Social Diffusion: Long-term Multiple Human Motion Anticipation

Julian Tanke*[1], Linguang Zhang[2], Amy Zhao[2], Chengcheng Tang[2], Yujun Cai[2], Lezi Wang[2], Po-Chen Wu[2], Juergen Gall[1,3], and Cem Keskin[2]

[1]University of Bonn [2]Reality Labs Research
[3]Lamarr Institute for Machine Learning and Artificial Intelligence
{tanke,gall}@iai.uni-bonn.de
{linguang,xamyzhao,chengcheng.tang,yujuncai,wanglezi,pochenwu,cemkeskin}@meta.com

Abstract

We propose Social Diffusion, a novel method for short-term and long-term forecasting of the motion of multiple persons as well as their social interactions. Jointly forecasting motions for multiple persons involved in social activities is inherently a challenging problem due to the interdependencies between individuals. In this work, we leverage a diffusion model conditioned on motion histories and causal temporal convolutional networks to forecast individually and contextually plausible motions for all participants. The contextual plausibility is achieved via an order-invariant aggregation function. As a second contribution, we design a new evaluation protocol that measures the plausibility of social interactions which we evaluate on the Haggling dataset, which features a challenging social activity where people are actively taking turns to talk and switching their attention. We evaluate our approach on four datasets for multi-person forecasting where our approach outperforms the state-of-the-art in terms of motion realism and contextual plausibility.

1. Introduction

Understanding and anticipating social interactions in groups of people is a challenging and highly relevant topic [35, 9, 32, 45, 47, 5, 34]. For instance, it is essential for socially-compliant robots [44], but it is also relevant for neuroscience and social sciences since it allows to develop computational models on how the behavior of other persons is perceived and how it changes the own behavior.

Forecasting realistic social interactions, however, is very challenging for two reasons. First, social interactions tend to last for tens of seconds [33] or even minutes - much longer than the prediction from most of the existing human motion anticipation models [23, 69, 3, 22, 27, 42, 49, 11, 25, 37, 4, 14, 41, 40]. Second, social interactions consist of interdependent motions [54, 36], which requires modeling the relationships among all individuals. For example, in conversational turn-taking, a person's turn to talk highly depends on the start/end of the others' speaking. While multi-person motion anticipation has emerged as a new topic, current approaches [23, 69, 3] do not pay much attention on complex social interactions. For instance, they do not preserve the social role of individuals in a group such that the interactions become socially implausible over time.

To address the limitations of existing models, we propose Social Diffusion to predict motions of multiple people and ensure contextually plausible interactions, as shown in Fig. 1. To this end, we learn the distribution of human motion by leveraging a diffusion model [39, 28, 52, 58, 24, 59, 62, 73]. To enforce information exchange among people, which is critical to predicting contextually plausible interactions, we introduce an order-invariant aggregation function to aggregate motion features from all people. For inference, we feed back the input sequence to the signal during the reverse-diffusion steps to condition the motion generation on the past motion. Our method is fully convolutional which allows us to generate sequences of arbitrary size. This allows us to not just forecast the next few seconds of an input motion but also to forecast social interactions that last longer. Furthermore, our approach is very flexible in the sense that the number of persons during training and inference can differ. To the best of our knowledge, our approach represents the first diffusion model that produces multi-person motions at the same time.

As a second contribution, we propose a new evaluation protocol for social interactions based on Symbolic Social Cues, which measures whether the forecast motion is socially plausible. Our key observation is that the probabilities of transitions between social interaction states are highly

*Work done partially while Julian was at Reality Labs Research.

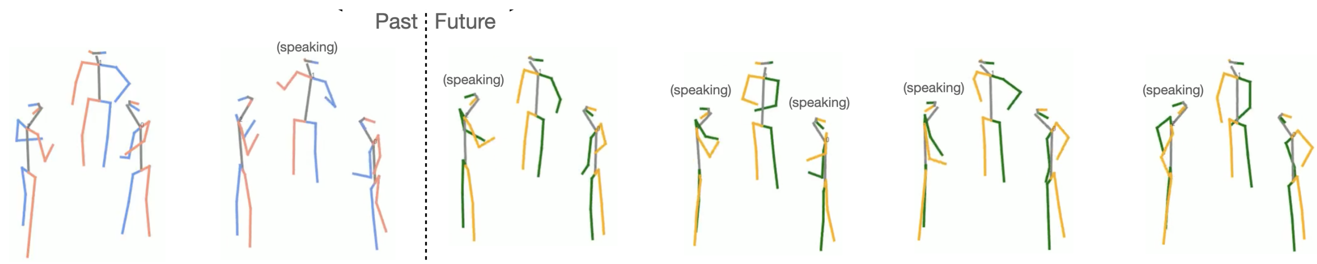


Figure 1: We propose an approach for multi-person motion anticipation: given a sequence of human social interactions (blue-red skeletons), the proposed model forecasts multi-person motions where the social roles are preserved and the interactions are socially plausible.

correlated with the plausibility of predicted social interactions. In a conversation, for example, a person usually starts talking only when a peer stops talking. To evaluate predicted motions, we first build the state transition graph by extracting states from the motions. We then treat the state transition graph as a probability distribution and compare it to the real data distribution.

For evaluation, we use the Haggling dataset [30] which comprises 175 videos of well-defined triadic social interactions. In contrast to other existing multi-person human motion datasets [43, 66, 69], the persons have different social roles that impact their behavior. We furthermore evaluate our approach on the MuPoTS-3D [43], 3DPW [67], and CMU-Mocap [1, 69] dataset. On all four datasets, our approach outperforms the state of the art for multi-person human motion forecasting.

In summary, our contribution is two-fold:

1. We propose Social Diffusion, the first stochastic multi-person motion anticipation model that outperforms the state of the art on common multi-person motion anticipation datasets.

2. We propose a novel social interaction evaluation protocol that considers not only the validity of poses but also the plausibility of social interactions.

2. Related Work

Single Person Human Motion Prediction Human motion prediction, which typically refers to generating motion sequences given a prefix segment, has been extensively studied in the past few years. Most recent papers focus on single person 3D predictions. Due to the inherent temporal nature of human motion, various temporal-based methods, such as recurrent neural networks (RNNs) [18, 22, 27, 42, 49, 48, 61, 72], temporal convolutional networks [11, 25, 37], transformers [4, 14, 12] and graph neural networks (GNNs) [14, 41, 40] have been used for this task. For long-term predictions, auto-regressive models that operate in the discrete space [46, 38, 61] have shown success, where a long prediction sequence can be obtained without converging to mean poses. Recently, anchor-based methods [16, 68] have been proposed; these focus on forecasting characteristic anchor poses rather than the entire sequence auto-regressively. To achieve better interactions with environments, [7, 70, 68, 74] proposed various ways to include contextual information into human motion predictions. Since generated motions can be controlled with a high-level guidance such as action class or text, some approaches [50, 13] used Variational Auto-Encoders [58] to solve this problem. Going beyond single human prediction, we predict the motions and interactions of multiple humans.

