STOCHASTIC MULTI-PERSON 3D MOTION FORECASTING

Sirui Xu  Yu-Xiong Wang*  Liang-Yan Gui*
University of Illinois at Urbana-Champaign
{siruixu2, yxw, lgui}@illinois.edu
https://sirui-xu.github.io/DuMMF

ABSTRACT

This paper aims to deal with the ignored real-world complexities in prior work on human motion forecasting, emphasizing the social properties of multi-person motion, the diversity of motion and social interactions, and the complexity of articulated motion. To this end, we introduce a novel task of stochastic multi-person 3D motion forecasting. We propose a dual-level generative modeling framework that separately models independent individual motion at the local level and social interactions at the global level. Notably, this dual-level modeling mechanism can be achieved within a shared generative model, through introducing learnable latent codes that represent intents of future motion and switching the codes’ modes of operation at different levels. Our framework is general; we instantiate it with different generative models, including generative adversarial networks and diffusion models, and various multi-person forecasting models. Extensive experiments on CMU-Mocap, MuPoTS-3D, and SoMoF benchmarks show that our approach produces diverse and accurate multi-person predictions, significantly outperforming the state of the art.

1 INTRODUCTION

One of the hallmarks of human intelligence is the ability to predict the evolution of the physical world over time given historical information. For example, humans naturally anticipate the flow of people in public areas, react, and plan their own behavior based on social rules, such as avoiding collisions. Effective forecasting of human motion has thus become a crucial task in computer vision and robotics, e.g., in autonomous driving (Paden et al., 2016) and robot navigation (Rudenko et al., 2018). This task, however, is challenging. First, human motion is structured with respect to both body physics and social norms, and is highly dependent on the surrounding environment and its changes. Second, human motion is inherently uncertain and multi-modal, especially over long time horizons.

Previous work on human motion forecasting often focuses on simplified scenarios. Perhaps the most widely adopted setting is on stochastic local motion prediction of a single person (Mao et al., 2021; Yuan & Kitani, 2020), which ignores human interactions with the environment and other people in the environment. Another related task is deterministic multi-person motion forecasting (Wang et al., 2021b; Adeli et al., 2020; 2021; Guo et al., 2022). However, it does not take into account the diversity of individual movements and social interactions. In addition, stochastic forecasting of human trajectories in crowds (Alahi et al., 2014) has shown progress in modeling social interactions, e.g., with the use of attention models (Kosaraju et al., 2019; Vemula et al., 2018; Zhang et al., 2019) and spatial-temporal graph models (Huang et al., 2019; Ivanovic & Pavone, 2019; Salzmann et al., 2020; Yu et al., 2020). Nevertheless, this task only considers motion and interactions at the trajectory level. Modeling articulated 3D poses involves richer human-like social interactions than trajectory forecasting which only needs to account for trajectory collisions.

To overcome these limitations, we introduce a novel task of stochastic multi-person 3D motion forecasting, aiming to jointly tackle the aforementioned aspects ignored in the previous work – the social properties of multi-person motion, the multi-modality of motion and social interactions, and

*Yu-Xiong Wang and Liang-Yan Gui contributed equally to this work.
Figure 1: Illustration of the multifaceted challenges in the proposed task of stochastic multi-person 3D motion forecasting. (a) Single-person fidelity: for each person, the predicted pose and trajectory should be realistic and consistent with each other, e.g., to avoid foot floating and skating. (b) Multi-person fidelity: multi-person motion in a scene inherently involves social interactions, e.g., to avoid motion collisions. (c) Overall diversity: long-term human motion is uncertain and stochastic; we address this intrinsic multi-modality, while existing work (Wang et al., 2021b; Adeli et al., 2020; 2021; Guo et al., 2022) simplifies to deterministic prediction.

the complexity of articulated motion. As shown in Figure 1, this task requires the predicted multiple motion sequences of multi-person to satisfy the following conditions – (a) single-person fidelity: for example, all single-person motion should be continuous, and the articulated properties should be preserved under the physical rules; (b) multi-person fidelity: the predicted motion should be socially-aware, with the consideration of the interactions between predictions from different people; and (c) overall diversity: the movement of the human body should be as varied as possible, but within the constraints of conditions (a) and (b).

Due to the substantially increased complexity of our task, it becomes challenging to optimize all three objectives simultaneously. We observe simply extending existing work such as on deterministic motion forecasting cannot address the proposed task. This difficulty motivates us to adopt a divide-and-conquer strategy, together with the observation that single-person fidelity and multi-person fidelity can be viewed as relatively independent goals, while there is an inherent trade-off between fidelity and diversity. Therefore, we propose a Dual-level generative modeling framework for Multi-person Motion Forecasting (DuMMF). At the local level, we model motion for different people independently under relaxed conditions, thus satisfying single-person fidelity and diversity. Meanwhile, at the global level, we model social interactions by considering the correlation between all motion, thereby further improving multi-person fidelity. Notably, this dual-level modeling mechanism can be achieved within a shared generative model, through simply switching the modes of operation of the motion intent codes (i.e., latent codes of the generative model) at different levels. By optimizing these codes with level-specific objectives, we produce diverse and realistic multi-person predictions.

Our contributions can be summarized as follows. (a) To the best of our knowledge, we are the first to investigate the task of stochastic multi-person 3D motion forecasting. (b) We propose a simple yet effective dual-level learning framework to address this task. (c) We introduce discrete learnable social intents at dual levels to improve the realism and diversity of predictions. (d) Our framework is general and can be operationalized with various generative models, including generative adversarial networks and diffusion models, and different types of multi-person motion forecasting models. Notably, it can be generalized to challenging more-person (e.g., 18-person) scenarios that are unseen during training.

