mt-vae: Learning motion transformations to generate multimodal human dynamics. In *ECCV*, 2018. 6

[75] Wei Yang, Wanli Ouyang, Xiaolong Wang, Jimmy Ren, Hongsheng Li, and Xiaogang Wang. 3D human pose estimation in the wild by adversarial learning. In *CVPR*, 2018. 3

[76] Ze Yang, Sivabalan Manivasagam, Ming Liang, Bin Yang, Wei-Chiu Ma, and Raquel Urtasun. Recovering and simulating pedestrians in the wild. In *CoRL*, 2021. 1

[77] Ziyi Yin, Ruijin Liu, Zhiliang Xiong, and Zejian Yuan. Multimodal transformer networks for pedestrian trajectory prediction. In *IJCAI*, 2021. 3

[78] Ye Yuan and Kris Kitani. Diverse trajectory forecasting with determinantal point processes. *CVPR*, 2019. 1, 6

[79] Ye Yuan and Kris Kitani. DLow: Diversifying latent flows for diverse human motion prediction. In *ECCV*, 2020. 1, 2, 3, 4, 5, 6, 7, 8

[80] Ye Yuan, Umar Iqbal, Pavlo Molchanov, Kris Kitani, and Jan Kautz. Glamr: Global occlusion-aware human mesh recovery with dynamic cameras. In *ICCV*, 2022. 3

[81] Jimuyang Zhang, Minglan Zheng, Matthew Boyd, and Eshed Ohn-Bar. X-World: Accessibility, vision, and autonomy meet. In *ICCV*, 2021. 5

[82] Jimuyang Zhang, Ruizhao Zhu, and Eshed Ohn-Bar. SelfD: Self-learning large-scale driving policies from the web. In *CVPR*, 2022. 6

[83] Siwei Zhang, Qianli Ma, Yan Zhang, Zhiyin Qian, Marc Pollefeys, Federica Bogo, and Siyu Tang. EgoBody: human body shape, motion and social interactions from head-mounted devices. *ECCV*, 2021. 2

[84] Weicheng Zhang, Hao Cheng, Fatema T Johora, and Monika Sester. Forceformer: Exploring social force and transformer for pedestrian trajectory prediction. *arXiv.org, 2302.07583*, 2023. 1

[85] Xiaohan Zhang, Bharat Lal Bhatnagar, Sebastian Starke, Vladimir Guzov, and Gerard Pons-Moll. COUCH: towards controllable human-chair interactions. In *ECCV*, 2022. 2

[86] Yan Zhang, Michael J Black, and Siyu Tang. We are more than our joints: Predicting how 3D bodies move. In *ICCV*, 2021. 6

[87] Xingyi Zhou, Qixing Huang, Xiao Sun, Xiangyang Xue, and Yichen Wei. Towards 3D human pose estimation in the wild: a weakly-supervised approach. In *ICCV*, 2017. 3