Multi-person Anticipation Modeling multi-person interactions has been a long standing problem [46, 23, 2, 65, 20, 30, 46, 63, 75]. For instance, DR^{2}N [60] predicts the activities of multiple people given a past video sequence. For a given frame, personal relationships between candidates are estimated using a graph attention network (GAT) [65] while temporal relationships are predicted using recurrent neural networks. Recently, Wang *et al.* [69] addressed multi-person 3D motion trajectory prediction via a Multi-Range Transformers framework. Guo *et al.* [23] introduced Cross-Interaction Attention to jointly model highly expressive dance sequences. Joo *et al.* [30] introduced a triadic haggling game for social signal prediction, based on Panoptic Studio [29, 31]. In their work they predict the motion of a single person given the motion of the others. Similarly, [46] predicts motions based on the other actors' behavior. Crucially, both methods only predict the motion of a single individual, given other persons social signals.

Diffusion Models Diffusion models [24, 57, 59] belong to the family of probabilistic generative models, which convert the training data successively to Gaussian noise, and then learn to recover the data by reversing this noising process. Diffusion models have emerged as powerful deep generative models with breakthroughs in many applications, including image synthesis [24, 15], segmentation [10], and natural language processing [8, 26]. For conditioned generation, [15] introduced classifier-guided diffusion, and [55] takes the inpainted images as denoised by the model. More recently, [62, 73] have suggested diffusion models for motion generation; however, they are limited to single human

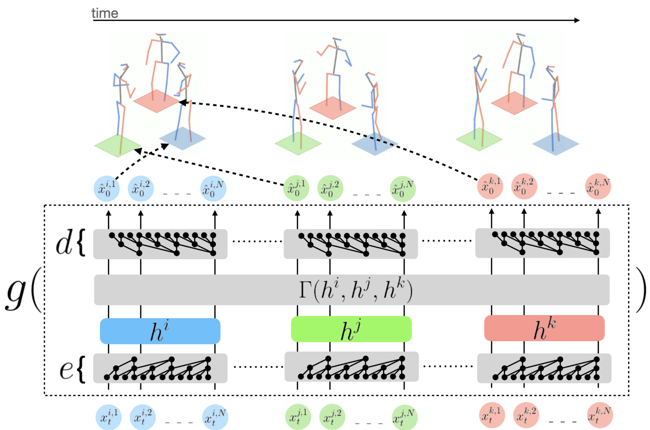


Figure 2: Model overview: the reverse diffusion process g consists of a fully convolutional causal encoder e and a fully convolutional causal decoder d that produces a denoised motion sequence $\hat{\mathbf{x}}_0^i \equiv \hat{\mathbf{x}}_0^{i,1:N}$ for person i, given the bottleneck state h^i and the aggregation function $\Gamma(\{\mathbf{h}_j \mid j \in 1, ..., p\})$ over all people in the scene.

3D motion prediction. To the best of our knowledge, we are the first to build a stochastic multi-person motion anticipation model that can predict very long-term future motions.

3. Social Diffusion Model

As illustrated in Fig. 1, we aim to forecast the motion of multiple persons that interact with each other. The forecast motion should be realistic and socially plausible. For instance, not all persons should talk at the same time. Formally, we represent a human motion sequence with p people of length N as $\mathbf{X}^{1:N} \in \mathbb{R}^{N \times p \times \delta}$ where δ represents the dimension of the individual pose vector at a given frame. Our goal is then to predict the future motion $\hat{\mathbf{X}}^{n+1:N}$ for all people, given their past motions $\mathbf{X}^{1:n}$:

$$\hat{\mathbf{X}}^{n+1:N} = \text{SDM}(\mathbf{X}^{1:n}). \tag{1}$$

Before we describe the proposed Social Diffusion Model (SDM) in Section 3.2, we will briefly describe a generic diffusion model [24] in Section 3.1. In Section 4, we will then introduce the social interaction evaluation protocol.

3.1. Diffusion Model

A latent representation $\mathbf{X}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$ is obtained via a T step Markov Gaussian noising process $q(\mathbf{X}_T|\mathbf{X}_0)$ where $\mathbf{X}_0 \equiv \mathbf{X}^{1:N}$ is a real motion sequence from the training set. The Markov Gaussian noising process can be written in closed form:

$$q(\mathbf{X}_t|\mathbf{X}_0) = \mathcal{N}(\mathbf{X}_t; \sqrt{\alpha_t}\mathbf{X}_0, (1-\alpha_t)\mathbf{I}) \tag{2}$$

where $\alpha_t \in (0, 1)$ is a step-dependent fixed hyperparameter. To sample from the generative model, we learn to invert the noising step using the generator function g:

$$\hat{\mathbf{X}}_{0,t} = g(\mathbf{X}_t, t) \tag{3}$$

The key contribution of our model is the novel generator function g, which models social interactions over time and will be described in Section 3.2. Following [24, 62, 51], the loss during training is defined by:

$$\mathcal{L} = \mathbb{E}_{\mathbf{X}_0 \sim p(\mathbf{X}), t \sim [1,T]}\Big[\big|\big|\mathbf{X}_0 - g(\mathbf{X}_t, t)\big|\big|\Big] \tag{4}$$

For inference, we reverse-iterate over Equation (3), starting at sampling step T and latent representation $\mathbf{X}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$. At each iteration t, we slowly denoise the motion sequence using (2) and (3):

$$\hat{\mathbf{X}}_{0,t-1} = g(q(\hat{\mathbf{X}}_{t-1}|\hat{\mathbf{X}}_{0,t}), t-1) \tag{5}$$

The final denoised motion is obtained when $t = 1$.

3.2. Multi-person Motion Generator

The diffusion model described in Section 3.1 produces unconditioned motion. In order to use the model for motion forecasting, we need to condition the model on the observed motion sequence of all persons $\mathbf{X}_0^{1:n} = \mathbf{X}^{1:n}$. To this end, we modify the inference sampling (5) to also include past motion as follows:

$$\hat{\mathbf{X}}_{0,t-1} = g(q(\hat{\mathbf{X}}_{t-1}|\mathbf{X}^{1:n} \cup \hat{\mathbf{X}}_{0,t}^{n+1:N}), t-1) \tag{6}$$

The reverse diffusion process $g(\mathbf{X}_t, t)$ consists of three components, a causal temporal convolutional encoder e, a causal temporal convolutional decoder d, and an order-invariant function Γ that aggregates the interaction of the different persons, see Figure 2.