2 Related Work

Stochastic Human Motion Forecasting. There have been many advances in stochastic human motion forecasting, many of which (Walker et al., 2017; Yan et al., 2018; Barsoum et al., 2018) are based on the adaptation and improvement of deep generative models such as variational autoencoders
(VAEs) (Kingma & Welling, 2013), generative adversarial networks (GANs) (Goodfellow et al., 2014), normalizing flows (NFs) (Rezende & Mohamed, 2015), and diffusion models (Sohl-Dickstein et al., 2015; Song et al., 2020; Ho et al., 2020). Some recent approaches (Bhattacharyya et al., 2018; Dilokthanakul et al., 2016; Gurumurthy et al., 2017; Yuan & Kitani, 2019; 2020; Zhang et al., 2021; Mao et al., 2021; Xu et al., 2022b; Petrovich et al., 2022) emphasize on the promotion of diversity. Mao et al. (2021) sequentially generate the different parts of a pose for better controllability of diversity. Xu et al. (2022b) introduce learnable anchors in the latent space to guide the samples with sufficient diversity. Although these methods can predict very diverse human motion sequences, most of them are limited to local motion and ignore the global trajectory. Some of their produced motion sequences are actually unrealistic; in particular, incorporating global trajectories may result in severe foot skating. Predicting human motion under the constraint of scene context (Cao et al., 2020; Hassan et al., 2021; Zhang & Tang, 2022) has recently been explored, where the effect of global trajectories and scenes on human motion is considered. Instead of predicting single-person movement, our work focuses on diverse multi-person movements and social interactions.

Multi-Person Forecasting. So far, research on multi-person forecasting has mainly focused on global trajectory forecasting (Helbing & Molnar, 1995; Mehran et al., 2009; Pellegrini et al., 2009; Yamaguchi et al., 2011; Zhou et al., 2012; Alahi et al., 2014; 2016; Lee et al., 2017; Gupta et al., 2018; Sadeghian et al., 2019; Amirian et al., 2019; Kosaraju et al., 2019; Sun et al., 2020; Mangalam et al., 2020; Sun et al., 2021; Kothari et al., 2021; Tsao et al., 2022). Gupta et al. (2018) utilize a winner-takes-all loss (Rupprecht et al., 2017) with a socially-aware GAN. We also adopt these two loss functions in our dual-level modeling. Kothari et al. (2021) introduce discrete social anchors and revise the distribution of trajectories by the predefined anchors. Unlike their hand-crafted anchors, our discrete social intent codes are learnable components obtained through the dual-level optimization. Some recent attempts address deterministic multi-person 3D motion forecasting. Adeli et al. (2020; 2021) use additional contextual information to aid multi-person forecasting and propose a Social Motion Forecasting (SoMoF) Benchmark. Guo et al. (2022) utilize cross-attention for two-person motion forecasting, while Wang et al. (2021b) introduce a multi-range transformer architecture that can be generalized to more persons. Please refer to Sec. B of the Appendix for additional discussion.

3 Methodology

In this section, we explain the proposed dual-level generative modeling framework (DuMMF) for our task. As illustrated in Figure 2, our key insight is to decouple the modeling of independent individual movements at the local level and social interactions at the global level (Sec. 3.1). Within a shared forecasting model, we achieve this by (a) introducing learnable latent codes that represent intents of future movement (Sec. 3.2), (b) switching the codes’ modes of operation at different levels (Sec. 3.1 and Sec. 3.2), and (c) training with level-specific objectives (Sec. 3.3). Our framework is general and can be operationalized with different types of motion forecasting models, and we summarize the multi-person motion predictors used in this paper (Sec. 3.4).

Problem Formulation. We denote the input motion sequence of length $T_p$ for $N$ persons in a scene as $\{X_n\}_{n=1}^N$, where $X_n[t]$ is the pose of $n$-th person at time step $t$. We aim to predict $M$ future motion sequences of length $T_f$, denoted as $\{\{Y_{mn}\}_{n=1}^N\}_{m=1}^M$, where $Y_{mn} = [Y_{mn}[T_{h+1}], \ldots, Y_{mn}[T_{h+T_f}]]$ is the $m$-th predicted motion of $n$-th person. We use the 3D coordinates to represent the absolute joint position of $V$ joints, hence $\forall n, m, t, X_n[t], Y_{mn}[t] \in \mathbb{R}^{V \times 3}$. We assume to be given the ground truth motion of $N$ persons as $\{Y_n\}_{n=1}^N$. Our goal is to forecast multiple realistic yet diverse future motion sequences, such that (a) all the $M$ predictions represent human-like motion, simultaneously satisfying single-person fidelity and multi-person fidelity; (b) the predictions are diverse (overall diversity); and (c) one of the predicted sequences is as close to the ground truth as possible.

3.1 Dual-Level Stochastic Multi-Person Motion Forecasting: Overview

Basic Generative Modeling Framework. Stochastic multi-person future motion can be modeled as a joint distribution using deep generative models (Goodfellow et al., 2014). Accordingly, we denote this joint distribution of the future motion of $N$ persons as $p(Y_1, Y_2, \ldots, Y_N | X_1, X_2, \ldots, X_N)$, where all future movements $\{Y_n\}_{n=1}^N$ are conditioned on the past sequences $\{X_n\}_{n=1}^N$ of all persons. Typically, we can use a latent code $z \sim p(z)$ to reparameterize this joint distribution as $p(\{Y_n\}_{n=1}^N | \{X_n\}_{n=1}^N) = \int p(\{Y_n\}_{n=1}^N | z, \{X_n\}_{n=1}^N)p(z)dz$. Here, the latent code $z$ can be inter-
We abstract the encoder of a multi-person predictor as the combination of a local branch that encodes single-person motion and a global branch that encodes multi-person motions (see Table A of the Appendix for instantiations).

Future social interactions requires all latent codes to be the same. The latent codes comprise both discrete intent codes, which are learned from the data and represented as a set, and continuous intent codes.

Figure 2: Overview of our proposed dual-level generative modeling with motion inputs of three persons as an illustration. (a) We combine encoded multi-person embeddings with independent intent codes at the level of modeling individual motion. (b) Differently, the global level of modeling going beyond the local individual level, the generative model in the Appendix for instantiations).