The encoder e and the decoder d process each individual sequence independently while Γ ensures that information flows between all persons in the scene. Formally, given $\mathbf{x}_t^i = \hat{\mathbf{X}}_t^{i,1:N} \in \mathbb{R}^{N \times \delta}$, which is the motion sequence for a single person i at scheduled noising step t, we obtain for each person i in the scene a bottleneck encoding $\mathbf{h}^i$:

$$\mathbf{h}^i = e(\mathbf{x}_t^i, t) \;\; \forall i \in 1, ..., p. \tag{7}$$

The encoder e consists of a 4-layer temporal convolutional network where each layer progressively reduces the temporal resolution by half via striding. In each layer, the noising step t is fed via sinusoidal positional encoding [64]. To produce the denoised motion $\hat{\mathbf{x}}_0^i$, the decoder d can be utilized as follows:

$$\hat{\mathbf{x}}_0^i = d\big(\mathbf{h}^i, t, \Gamma(\mathbf{h}^j \; \forall j \in 1, ..., p)\big) \tag{8}$$

where d is a 4-layer temporal convolutional network and each layer progressively doubles the temporal resolution via linear upsampling. As for the encoder, the noising step t is fed via sinusoidal positional encoding to each layer. In addition, the output of the order-invariant aggregation function Γ is concatenated to $\mathbf{h^i}$ before passing it to the first convolutional layer. The estimated motion sequences $\hat{\mathbf{x}}_0^i$ of each person i at noising step t are then concatenated to obtain $\hat{\mathbf{X}}_{0,t}$ and the approach proceeds to the next step $t-1$.

The order-invariant aggregation function Γ passes information from other people in the scene. In our experiments, we evaluate two aggregation functions, averaging ($\Gamma_{\mathbb{E}}$) over all people and multi-headed attention [64] (Γ_{attn}):

$$\Gamma_{\text{attn}}(\mathbf{H}) = \text{MultiHead}(\mathbf{H}) \tag{9}$$

$$\Gamma_{\mathbb{E}}(\mathbf{H}) = \frac{1}{p}\sum_{i=1}^{p}\mathbf{h}^i \tag{10}$$

where $\mathbf{H} = \{h^i\}_{i=1}^p$ are the embeddings of all layers of the encoder and for all people in the scene. $\text{MultiHead}(\mathbf{H})$ calculates the self-attention across all people for a given frame.

3.3. Implementation Details

We follow state-of-the-art diffusion models [24, 39, 62] and use the cosine variance schedule. We set the number of diffusion steps to $T = 1000$. The encoder e consists of four layers of convolutional blocks with kernel size 3 and stride 2. The decoder d consists of four layers of convolutional blocks with additional upsampling layers that upsample the input sequence by factor two using linear interpolation. We standardize the training data to have zero mean and standard deviation one. We normalize all poses by splitting pose and global translation: each pose is transformed into a hip-centric coordinate frame and the pose is concatenated with the global rotation and translation to form a δ dimensional pose vector.

4. Symbolic Social Cues Protocol

We are interested in anticipating the social interactions among multiple people. Multi-person social interactions consist of several intricate and complex behaviours such as paying attention to a specific person [56] and turn-taking [54, 36], which usually take tens of seconds or even minutes. Current state-of-the-art multi-person motion anticipation methods [3, 69] calculate the Mean Per Joint Positional Error (MPJPE) using the ground-truth sequence, which is only meaningful for short time horizons of around one second [6, 21, 53, 61, 71]. More important, however, is that it does not measure the realism of social interaction.

We thus propose the Symbolic Social Cues Protocol (SSCP), which divides the social interactions into a set of discrete interaction classes. In SSCP, we define a social signal function

$$C^{1:N} = s(\mathbf{X}^{1:N}) \tag{11}$$

which takes as input a multi-person motion sequence $\mathbf{X}^{1:N} \in \mathbb{R}^{N\times p\times\delta}$ and produces a discrete symbolic representation $C^{1:N} = \{c_n\}_{n=1}^N$, where $c_n \in \{1, ..., m\}$ and m represents the total number of symbolic states. A symbolic state is a unique summary of the current state of interaction, e.g., a person is talking and another person is listening. Given a test set $\mathcal{X} = \{\mathbf{X}_i^{1:N}\}_{i=1}^K$ with K sequences, we can now calculate the probability distribution p_{sscp} over the social state transitions.

$$p_{\text{sscp}}(\mathcal{X}) = \frac{1}{\zeta}\sum_{i=1}^{K}\text{stm}(s(\mathbf{X}_i^{1:N})) \tag{12}$$

where $\text{stm}(C_i^{1:N})$ produces the $m \times m$ state transition matrix for the discrete symbolic sequence $C^{1:N}$ and $\zeta = \sum_{i=1}^K \sum_{m',m''} \text{stm}(C^{1:N})_{m',m''}$ is a normalization constant to ensure that p_{sscp} is a valid probability distribution.

To evaluate a motion anticipation model f, we predict the future motion for all K test sequences from a fixed start frame n until the end of the sequence N:

$$\hat{\mathcal{X}}^{n+1:N} = \{f(\mathbf{X}_i^{1:n})\}_{i=1}^K \tag{13}$$

We can now calculate the distance between the generated and ground-truth social motion distribution:

$$D_{\text{JSD}}\big(p_{\text{sscp}}(\mathcal{X}^{n:N}), p_{\text{sscp}}(\hat{\mathcal{X}}^{n:N})\big). \tag{14}$$

D_{JSD} is the squared Jensen–Shannon distance [17, 19]:

$$D_{\text{JSD}}(p||q) = \sqrt{\frac{\left(D_{\text{KL}}(p||\frac{p+q}{2}) + D_{\text{KL}}(q||\frac{p+q}{2})\right)}{2}} \tag{15}$$

where p and q are probability distributions and D_{KL} is the Kullback-Leibler divergence. Note that we compare the generated motion $\hat{\mathcal{X}}^{n+1:N}$ to the test set $\mathcal{X}^{n+1:N}$ with the same start frame as the data distribution might shift across time.

5. Experiments

We evaluate our approach on four datasets. Following [69], we report the Mean Per Joint Positional Error (MPJPE) in global and local aligned coordinates for the multi-person human motion datasets MuPoTS-3D [43], 3DPW [67] and CMU-Mocap [1, 69]. MuPoTS-3D [43] contains recordings of 2 to 3 persons in workout settings. Interactions between the persons are rare. 3DPW [67] contains recordings of 1 to 2 persons and the sequences cover a wide range of different activities. The level of interactions range from no interaction and little interaction, like two persons walking, to close interactions such as dancing. CMU [69] combines the motion of different sequences from the CMU-Mocap dataset [1]. The composition of motion

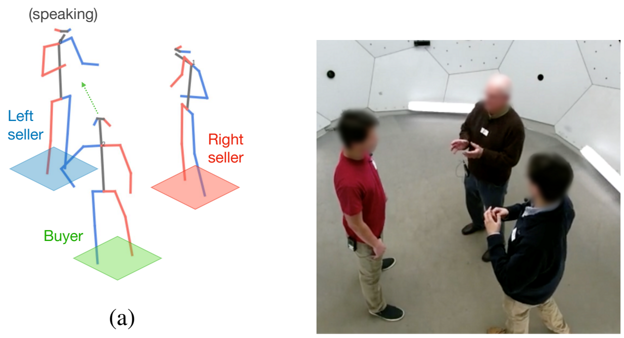


Figure 3: A sample frame from the Haggling dataset [30] for evaluating social interactions. (a): 3D poses in a haggling sequence. Blue limbs represent the left body side while red limbs represent the right body side. The buyer's attention is indicated as green arrow. (b): a sample video frame from a haggling sequence.

Method	CMU-Mocap			MuPoTS-3D			3DPW		
	1s	2s	3s	1s	2s	3s	1s	2s	3s
LTD [41]	1.37	2.19	3.26	1.19	1.81	2.34	4.67	7.10	8.71
HRI [40]	1.49	2.60	3.07	0.94	1.68	2.29	4.07	6.32	8.01
SP [2]	1.15	2.71	3.90	0.92	1.67	2.51	4.17	7.17	9.27
MRT [69]	0.96	1.57	2.18	**0.89**	1.59	2.22	3.87	6.12	7.83
Ours	**0.74**	**1.06**	**1.34**	1.15	**1.29**	**1.44**	**1.64**	**2.72**	**3.55**

Table 1: MPJPE $\downarrow$ in dm on different datasets.

sequences, however, does not reflect realistic interactions. For these datasets, 1 second of human motion is observed and 1-3 seconds need to be forecast. As our model is generative, we sample 8 samples for each test sequence and report the average over all 8 samples. The source code including trained evaluation models and the dataset transformation script is publicly available[1].

5.1. Haggling Forecasting Dataset

Since the three datasets contain multiple persons, but very few social interactions, we prepared a new dataset for multi-person forecasting in the context of social interactions. For this, we utilize the Haggling dataset [30] where 122 participants play a social game with two sellers trying to sell their products to a buyer. Each game lasts one minute and contains interesting triadic interactions such as turn-taking and attention changes. A sample scene is shown in Figure 3. The dataset consists of 135 training sequences and 40 test sequences, sampled at 30Hz. Some of the 3D poses in the dataset are noisy as they have been estimated [29, 31] and we manually correct them. In total, the dataset consists of $234,907$ training and $69,951$ test frames, each with three people. For more details, please see the supplementary material.

For evaluation, we take the first 10% of a sequence as observation and forecast the remaining 90% of the sequence, but we also evaluate the motion at intermediate frames ranging from frame 1 to frame 1300. It needs to be noted that long-term forecasting is highly relevant for neuroscience and social sciences since it allows to develop computational models on how the behavior of other persons is perceived and how it changes the own behavior. For measuring the plausibility of the forecast motion of each individual, we use the Normalized Directional Motion Similarity (NDMS) [61] since the measure not only considers static poses but also the motion of the forecast sequence. Furthermore, it can be applied to sequences of any length. NDMS, however, does not measure if the social interactions are plausible.

For evaluating the social motion quality, we utilize the proposed Symbolic Social Cues Protocol described in Section 4. To this end, we need to specify the classes of social interactions. A haggling activity is composed of the sellers trying to convince the buyer to purchase their products and the buyer switching attention between the sellers. Throughout the game, certain social patterns emerged [56, 36]:

1. Most of the time, only a single person speaks
2. For almost all frames, at least one person is talking
3. The sellers speak roughly the same amount of time while the buyer seldom talks
4. The buyer pays attention (looks at) to whoever talks
5. The sellers take turns to speak but sometimes interrupt each other

Given the well-defined structure of the task and the emerging social behaviors, we reduce the haggling game to two key signals:

- **talking**: defines who is talking
- **attention**: defines who of the two sellers has the buyer's attention

Given all possible combinations (e.g., both sellers can talk at the same time, or nobody talks), we end up with 16 possible states for each frame and formulate the social interactions as a symbolic representation over time. Note that we have to distinguish between left/right seller to catch events such as attention switching. Please see the supplementary material for more details.

We define the social signal function $s(\mathbf{X}^{1:N})$ as Equation (11), which takes a sequence of multi-person motion $\mathbf{X}^{1:N}$ as input and generates one of the 16 distinct states per frame. For the social signal function s to work on any haggling motion sequence, we need to determine three pieces of information:

1. who the buyer is,
2. whom the buyer is paying attention to,
3. whether someone is speaking.

[1] https://github.com/jutanke/social_diffusion

Method	CMU-Mocap						MuPoTS-3D						3DPW					
	Root			Pose			Root			Pose			Root			Pose		
	1s	2s	3s	1s	2s	3s	1s	2s	3s	1s	2s	3s	1s	2s	3s	1s	2s	3s
LTD [41]	0.97	1.73	2.62	0.98	1.21	1.37	0.89	1.39	1.91	0.88	1.14	1.31	4.28	6.79	8.41	1.54	1.76	1.98
HRI [40]	0.96	2.06	3.11	1.05	1.37	1.58	**0.66**	1.30	2.16	0.73	1.07	1.30	3.67	6.42	8.64	$\underline{1.43}$	$\underline{1.75}$	1.94
SP [2]	0.96	2.01	2.96	1.03	1.41	1.71	0.96	1.38	2.21	0.72	1.08	1.30	3.76	6.86	9.07	1.60	1.95	2.15
MRT [69]	**0.60**	$\underline{1.12}$	$\underline{1.71}$	$\underline{0.79}$	$\underline{1.05}$	$\underline{1.22}$	$\underline{0.67}$	**1.25**	$\underline{1.86}$	$\underline{0.69}$	$\underline{0.99}$	$\underline{1.19}$	$\underline{3.42}$	$\underline{5.69}$	$\underline{7.30}$	1.52	$\underline{1.75}$	$\underline{1.93}$
Ours	$\underline{0.72}$	**1.10**	**1.44**	**0.38**	**0.46**	**0.49**	1.14	$\underline{1.28}$	**1.42**	**0.59**	**0.64**	**0.67**	**1.66**	**2.76**	**3.59**	**0.94**	**1.03**	**1.06**

Table 2: MPJPE ↓ in dm for root joint and pose. The lowest error is in bold and the second lowest is underscored.

Frame	1	5	10	15	20	25	30	60	120	250	500	750	1000	1300
MRT [69]	0.624	0.278	0.194	0.212	0.224	0.215	0.215	0.218	0.205	0.180	0.129	0.079	0.062	0.047
Ours ($\Gamma_{\varnothing}$)	**0.644**	**0.280**	**0.206**	$\underline{0.215}$	$\underline{0.225}$	0.226	0.229	**0.233**	**0.229**	**0.226**	0.229	0.227	0.225	**0.226**
Ours (Γ_{attn})	0.639	**0.280**	0.199	0.213	0.224	**0.229**	**0.232**	0.227	**0.229**	$\underline{0.223}$	**0.233**	$\underline{0.228}$	**0.233**	0.223
Ours* ($\Gamma_{\mathbb{E}}$)	$\underline{0.640}$	0.279	$\underline{0.204}$	**0.216**	**0.227**	$\underline{0.228}$	$\underline{0.230}$	**0.233**	0.227	0.222	$\underline{0.230}$	**0.229**	$\underline{0.230}$	**0.226**

Table 3: Per-frame average NDMS ↑ score on the Haggling dataset. The highest score is in bold and the second highest is underscored.