Given the complexity of our task, it becomes challenging to simultaneously ensure all objectives (i.e., the fidelity of a single person, the fidelity of multiple persons, and the overall diversity). To overcome this difficulty, we introduce a dual-level modeling mechanism that explicitly decomposes the task objectives into local modeling of independent individual movements and global modeling of social interactions. Notably, we achieve this by simply switching the modes of operation for the latent codes $z$ w.r.t. different levels of modeling, without any change to the model architecture $G$.

**Local-Level Modeling: Individual Motion.** At this level, the generative model $G_\theta$ models all human bodies as independent of each other, and we aim to improve the overall diversity and the single-person fidelity, alleviating problems such as predicting unrealistic poses. Here, the joint distribution of future human motions can be rewritten in the form of all single-person marginal distributions, i.e., $p(\{Y_n\}_{n=1}^N | \{X_n\}_{n=1}^N) = \prod_{n=1}^N p(Y_n | X_n)$. To this end, as shown in Figure 2(a), we consider leveraging $N$ different individual intents $z_1, z_2, \ldots, z_N$ independently drawn from $p(z)$ to generate independent future movements, denoted as

$$z_1, \ldots, z_N \sim p(z), \ \{\hat{Y}_n\}_{n=1}^N = G_\theta(\{z_n\}_{n=1}^N, \{X_n\}_{n=1}^N). \quad (2)$$

**Global-Level Modeling: Social Interactions.** Going beyond the local individual level, the generative model $G_\theta$ at the global level takes into account the social behavior of multiple people to model their joint distribution. The goal is to further improve the multi-person fidelity, e.g., promoting the overall accuracy. As illustrated in Figure 2(b), to maintain the network architecture $G$ unchanged, we still use $N$ individual intents as input. However, different from the local level, we constrain these $N$ individual intents to be the same, representing social intents that stand for correlations between the intents of multiple persons. Formally, we have

$$z \sim p(z), \ \{\hat{Y}_n\}_{n=1}^N = G_\theta(\{z\}_{n=1}^N, \{X_n\}_{n=1}^N). \quad (3)$$
Note that, without additional constraints, this dual-level modeling scheme by itself is not guaranteed to enforce the latent codes to behave in the designed manner. To this end, we introduce learnable latent intent codes $z$ (Sec. 3.2), jointly optimize the codes $z$ and the forecasting model $G$ guided by the level-specific training objectives (Sec. 3.3).

### 3.2 Discrete Learnable Human Intent Codes

Intuitively, an arbitrary, albeit identical, individual intent in Eq. 3 may not adequately lead to a valid social intent. We thus hypothesize that a social intent is formed when all individual intents are the same and belong within some range of “options.” This can typically be achieved through discrete choice models (Aguiirregabiria & Mira, 2010; Ryan & Gerard, 2003; Bhat et al., 2008; Leonardi, 1984) – an effective tool that predicts choices from a set of available options created by hand-crafted rules. Here, we formulate the correlation of multiple persons at the global level by using the same discrete code. However, the intent options for social interactions are more subtle and difficult to define manually than those in other applications such as for trajectories (Kothari et al., 2021). Therefore, we use a set of learnable codes $z^d \in \{Z^m\}_{m=1}^M$ to represent social intents, inspired by Xu et al. (2022b). However, we introduce different training strategies that are tailored to this new task (Sec. 3.3). Our motivations are: (a) Subject to physical constraints and social laws, the intents of future movements should share some deterministic patterns. For example, all intents should avoid imminent collisions anticipated from the history, even if these intents refer to different motions. We assume that such deterministic properties shared by social intents can be represented by a set of shareable codes learned directly from the data. (b) It will be easier for the predictor to identify and implement different levels of functionality by jointly optimizing discrete intents and the predictor. To further enhance the expressiveness of the codes, we retain the original continuous Gaussian noise $z^c \sim p(z)$ of the generative model $G_\theta$ and bundle the discrete intent with the noise to represent the final intent, as shown in Figure 2(c). Now the global-level modeling of social interactions in Eq. 3 is reformulated as

$$z_1^1, \ldots, z_n^1 \sim P(z), \ z^d \in \{Z^m\}_{m=1}^M, \ \{\tilde{Y}_n\}_{n=1}^N = G_\theta(z_1^1 + z^d_{m=1}^{n=1}, \{X_n\}_{n=1}^N).$$

And correspondingly, the local-level modeling of individual motion in Eq. 2 becomes

$$z_1^1, \ldots, z_n^1 \sim P(z), \ z^d_1, \ldots, z^d_n \in \{Z^m\}_{m=1}^M, \ \{\tilde{Y}_n\}_{n=1}^N = G_\theta(z_1^1 + z^d_{m=1}^{n=1}, \{X_n\}_{n=1}^N).$$

### 3.3 Training Under the Guidance of Level-Specific Objectives

Using only the same discrete intent does not naturally and necessarily inherit the multi-person correlation. Thus, we optimize both the parameters of the predictor and the discrete codes with a level-specific training strategy. We jointly train both levels of individual movement modeling and social interaction modeling, but with each level guided by its own objective. In each forward pass, we explicitly produce different output predictions from different intents of the two levels. And then, in the backward pass, the discrete intent codes $z^d$ are optimized separately at different levels, while the parameters $\theta$ of the forecasting model $G$ are updated based on the fused losses from the two levels.

**Local-Level Training.** At the local level, we train the model without social interactions. Given independent multi-person motion data $\{\{X_n\}_{n=1}^N, \{Y_n\}_{n=1}^N\}$, we first randomly sample the discrete intent codes and merge them with independently sampled continuous intent codes into $M \times N$ different latent codes $\{z^m_n\}_{n=1}^N$. We then use each intent and the past motion of each person to predict $M$ future motion sequences $\{\tilde{Y}_n\}_{n=1}^N$ for each person $\{X_n\}_{n=1}^N$. Local-level objectives consider single-person fidelity and overall diversity respectively.