To solve (1) we train a simple buyer detection network, consisting of three layers of bi-directional Gated Recurrent Units, which gets as input a haggling motion and outputs the likelihood of each participant being the buyer. In Table 5, we report our accuracy of this approach. We see that the buyer detector correctly identifies the buyer all the time on the test set.

For (2), we define the buyer's attention as whomever they look at, which can be easily calculated from the 3D body pose:

$$\underset{i \in \{\text{left}, \text{right}\}}{\operatorname{argmin}} \left[n^T \left(\frac{d_i}{|d_i|} \right) \right] \tag{16}$$

where d_{left} and d_{right} are the directional vectors from the buyer nose to the left and right seller nose, respectively, projected onto the ground plane and n is the 2D unit vector that is perpendicular to the left eye→right eye vector of the buyer, projected onto the ground plane.

Last but not least, to determine if someone is speaking we utilize an off-the-shelf action classification network consisting of three layers of bi-directional Gated Recurrent Units. For training, we use the annotation of the Haggling dataset [30] which indicates if a person is speaking or not. The classifier achieves 87% accuracy in speech detection on the test set.

5.2. Multi-Person Forecasting

We first report the results for the multi-person human motion datasets MuPoTS-3D [43][2], 3DPW [67][1] and CMU-Mocap [1, 69]. We follow [69] and report the Mean Per Joint Positional Error (MPJPE) using global coordinates in Table 1. Our approach outperforms the methods LTD [41], HRI [40], SP [2], and MRT [69] by a large margin. While some methods perform better for the first second on MuPoTS-3D [43], our approach achieves a much lower error for all other settings and datasets. On the most difficult dataset 3DPW [67], the error is reduced by 57.6%, 55.6%, and 54.7% for 1, 2, and 3 seconds, respectively. As in [69], we also report the error of the position of the root joint and pose error in local coordinates, i.e., setting the root position for all frames to zero, in Table 2. The results show that our approach forecasts by far the most accurate poses and outperforms the state of the art by a large margin on all datasets. Only the position of the root joint is slightly better estimated by other methods at the beginning of the datasets MuPoTS-3D and CMU-Mocap. At 3 seconds, our approach also achieves the lowest root joint error for all three datasets.

[2]Data access and processing was conducted at University of Bonn

5.3. Multi-Person Forecasting in the Context of Social Interactions

For the remaining experiments, we evaluate our approach on the newly prepared Haggling dataset since it contains more social interactions as the other datasets. We compare our approach to Multi-Range Transformers (MRT) [69], which performed better than other approaches in Section 5.2. We used the publicly available source code and adjusted the approach to work with 30Hz. We kept all other settings as is.

We first report the per-frame NDMS (higher is better) at different frames in Table 3. Our experiments show that MRT [69] is capable of generating realistic motion for a few seconds. However, after 120 frames ($4s$) the NDMS score drops significantly. This is caused by the auto-regressive motion forecasting strategy adopted by MRT, which results in error accumulation over time. In contrast, our method continues to predict motions with good NDMS scores well into the future. As discussed in Section 3.2, we compare different variants of the aggregation function Γ. In terms of Γ, there are no major differences in terms of NDMS, but they all perform much better than MRT [69]. In Table 4, we report the average NDMS score over the entire forecast sequence. We observe that $\Gamma_{\varnothing}$, i.e., using no aggregation

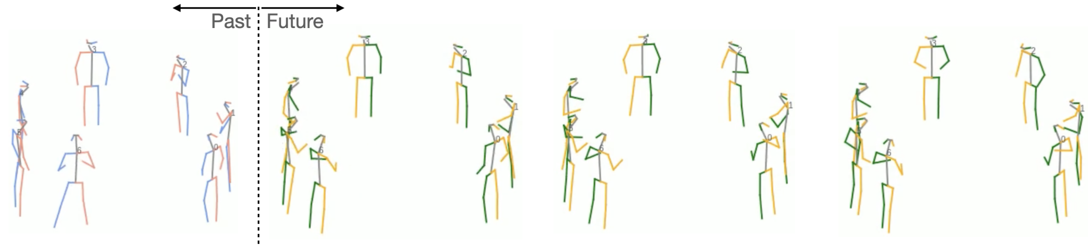
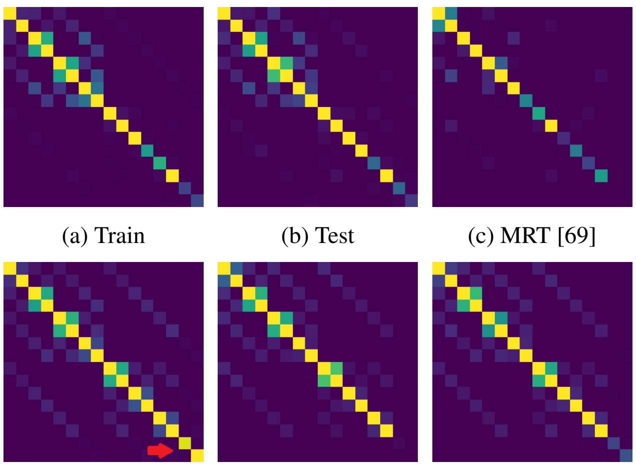


Figure 4: Our model is capable of generating realistic motion for 7 people from the *Ultimatum* sequence of Panoptic Studio [29], even though it was trained only on the triadic Haggling dataset. The *Ultimatum* sequence shares similarities with the Haggling dataset such as persons taking turns and talking to each other.



(a) Train (b) Test (c) MRT [69]

(d) Ours ($\Gamma_{\varnothing}$) (e) Ours (Γ_{attn}) (f) Ours ($\Gamma_{\mathbb{E}}$)

Figure 5: State transition matrices for the states defined in Section 5.1 for the ground-truth of the train (a) and test set (b). The other transition matrices are obtained by the forecast of the current state-of-the-art model MRT [69] on the test set (c) and our model without context (d), with context via attention (e), and with context via averaging (f). The more similar the transition matrix is to the test set (b), the closer it matches the test motion.