**Global-Level Training.** Meanwhile, training is also conducted at the global level to enable the modeling of social interactions. The difference from the local setting is that we incorporate the discrete and continuous codes into only $M$ distinct latent codes $\{z^m_n\}_{n=1}^N$: hence, the discrete latent codes $z^m_n$ of all $N$ individuals are the same for a certain $m$-th prediction. Here, we introduce learning objectives to facilitate multi-person fidelity and accuracy.

Please refer to Sec. C of the Appendix for more detail on the aforementioned level-specific learning objectives. Also, Sec. G of the Appendix further demonstrates that these learning objectives are important, and in some cases critical, to the fidelity and diversity of multi-person motion.

**Inference.** During inference, we only use the global-level strategy by sampling the same intent for all individuals present in the scene. Note that the uncertainty of human motion substantially
Figure 3: Qualitative results of DuMMF with a DDPM. We demonstrate the generalizability of our method to handle a significantly more complex scenario with 18 persons. Note that our model is trained only on 3-person data. We visualize the predicted final poses at 2 seconds.

Table 1: Quantitative results of DuMMF with a DDPM. Both the baseline and our models are trained using SMPL-X representations on AMASS, and we convert them to skeletons for evaluation. Using the same backbone and generative model, our DuMMF framework significantly provides more accurate predictions with more intents.

<table>
<thead>
<tr>
<th>Method, # of Intents</th>
<th>@25 frames</th>
<th>@50 frames</th>
<th>@75 frames</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ADE ↓</td>
<td>FDE ↓</td>
<td>ADE ↓</td>
</tr>
<tr>
<td>DDPM (Ho et al., 2020), N/A</td>
<td>2.021</td>
<td>3.766</td>
<td>5.434</td>
</tr>
<tr>
<td>DuMMF (Ours), 1</td>
<td>2.405</td>
<td>4.814</td>
<td>3.992</td>
</tr>
<tr>
<td>+ DuMMF (Ours), 3</td>
<td>1.456</td>
<td>3.316</td>
<td>2.700</td>
</tr>
</tbody>
</table>

3.4 Network Architecture

Our dual-level modeling framework in conjunction with the discrete learnable intent codes is general and, in principle, does not rely on specific network architectures or generative models. To demonstrate this, we combine our framework with various types of deterministic multi-person motion predictors and different generative models, yielding consistent and significant improvements across all baselines (see Sec. 4). As shown in Figure 2, we abstract the encoder of the multi-person motion predictor into two parts: the local part is responsible for encoding single-person motion, while the global part is responsible for encoding multi-person motion and its interactions. A summary of the multi-person predictors used in the paper is given in Table A of the Appendix.

4 Experiments

Datasets. In the main paper, we show the evaluation on two motion capture datasets, CMU-Mocap (CMU) and MuPoTS-3D (Mehta et al., 2018). CMU-Mocap consists of movement sequences with up to two subjects for each scene. It contains 2,235 recordings performed by 144 different subjects, eight of which include double-person motion. We directly adopt these two-person motions for comparisons in two-person scenarios. For skeletal representation, we follow Wang et al. (2021b) for the train/test split and the preprocess to mix single-person and double-person motion together to synthesize a 3-person motion in each scene. Moreover, we construct an even more challenging scenario where we mix up to more people per scene. Please refer to Sec. H of the Appendix for the detail of this generalized setting. For meshes generated from SMPL-X representations (Pavlakos et al., 2019), we extract the multi-person data in CMU-Mocap from AMASS (Mahmood et al., 2019).
Figure 4: Qualitative results of DuMMF with a CGAN model. The leftmost column is the ground truth starting poses, and the second column is the ground truth poses three seconds later. We show our five sampled 3-second predictions in the middle. Our model produces diverse predictions, with one closer to the ground truth (highlighted by the red box) compared with MRT (rightmost column).

Table 2: Quantitative comparison between our DuMMF and deterministic forecasting baselines and their CGAN stochastic variants, using skeletal representations on CMU-Mocap. The number of intents is set to 5 for stochastic forecasting in 3-person (top) and 2-person (bottom) scenarios. DuMMF significantly improves multi-person accuracy and diversity across various architectures and deterministic predictors. Additionally, our discrete and continuous intent codes are complementary to each other in most cases.

<table>
<thead>
<tr>
<th>Architecture</th>
<th>Predictor</th>
<th>Description</th>
<th>Variants</th>
<th>MRT (Top)</th>
<th>Deterministic</th>
<th>CGAN</th>
<th>CGAN+DuMMF (Ours)</th>
<th>CGAN+DuMMF (Ours)</th>
<th>CGAN+DuMMF (Ours)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transformer</td>
<td>MRT (Wang et al., 2021a)</td>
<td>w/o Separation</td>
<td>w/o Continuous</td>
<td>Full</td>
<td>N/A</td>
<td>0.057</td>
<td>1.176</td>
<td>0.894</td>
<td>1.180</td>
</tr>
<tr>
<td>Transformer</td>
<td>XIA (Guo et al., 2022)</td>
<td>w/o Separation</td>
<td>w/o Continuous</td>
<td>Full</td>
<td>N/A</td>
<td>0.041</td>
<td>1.098</td>
<td>1.183</td>
<td>1.183</td>
</tr>
</tbody>
</table>

and follow the same strategy to mix single-person and double-person motion together. MuPoTS-3D consists of more than 8,000 frames with up to three subjects. We convert the data to the same 15-joint human skeleton and length units as CMU-Mocap, and evaluate the generalization on MuPoTS-3D of a model trained only on CMU-Mocap. We also report our performance on the SoMoF benchmark (Adeli et al., 2020; 2021) in Sec. G of the Appendix.

Metrics. For evaluating multi-person motion accuracy and diversity, we adopt the common metrics used in stochastic forecasting (Mao et al., 2021; Yuan & Kitani, 2020; 2019; Salzmann et al., 2020; Kothari et al., 2021) as follows. For accuracy measurement, we follow the Best-of-N (BoN) evaluation and use (a) Average Displacement Error (ADE): the average $\ell_2$ distance over time between the ground truth and the prediction closest to the ground truth; (b) Final Displacement Error (FDE): the $\ell_2$ distance between the final pose of the ground truth and the last predicted pose closest to the ground truth. For diversity measurement, we employ (c) Final Pairwise Distance (FPD): the average $\ell_2$ distance between all predicted final pose pairs. We disentangle the local pose and the global trajectory of the motion and measure their accuracy and diversity separately by defining the following metrics: rootADE, rootFDE, poseADE, poseFDE, rootFPD, and poseFPD. To measure three different aspects including single-person fidelity, multi-person fidelity, and overall diversity comprehensively, we provide a summary and analysis of all metrics for this novel task in Table F and Sec. F of the Appendix. We include results on all above metrics in Sec. G of the Appendix.
Figure 5: Qualitative results of DuMMF with a DDPM model. We visualize the predicted frames for two different 3-person input motions, which are listed on the left and right respectively. For each input, we generate four sampled motions, arranged in a single column and listed sequentially in time at 0, 1, 2, and 3 seconds. Our method effectively produces diverse human-like social interactions.

**Baselines.** Given that we propose a new task, there is no readily available baseline from existing work. For comparison purposes, we customize several baselines from deterministic multi-person motion forecasting which we summarize in Table A, and extend them to stochastic forecasting. (a) **RNN+SC-MPF** (Adeli et al., 2020): We use Social Pooling (Alahi et al., 2016) to integrate the multi-person information. We follow the implementation in Adeli et al. (2020) and reproduce the results. For the Social Pooling module, we select the maximum pooling for its best performance as ablated in Adeli et al. (2020). (b) **Transformer+SC-MPF:** For a fair comparison, we introduce a transformer-based encoder architecture to combine with the Social Pooling module. (c) **XIA** (Guo et al., 2022): We follow their implementation and use Cross Interaction Attention (XIA) to encode 2-person information. (d) **MRT** (Wang et al., 2021b): We use multi-range transformer (MRT) as our main architecture for its state-of-the-art performance. We further extend MRT to 2-person scenarios to compare with XIA. For stochastic forecasting, we formulate a conditional generative adversarial network (CGAN) (Mirza & Osindero, 2014) and a denoising diffusion probabilistic model (DDPM) (Ho et al., 2020), with or without dual-level modeling and discrete intents.

**Implementation Details.** For the skeletal representation, following Wang et al. (2021b), we train the model to predict a 15-frame sequence of 3 people given the ground truth 15, 30, and 45 past frames at 15Hz, as the encoder (RNN/Transformer) accepts different input lengths. We use a 6-layer transformer (or RNNs), where we set the feature dimension to 128. For evaluation, we recursively predict the next 15 frames 3 times given all past frames generated, as illustrated in Sec. 3.3. Thus, given the number of intents to be $M$, the model outputs $M$, $M^2$, and $M^3$ different predictions in sequence. For the SMPL representation, we train the model to predict a 25-frame sequence of 3 people given the 10 past frames at 30Hz. We use an 8-layer transformer, where we set the feature dimension to 512. For evaluation, similarly, we use 10 frames as the past motion and recursively predict the next 25 frames. Additional implementation details are provided in Sec. E of the Appendix.

**Quantitative Results.** We compare our method with a pure DDPM in Table 1. Our improvement is significant. Notably, even in the case of a single intent, where we only evaluate one prediction, our method outperforms DDPM in long-term generation. As the number of intents increases, our method provides more accurate results, especially in long-term prediction. In Table 2, we demonstrate that our DuMMF framework benefits all predictor variants. We observe that the simple combination of deterministic predictors and CGAN results in very low diversity and accuracy. By contrast, our full approach significantly outperforms CGAN baselines on both diversity and accuracy across all backbones. In Table D of the Appendix, we further show that DuMMF achieves the best generalization results across all predictors on MuPoTS-3D, highlighting its generality and superiority.

**Ablation: Effectiveness of Dual-Level Modeling.** Tables 2 and 3 show the effectiveness of our dual-level framework. First, we investigate the settings with only single-person motion modeling or only social interaction modeling. In Table 3, compared with our full method, modeling independent multi-person motion (‘w/o Social’) provides higher diversity but leads to inaccurate poses, since the social restrictions are not considered. With only social interaction modeling, the model (‘w/o Individual’) cannot output sufficiently diverse predictions, which also makes predictions inaccurate,
Table 3: Ablation study of our DuMMF with a CGAN and MRT (Wang et al., 2021b) using skeletal representations on CMU-Mocap and MuPoTS-3D. We report both accuracy and diversity for root and pose separately. The results show the effectiveness of our dual-level modeling along with discrete motion intents, and complementariness of local-level and global-level modeling.

<table>
<thead>
<tr>
<th>Method</th>
<th>@ t = 3s</th>
<th>CMU-Mocap (CMU)</th>
<th>Generalized to MuPoTS-3D (Mehta et al., 2018)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>ROOT</td>
<td>POSE</td>
</tr>
<tr>
<td></td>
<td></td>
<td>FPD ↑ ADE ↓ FDE ↓</td>
<td>FPD ↑ ADE ↓ FDE ↓</td>
</tr>
<tr>
<td>MRT (Wang et al., 2021b)</td>
<td>N/A</td>
<td>0.753 0.946</td>
<td>N/A</td>
</tr>
<tr>
<td>Ours, Deterministic</td>
<td>N/A</td>
<td>0.738 0.916</td>
<td>N/A</td>
</tr>
<tr>
<td>w/o Individual</td>
<td>0.178</td>
<td>0.733 0.896</td>
<td>0.339</td>
</tr>
<tr>
<td>w/o Separation</td>
<td>1.664</td>
<td>0.751 0.909</td>
<td>0.977</td>
</tr>
<tr>
<td>w/o Social</td>
<td>0.167</td>
<td>0.747 0.911</td>
<td>0.293</td>
</tr>
<tr>
<td>Full</td>
<td>0.249</td>
<td>0.734 0.898</td>
<td>0.562</td>
</tr>
<tr>
<td>w/o Discrete</td>
<td>0.114</td>
<td>0.761 0.949</td>
<td>0.196</td>
</tr>
<tr>
<td>w/o Continuous</td>
<td>0.298</td>
<td>0.746 0.965</td>
<td>0.384</td>
</tr>
<tr>
<td>Full</td>
<td>0.249</td>
<td>0.731 0.898</td>
<td>0.562</td>
</tr>
</tbody>
</table>