	Train	[69]	Ours ($\Gamma_{\varnothing}$)	Ours (Γ_{attn})	Ours*($\Gamma_{\mathbb{E}}$)
NDMS ↑	-	0.1015	**0.2301**	0.2270	<u>0.2297</u>
SSCP ↓	0.0999	0.4839	0.3576	<u>0.3278</u>	**0.3252**

Table 4: Per-frame average NDMS ↑ and SSCP ↓ on the Haggling dataset.

information yields a slightly higher NDMS score than the other versions, but the differences are very small. However, Γ_{attn} and $\Gamma_{\mathbb{E}}$ forecast much more plausible social interactions (lower SSCP) as we will discuss next.

Test set	MRT [69]	Ours ($\Gamma_{\varnothing}$)	Ours (Γ_{attn})	Ours* ($\Gamma_{\mathbb{E}}$)
1.0	0.3250	0.8812	<u>0.8875</u>	**0.8938**

Table 5: Accuracy ↑ of the buyer detection network. For MRT [69], the model randomly selects a person to be the buyer, as there is a $1/3$ chance of selecting the buyer with random chance.

5.3.1 Social Motion Evaluation

The SSCP scores are presented in Table 4 where a lower value corresponds to more plausible forecast social interactions. All our proposed variants outperform the state-of-the-art approach MRT [69]. Using no aggregation information ($\Gamma_{\varnothing}$) performs worse than the aggregation functions Γ_{attn} (9) and $\Gamma_{\mathbb{E}}$ (10), which is expected since the aggregation function passes information from other people in the scene. The average aggregation ($\Gamma_{\mathbb{E}}$) performs slightly better than the more complex multi-headed attention approach (Γ_{attn}). We conjecture that averaging bottleneck encodings over all people introduces an inductive bias to pay the same attention to everyone, which works well for modelling the haggling game. For completeness, we also report the SSCP score of the training set.

We can draw some insights about what motion each model generates by looking at the state probability transition matrices in Figure 5. For example, MRT [69] (Figure 5d) produces mostly self-loops (diagonal of transition matrix) indicating that the motion gets stuck over time. When no context information is provided ($\Gamma_{\varnothing}$), our method produces motions where all three people are talking at the same time, as can be seen in Figure 5d, where the last two entries (red arrow) in the transition matrix represent states with all three people talking. This is sensible as the model sees two sellers and only one buyer during training and thus it is more likely to produce motion that resembles a seller, who talks most of the time. When the context is provided, our approach overcomes this limitation as expected and rarely produces motion where all three people are talking at the same time as shown in Figures 5e and 5f.

These observations are also confirmed when measuring the buyer detection accuracy on the forecast motion, which is reported in Table 5. The detector fails to identify the correct buyer in the sequences that are forecast by MRT [69] and it nearly chooses the buyer at random with $1/3$ accuracy. This confirms that MRT does not forecast socially consistent sequences where the social role of the persons, namely buyer or seller, is preserved. In contrast, our method predicts motion where the buyer can be easily determined most of the time. As for the SSCP scores reported in Table 4, $\Gamma_{\mathbb{E}}$ performs best.

In summary, the aggregation function $\Gamma_{\mathbb{E}}$ outperforms the other aggregation functions on the Haggling dataset [30] as it produces the most socially plausible motion according to our Symbolic Social Cues Protocol while also generating highly plausible 3D body motion.

5.4. Ablation Study

Average velocity over time Freezing or unrealistically expanding motion are common failures in human motion anticipation. While NDMS [61] penalizes in contrast to MPJPE errors in the velocity, visualizing the average motion velocity can give interesting insights. In Figure 6, we plot the average velocity over all frames for all our summary function variants, the test set, and the state-of-the-art method MRT [69]. Note that the beginning of the sequence has a higher velocity due to people walking into the scene. We observe that MRT suffers from error accumulation caused by the auto-regressive inference scheme. The velocity produced by our motion tightly follows the test set velocity for roughly 250 frames after which the test set velocity is slightly larger. We attribute this to the higher degree of stochasticity of real motion, which results in sudden jerks and swings that increase the average velocity.

NDMS score over time In Figure 7, we visualize how the NDMS scores of all proposed variants and MRT evolve over time. For reference, we also calculate the NDMS score of the training data which is guaranteed to be realistic. Note that NDMS is 1 for the observed part of the test sequences. As shown in the figure, our method achieves almost the same level of realism as the training data while the quality of MRT slowly degenerates over time.

Anticipating more than three people We have trained and evaluated our model on the Haggling dataset where each sequence consists of triadic interaction. However, the fully convolutional nature of our approach as well as the order-invariance of the summarization function Γ allow us to forecast any number of people. To demonstrate this capability, we predict 7 people from the *Ultimatum* sequence of Panoptic Studio [29] using only the model trained on the Haggling dataset. This works well because the Haggling and *Ultimatum* sequences share many social behaviors, such as turn-taking, talking, and paying attention while

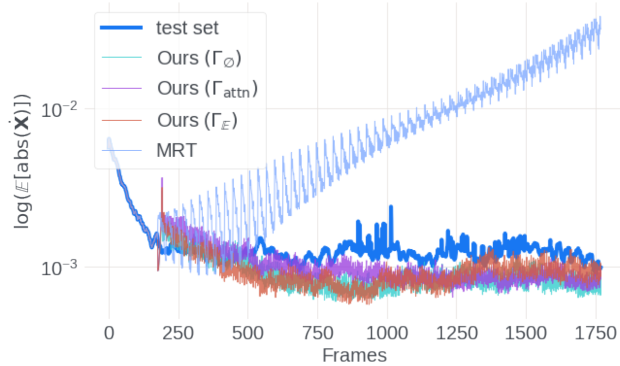


Figure 6: Average velocity over time for the entire test motion $\mathcal{X}$ and generated motions $\hat{\mathcal{X}}$. The x-axis represents the frames while the y-axis represents the (log) average velocity of the data.

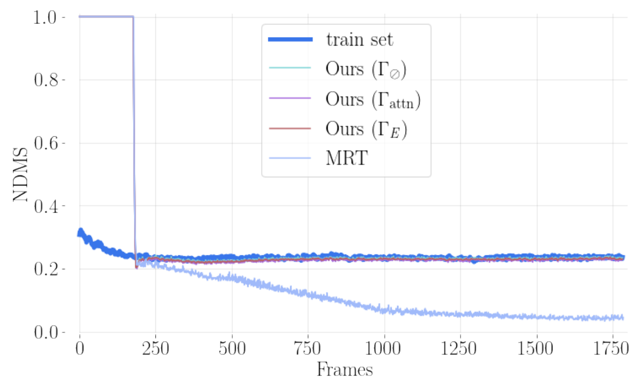


Figure 7: Average NDMS [61] score ↑ of all proposed variants, MRT, and the train set over time.

standing in a circle. Our results can be seen in Figure 4 where our model is able to predict realistic motion for 7 people, even though it was trained only on the triadic Haggling dataset. More results are provided in the supplementary material.