as more varied outputs have a better chance of covering the ground truth. Note that the results of ‘CGAN’ and ‘w/o Separation’ in Table 2 and ‘w/o Separation’ in Table 3 are worse since they simply use all the learning objectives together without disentangling the two levels of modeling, while ‘w/o Separation’ is slightly better due to the use of discrete intents. Under our dual-level framework with level-specific motion intents and learning objectives, the model can more effectively incorporate the benefits of both levels, thus leading to improved accuracy and diversity. In Sec. G of the Appendix, we further demonstrate that our dual-level benefits different predictor variants.

Ablation: Effectiveness of Discrete Human Intents. In Tables 2 and 3, we also demonstrate that discrete human intents are effective and crucial. We observe the best results when using both discrete and continuous intents, indicating that they are complementary. In the absence of discrete intents (‘w/o Discrete’), the performance is only comparable with the baseline (‘CGAN’). Importantly, with the help of discrete intents, the improvement of dual-level modeling (‘Full’) over ‘w/o Separation’ is more pronounced, compared with the improvement of ‘w/o Discrete’ over ‘CGAN’. Therefore, discrete learnable intents are essential for effectively integrating the advantages of both levels during training. The performance without continuous intents (‘w/o Continuous’) is slightly worse than the full method. Our hypothesis is that relying solely on discrete intents is limiting, because they only support a finite number of outputs. In Sec. G of the Appendix, we further investigate how the number of discrete intents impacts stochastic forecasting.

Qualitative Results. Consistent with the quantitative evaluation above, we observe that our method provides diverse multi-person motion, and produces predictions closer to the ground truth compared with the deterministic method MRT in Figure 4. In Figure 5, we qualitatively show that our generated results in meshes reflect real-world diversity of social interactions. Furthermore, we provide qualitative results for more-person scenarios in Figure 3, and others in Figure A and Figure B of the Appendix. Please refer to Sec. H of the Appendix for more detail on the more-person setting.

Limitation. Although our dual-level framework has proven effective in producing high-quality and diverse predictions, we have observed artifacts such as foot skating in some predicted motion sequences. This is because our model relies solely on loss functions to constrain motion, rather than explicitly modeling articulated motion. As this is a common issue in learning-based methods, we plan to exploit a physical simulator to further improve the plausibility of our predicted motion.

5 CONCLUSION

We formulate a novel task called stochastic multi-person 3D motion forecasting, which better reflects the real-world human motion complexities. To simultaneously achieve single-person fidelity, social realism, and overall diversity, we propose a dual-level generative modeling framework (DuMMF) with learnable latent intent codes. Compared with prior work on deterministic or single-person prediction, our model learns to generate diverse and realistic human motion and interactions. Notably, our framework is model-agnostic and generalizes to unseen more-person scenarios.
Ethics Statement. Our proposed technique is useful in many applications, such as self-driving to avoid crowds. The potential negative societal impacts include: (a) our approach can be used to synthesize highly realistic human motion, which might lead to the spread of false information; (b) our approach requires real behavioral information as input, which may raise privacy concerns and result in the disclosure of sensitive identity information. Nevertheless, our model operates on the processed human skeleton representation that contains minimal identifying information, unlike raw data. On the positive side, this can be seen as a privacy-enhancing feature.

Acknowledgement. This work was supported in part by NSF Grant 2106825, NIFA Award 2020-67021-32799, the Jump ARCHES endowment through the Health Care Engineering Systems Center, the National Center for Supercomputing Applications (NCSA) at the University of Illinois at Urbana-Champaign through the NCSA Fellows program, the IBM-Illinois Discovery Accelerator Institute, the Illinois-Inspir Partnership, and the Amazon Research Award. This work used NVIDIA GPUs at NCSA Delta through allocation CIS220014 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by NSF Grants #2138259, #2138286, #2138307, #2137603, and #2138296.

REFERENCES

CMU graphics lab motion capture database. http://mocap.cs.cmu.edu/.


German Barquero, Johnny Núñez, Sergio Escalera, Zhen Xu, Wei-Wei Tu, Isabelle Guyon, and Cristina Palmero. Didn’t see that coming: a survey on non-verbal social human behavior forecasting. In Understanding Social Behavior in Dyadic and Small Group Interactions. PMLR, 2022.


Tianpei Gu, Guangyi Chen, Junlong Li, Chunze Lin, Yongming Rao, Jie Zhou, and Jiwen Lu. Stochastic trajectory prediction via motion indeterminacy diffusion. In CVPR, 2022. 15

Wen Guo, Xiaoyu Bie, Xavier Alameda-Pineda, and Francesc Moreno-Noguer. Multi-person extreme motion prediction. In CVPR, 2022. 1, 2, 3, 7, 8, 14, 15, 17

Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social GAN: Socially acceptable trajectories with generative adversarial networks. In CVPR, 2018. 3


Mohamed Hassan, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, and Michael Black. Stochastic scene-aware motion prediction. In ICCV, 2021. 3

Dirk Helbing and Peter Molnar. Social force model for pedestrian dynamics. Physical review E, 1995. 3


Yingfan Huang, Huikun Bi, Zhaoxin Li, Tianlu Mao, and Zhaowei Wang. STGAT: Modeling spatial-temporal interactions for human trajectory prediction. In ICCV, 2019. 1


Parth Kothari, Brian Sifringer, and Alexandre Alahi. Interpretable social anchors for human trajectory forecasting in crowds. In CVPR, 2021. 3, 5, 7, 14


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. 3
Wei Mao, Miaomiao Liu, and Mathieu Salzmann. Generating smooth pose sequences for diverse human motion prediction. In CVPR, 2021. 1, 3, 6, 7, 15, 18


Brian Paden, Michal Cáp, Sze Zheng Yong, Dmitry S. Yershov, and Emilio Frazzoli. A survey of motion planning and control techniques for self-driving urban vehicles. IEEE Transactions on Intelligent Vehicles, 2016. 1


Mathis Petrovich, Michael J. Black, and Gül Varol. TEMOS: Generating diverse human motions from textual descriptions. In ECCV, 2022. 3


Armin Saadat, Nima Fathi, and Saeed Saadatnejad. Towards human pose prediction using the encoder-decoder LSTM. In ICCVW, 2021. 14


Tim Salzmann, Boris Ivanovic, Punarjay Chakravarty, and Marco Pavone. Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In ECCV, 2020. 1, 7, 14


Jianhua Sun, Qinhong Jiang, and Cewu Lu. Recursive social behavior graph for trajectory prediction. In CVPR, 2020. 3


Anirudh Vemula, Katharina Muelling, and Jean Oh. Social Attention: Modeling attention in human crowds. In *ICRA*, 2018. 1


Jiashun Wang, Huazhe Xu, Medhini Narasimhan, and Xiaolong Wang. Multi-person 3d motion prediction with multi-range transformers. In *NeurIPS*, 2021b. 1, 2, 3, 6, 7, 8, 9, 14, 15, 17, 18, 19, 20


Yi Xu, Lichen Wang, Yizhou Wang, and Yun Fu. Adaptive trajectory prediction via transferable gnn. In *CVPR*, 2022c. 15

Kota Yamaguchi, Alexander C Berg, Luis E Ortiz, and Tamara L Berg. Who are you with and where are you going? In *CVPR*, 2011. 3


Cunjun Yu, Xiao Ma, Jiawei Ren, Haiyu Zhao, and Shuai Yi. Spatio-temporal graph transformer networks for pedestrian trajectory prediction. In *ECCV*, 2020. 1


Yan Zhang, Michael J Black, and Siyu Tang. We are more than our joints: Predicting how 3d bodies move. In *CVPR*, 2021. 3, 18

In this Appendix, we include additional method details and experimental results that are not included in the main paper due to limited space as follows. 1) We provide a visualization video as additional qualitative results, and the details are explained in Sec. A. 2) We include a further discussion on related work in Sec. B. 3) We explain different generative models incorporated in our proposed dual-level modeling framework in Sec. D; 4) We provide additional details of the experimental implementation in Sec. E and the summary of evaluation metrics in Sec. F. 5) To elaborate the effectiveness of our method, we provide additional ablation experiments with qualitative and quantitative analysis in Sec. G, and we evaluate our approach in more challenging scenarios with a significantly increased number of people in Sec. H.

A Visualization Video

In addition to Figure 4 and Figure 5 in the main paper and more qualitative results in this Appendix (Figures A and B), we provide a video to demonstrate more comprehensive visualizations of multi-person 3D motion forecasting at https://sirui-xu.github.io/DuMMF/images/demo.mp4. In this video, we illustrate that our method DuMMF generates diverse multi-person motion and social interactions, as well as taking into account both single-person and multi-person fidelity. We also show that our model is scalable and provide effective predictions in more challenging scenarios with a significantly increased number of people and associated more complex interactions. We also highlight the impact and effectiveness of our dual-level modeling framework.

B Additional Discussion on Related Work

As we demonstrate in the main paper, our proposed stochastic multi-person motion forecasting needs to simultaneously take into account single-person pose fidelity, consistency of pose and trajectory, social interactions between poses, and overall diversity of motion, while prior work including stochastic multi-person trajectory forecasting (Alahi et al., 2014; Kothari et al., 2020; 2021) and deterministic multi-person motion forecasting (Adeli et al., 2020; 2021; Guo et al., 2022; Wang et al., 2021b) focuses on simplified scenarios. A survey (Barquero et al., 2022) summarizes the progress in this area. Our problem reveals real-world complexity with substantially increased, multi-faced challenges which have not been, and cannot be, jointly tackled by any of the prior literature. In particular, in our stochastic scenario, the fidelity and interactions need to be both satisfied and diversified for all the predictions, which is very challenging and cannot be addressed by a simple extension of existing work on stochastic multi-person trajectory forecasting and deterministic multi-person motion forecasting as shown in Sec. G.

For deterministic multi-person motion forecasting, in Sec. 2 of the main paper, we have highlighted our key difference with Adeli et al. (2020; 2021); Guo et al. (2022), and here we discuss in more detail. Adeli et al. (2020) introduce a social pooling module to integrate social information. Adeli et al. (2021) propose a graph attentional network to jointly model human-human and human-object interactions. Both methods utilize additional contextual information to aid deterministic prediction with their proposed Social Motion Forecasting (SoMoF) benchmark (Adeli et al., 2020; 2021). Guo et al. (2022) propose a cross-interaction attention mechanism to predict cross dependencies between two pose sequences, making it applicable only to 2-person scenarios.

There are many methods achieving strong performance with simple and non-transformer-based frameworks on the SoMoF leaderboard. Specifically, futuremotion_ICCV21 (Wang et al., 2021a) develops a simple but effective framework based on a combination of graph convolutional networks and multiple motion optimization techniques. DViTA (Parsaeifard et al., 2021) decouples the human pose into a trajectory and local pose and employs a VAE to learn a representation of the local pose dynamics. LSTMV_LAST (Saadat et al., 2021) proposes a sequence-to-sequence LSTM via keypoint-based representation.