6. Conclusion

In this work, we present Social Diffusion, a stochastic multi-person motion anticipation model. The approach not only forecasts realistic motions on the individual level, but also plausible social interactions where the social roles of individuals are preserved over time. The approach is very flexible. It can be used for short and long-term forecasting and can be applied to larger groups than observed during training. As a second contribution, we proposed a new evaluation protocol to measure the realism of forecast social interactions. We furthermore derived a dataset for multi-person social interaction forecasting from the Haggling dataset [30] where the persons have different social roles that impact their behavior. We evaluated our approach on four multi-person datasets and demonstrated that our approach outperforms the state-of-the-art for short-term and long-term anticipation both in realism of forecast motion and social interaction. The approach has still some limitations. For instance, the global positions of the root joints can be better estimated. Future directions also include extending the model to predict motions of dynamic groups of people, e.g., at a cocktail party where any individual can freely disengage from the current conversation group and join another one.

Acknowledgement

Juergen Gall has been supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) GA 1927/4-2 (FOR 2535 Anticipating Human Behavior), by the Federal Ministry of Education and Research (BMBF) under grant no. 01IS22094E WEST-AI, the project iBehave (receiving funding from the programme "Netzwerke 2021", an initiative of the Ministry of Culture and Science of the State of Northrhine Westphalia), and the ERC Consolidator Grant FORHUE (101044724). The sole responsibility for the content of this publication lies with the authors.

References

[1] CMU. Carnegie-Mellon Mocap Database.

[2] Vida Adeli, Ehsan Adeli, Ian Reid, Juan Carlos Niebles, and Hamid Rezatofighi. Socially and contextually aware human motion and pose forecasting. *Robotics and Automation Letters*, 2020.

[3] Vida Adeli, Mahsa Ehsanpour, Ian Reid, Juan Carlos Niebles, Silvio Savarese, Ehsan Adeli, and Hamid Rezatofighi. Tripod: Human trajectory and pose dynamics forecasting in the wild. In *International Conference on Computer Vision*, 2021.

[4] Emre Aksan, Manuel Kaufmann, Peng Cao, and Otmar Hilliges. A spatio-temporal transformer for 3d human motion prediction. In *International Conference on 3D Vision*, 2021.

[5] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *Conference on Computer Vision and Pattern Recognition*, 2016.

[6] Sadegh Aliakbarian, Fatemeh Sadat Saleh, Mathieu Salzmann, Lars Petersson, and Stephen Gould. A stochastic conditioning scheme for diverse human motion prediction. In *Conference on Computer Vision and Pattern Recognition*, 2020.

[7] Eduardo Alvarado, Damien Rohmer, and Marie-Paule Cani. Generating upper-body motion for real-time characters making their way through dynamic environments. *Eurographics Symposium on Computer Animation*, 2022.

[8] Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. *Advances in Neural Information Processing Systems*, 2021.

[9] Timur Bagautdinov, Alexandre Alahi, François Fleuret, Pascal Fua, and Silvio Savarese. Social scene understanding: End-to-end multi-person action localization and collective activity recognition. In *Conference on Computer Vision and Pattern Recognition*, 2017.

[10] Dmitry Baranchuk, Ivan Rubachev, Andrey Voynov, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. *arXiv preprint arXiv:2112.03126*, 2021.

[11] Judith Bütepage, Michael J Black, Danica Kragic, and Hedvig Kjellstrom. Deep representation learning for human motion prediction and classification. In *Conference on Computer Vision and Pattern Recognition*, 2017.

[12] Yujun Cai, Lin Huang, Yiwei Wang, Tat-Jen Cham, Jianfei Cai, Junsong Yuan, Jun Liu, Xu Yang, Yiheng Zhu, Xiaohui Shen, et al. Learning progressive joint propagation for human motion prediction. In *European Conference on Computer Vision*, 2020.

[13] Yujun Cai, Yiwei Wang, Yiheng Zhu, Tat-Jen Cham, Jianfei Cai, Junsong Yuan, Jun Liu, Chuanxia Zheng, Sijie Yan, Henghui Ding, et al. A unified 3d human motion synthesis model via conditional variational auto-encoder. In *International Conference on Computer Vision*, 2021.

[14] Lujing Chen, Rui Liu, Xin Yang, Dongsheng Zhou, Qiang Zhang, and Xiaopeng Wei. Sttg-net: a spatio-temporal network for human motion prediction based on transformer and graph convolution network. *Visual Computing for Industry, Biomedicine, and Art*, 2022.

[15] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 2021.

[16] Christian Diller, Thomas Funkhouser, and Angela Dai. Forecasting characteristic 3d poses of human actions. *Conference on Computer Vision and Pattern Recognition*, 2022.

[17] Dominik Maria Endres and Johannes E Schindelin. A new metric for probability distributions. *Transactions on Information theory*, 2003.

[18] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent network models for human dynamics. In *International Conference on Computer Vision*, 2015.

[19] Bent Fuglede and Flemming Topsoe. Jensen-Shannon divergence and Hilbert space embedding. In *International Symposium on Information Theory*, 2004.

[20] Jiyang Gao, Chen Sun, Hang Zhao, Yi Shen, Dragomir Anguelov, Congcong Li, and Cordelia Schmid. Vectornet: Encoding hd maps and agent dynamics from vectorized representation. In *Conference on Computer Vision and Pattern Recognition*, 2020.

[21] Anand Gopalakrishnan, Ankur Mali, Dan Kifer, Lee Giles, and Alexander G Ororbia. A neural temporal model for human motion prediction. In *Conference on Computer Vision and Pattern Recognition*, 2019.

[22] Liang-Yan Gui, Yu-Xiong Wang, Xiaodan Liang, and José MF Moura. Adversarial geometry-aware human motion prediction. In *European Conference on Computer Vision*, 2018.

[23] Wen Guo, Xiaoyu Bie, Xavier Alameda-Pineda, and Francesc Moreno-Noguer. Multi-person extreme motion prediction. In *Conference on Computer Vision and Pattern Recognition*, 2022.

[24] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 2020.

[25] Daniel Holden, Jun Saito, Taku Komura, and Thomas Joyce. Learning motion manifolds with convolutional autoencoders. In *SIGGRAPH Asia technical briefs*. 2015.

[26] Emiel Hoogeboom, Didrik Nielsen, Priyank Jaini, Patrick Forré, and Max Welling. Argmax flows and multinomial diffusion: Learning categorical distributions. *Advances in Neural Information Processing Systems*, 2021.

[27] Ashesh Jain, Amir R Zamir, Silvio Savarese, and Ashutosh Saxena. Structural-rnn: Deep learning on spatio-temporal graphs. In *Conference on Computer Vision and Pattern Recognition*, 2016.

[28] Michael Janner, Yilun Du, Joshua B Tenenbaum, and Sergey Levine. Planning with diffusion for flexible behavior synthesis. *International Conference on Learning Representations*, 2022.

[29] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic Studio: A Massively Multiview System for Social Motion Capture. In *International Conference on Computer Vision*, 2015.

[30] Hanbyul Joo, Tomas Simon, Mina Cikara, and Yaser Sheikh. Towards social artificial intelligence: Nonverbal social signal prediction in a triadic interaction. In *Conference on Computer Vision and Pattern Recognition*, 2019.

[31] Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Scott Godisart, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social interaction capture. *Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[32] Julia Kantorovitch, Janne Väre, Vesa Pehkonen, Arto Laikari, and Heikki Seppälä. An assistive household robot–doing more than just cleaning. *Journal of Assistive Technologies*, 2014.

[33] Anne Keitel, Moritz M Daum, et al. The use of intonation for turn anticipation in observed conversations without visual signals as source of information. *Frontiers in psychology*, 2015.

[34] Łukasz Kidziński, Bryan Yang, Jennifer L Hicks, Apoorva Rajagopal, Scott L Delp, and Michael H Schwartz. Deep neural networks enable quantitative movement analysis using single-camera videos. *Nature communications*, 2020.

[35] Vineet Kosaraju, Amir Sadeghian, Roberto Martín-Martín, Ian Reid, Hamid Rezatofighi, and Silvio Savarese. Social-bigat: Multimodal trajectory forecasting using bicycle-gan and graph attention networks. *Advances in Neural Information Processing Systems*, 2019.

[36] Stephen C Levinson and Francisco Torreira. Timing in turn-taking and its implications for processing models of language. *Frontiers in psychology*, 2015.

[37] Chen Li, Zhen Zhang, Wee Sun Lee, and Gim Hee Lee. Convolutional sequence to sequence model for human dynamics. In *Conference on Computer Vision and Pattern Recognition*, 2018.

[38] Thomas Lucas*, Fabien Baradel*, Philippe Weinzaepfel, and Grégory Rogez. Posegpt: Quantization-based 3d human motion generation and forecasting. In *European Conference on Computer Vision*, 2022.

[39] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Conference on Computer Vision and Pattern Recognition*, 2022.

[40] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. History repeats itself: Human motion prediction via motion attention. In *European Conference on Computer Vision*, 2020.

[41] Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. Learning trajectory dependencies for human motion prediction. *International Conference on Compuver Vision*, 2019.

[42] Julieta Martinez, Michael J Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *Conference on Computer Vision and Pattern Recognition*, 2017.

[43] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3d pose estimation from monocular rgb. In *International Conference on 3D Vision*, 2018.

[44] Ronja Möller, Antonino Furnari, Sebastiano Battiato, Aki Härmä, and Giovanni Maria Farinella. A survey on human-aware robot navigation. *Robotics and Autonomous Systems*, 2021.

[45] Brendan Tran Morris and Mohan Manubhai Trivedi. A survey of vision-based trajectory learning and analysis for surveillance. *Transactions on Circuits and Systems for Video Technology*, 2008.

[46] Evonne Ng, Hanbyul Joo, Liwen Hu, Hao Li, Trevor Darrell, Angjoo Kanazawa, and Shiry Ginosar. Learning to listen: Modeling non-deterministic dyadic facial motion. In *Conference on Computer Vision and Pattern Recognition*, 2022.

[47] Sangmin Oh, Anthony Hoogs, Amitha Perera, Naresh Cuntoor, Chia-Chih Chen, Jong Taek Lee, Saurajit Mukherjee, JK Aggarwal, Hyungtae Lee, Larry Davis, et al. A large-scale benchmark dataset for event recognition in surveillance video. In *Conference on Computer Vision and Pattern Recognition*, 2011.

[48] Dario Pavllo, Christoph Feichtenhofer, Michael Auli, and David Grangier. Modeling human motion with quaternion-based neural networks. *International Journal of Computer Vision*, 2020.

[49] Dario Pavllo, David Grangier, and Michael Auli. Quaternet: A quaternion-based recurrent model for human motion. In *British Machine Vision Conference*, 2018.

[50] Mathis Petrovich, Michael J Black, and Gül Varol. Temos: Generating diverse human motions from textual descriptions. *European Conference on Computer Vision*, 2022.

[51] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

[52] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Conference on Computer Vision and Pattern Recognition*, 2022.

[53] Alejandro Hernandez Ruiz, Juergen Gall, and Francesc Moreno-Noguer. Human motion prediction via spatio-temporal inpainting. *International Conference on Computer Vision*, 2019.

[54] Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. A simplest systematics for the organization of turn taking for conversation. In *Studies in the organization of conversational interaction*. 1978.

[55] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *Transactions on Graphics*, 2022.

[56] Keren Shavit-Cohen and Elana Zion Golumbic. The dynamics of attention shifts among concurrent speech in a naturalistic multi-speaker virtual environment. *Frontiers in Human Neuroscience*, 2019.

[57] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, 2015.

[58] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 2015.

[59] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[60] Chen Sun, Abhinav Shrivastava, Carl Vondrick, Rahul Sukthankar, Kevin Murphy, and Cordelia Schmid. Relational action forecasting. In *Conference on Computer Vision and Pattern Recognition*, 2019.

[61] Julian Tanke, Chintan Zaveri, and Juergen Gall. Intention-based long-term human motion anticipation. In *International Conference on 3D Vision*, 2021.

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

[63] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 2017.

[64] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 2017.

[65] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *International Conference on Learning Representations*, 2017.

[66] Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *European Conference on Computer Vision*, 2018.

[67] 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 *European conference on computer vision*, 2018.

[68] Jingbo Wang, Yu Rong, Jingyuan Liu, Sijie Yan, Dahua Lin, and Bo Dai. Towards diverse and natural scene-aware 3d human motion synthesis. In *Conference on Computer Vision and Pattern Recognition*, 2022.

[69] Jiashun Wang, Huazhe Xu, Medhini Narasimhan, and Xiaolong Wang. Multi-person 3d motion prediction with multi-range transformers. *Advances in Neural Information Processing Systems*, 2021.

[70] Jiashun Wang, Huazhe Xu, Jingwei Xu, Sifei Liu, and Xiaolong Wang. Synthesizing long-term 3d human motion and interaction in 3d scenes. In *Conference on Computer Vision and Pattern Recognition*, 2021.

[71] Ye Yuan and Kris Kitani. Dlow: Diversifying latent flows for diverse human motion prediction. In *European Conference on Computer Vision*, 2020.

[72] Jason Y Zhang, Panna Felsen, Angjoo Kanazawa, and Jitendra Malik. Predicting 3d human dynamics from video. In *International Conference on Computer Vision*, 2019.

[73] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*, 2022.

[74] Yan Zhang, Mohamed Hassan, Heiko Neumann, Michael J Black, and Siyu Tang. Generating 3d people in scenes without people. In *Conference on computer vision and pattern recognition*, 2020.

[75] Yan Zhang and Siyu Tang. The wanderings of odysseus in 3d scenes. In *Conference on Computer Vision and Pattern Recognition*, 2022.