Complementing Sec. 2 of the main paper, we here discuss in more depth the trajectory-level multi-person forecasting. Trajectron++ (Salzmann et al., 2020) constructs a spatio-temporal scene graph where nodes represent individuals and edges represent their interactions, and global and local modeling is embedded in this graph structure. EvolveGraph (Li et al., 2020) introduces an effective dynamic mechanism to evolve interaction graphs, which can flexibly model dynamically changing interactions. Tra2tra (Xu et al., 2021) introduces a global social spatio-temporal attention neural network.
Table A: Summary of methods for encoding and integrating multi-person motions and interactions. We follow to abstract the encoder into local and global part as Figure 2(d). As we use the same decoder for baselines (Sec. 4 of the main paper), we only discuss encoders of three predictors in this paper. Note that XIA (Guo et al., 2022) can only be applied to two-person scenarios.

<table>
<thead>
<tr>
<th>Method</th>
<th>Local Encoder</th>
<th>Global Encoder</th>
</tr>
</thead>
<tbody>
<tr>
<td>SC-MPF (Adeli et al., 2020)</td>
<td>GRU((X_1))</td>
<td>MaxPool({GRU((X_n))}_{n=1}^N)</td>
</tr>
<tr>
<td>MRT (Wang et al., 2021b)</td>
<td>MultiHead((X_1, X_1))</td>
<td>MultiHead({(X_n)}<em>{n=1}^N, {(X_n)}</em>{n=1}^N)</td>
</tr>
<tr>
<td>XIA (Guo et al., 2022)</td>
<td>N/A</td>
<td>{MultiHead((X_1, X_2), MultiHead((X_2, X_1, X_1)))}</td>
</tr>
</tbody>
</table>

network to encode both spatial interactions and temporal features. GroupNet (Xu et al., 2022a) employs a multi-scale hypergraph neural network that models group-based interactions and facilitates more comprehensive relational reasoning. T-GNN (Xu et al., 2022c) introduces a transferable graph neural network that allows not only trajectory prediction but also domain alignment of potential distribution differences. MID (Gu et al., 2022) employs a diffusion model to model the variation of indeterminacy for trajectory prediction.

C  ADDITIONAL DETAILS OF LEVEL-SPECIFIC OBJECTIVES

We use \(\Delta\) to represent the residual of the motion sequence. For example, \(\Delta \hat{Y}_i^f = [\hat{Y}_i^f[T_h + 1] - \hat{X}_i[T_h], \hat{Y}_i^f[T_h + 2] - \hat{Y}_i^f[T_h + 1], \ldots, \hat{Y}_i^f[T_h + T_p] - \hat{Y}_i^f[T_h + T_p - 1]]\).

**Local-Level Objectives.** We adopt the multiple output loss (Guzmán-rivera et al., 2012) and extend it to the local reconstruction loss of multiple people \(\mathcal{L}_{IR}\), which is used to optimize the most accurate prediction of each person while maintaining diversity. We highlight the structure of the human skeleton by introducing the limb loss (Mao et al., 2021) \(\mathcal{L}_{L}\). Specifically,

\[
\mathcal{L}_{IR} = \frac{1}{N} \sum_{n=1}^{N} \min_{m=1,\ldots,M} \|\Delta \hat{Y}_n^m - \Delta Y_n\|_2^2, \tag{6}
\]

\[
\mathcal{L}_{L} = \frac{1}{N \times M} \sum_{n=1}^{N} \sum_{m=1}^{M} \|\hat{L}_n^m - L_n\|_2^2, \tag{7}
\]

where the vector \(L_n\) represents the ground truth distance between all pairs of joints that are physically connected in the \(n\)-th human body and \(\hat{L}_n^m\) includes the limb length for all the \(T_p\) poses in \(\hat{Y}_n^m\).

We further develop a multimodal reconstruction loss \(\mathcal{L}_{mmR}\) to provide additional supervision for all outputs \(\{\{\Delta Y_n^m\}_{n=1}^N\}_{m=1}^M\). We first construct pseudo future motion \(\{\hat{Y}_n^P\}_{p=1}^P\) for each historical sequence \(X_i\). Different from (Yuan & Kitani, 2020; Mao et al., 2021), we additionally consider translation \(T \in \mathbb{R}^3\) and rotation \(R \in \mathbb{R}^{3 \times 3}\) of the pose. Specifically, given a threshold \(\epsilon\), we cluster future motion with a similar start pose and train the model with their residuals as

\[
\{\hat{Y}_n^P\}_{p=1}^P = \{\hat{Y}_n^P|\min_{R,T} \|R(\hat{X}_n^P[T_h] - T) - X_i[T_h]\|_2 \leq \epsilon\}, \tag{8}
\]

\[
\mathcal{L}_{mmR} = \frac{1}{N \times P} \sum_{n=1}^{N} \sum_{p=1}^{P} \min_{m=1,\ldots,M} \|\Delta \hat{Y}_n^m - \Delta Y_n^p\|_2^2. \tag{9}
\]

To explicitly encourage diversity, we adopt a diversity-promoting loss (Yuan & Kitani, 2020), which directly promotes the pairwise distance between the predictions of a single person. We decompose this loss into two parts, promoting the diversity of local pose and the global root separately. Supposing that \(\hat{Y}_n^m(l)\) and \(\hat{Y}_n^m(g)\) are the local pose and the global root joint extracted from the global pose \(\hat{Y}_n^m\), respectively, and \(\alpha\) and \(\beta\) are two hyperparameters, this diversity-promoting loss is denoted as

\[
\mathcal{L}_D = \frac{1}{N \times M (M - 1)} \sum_{n=1}^{N} \sum_{m=1}^{M} \sum_{k=m+1}^{M} \left[\exp\left(\frac{\|\hat{Y}_n^m(g) - \hat{Y}_n^k(g)\|_2^2}{\alpha}\right) + \exp\left(\frac{\|\hat{Y}_n^m(l) - \hat{Y}_n^k(l)\|_2^2}{\beta}\right)\right]. \tag{10}
\]
Table B: Summary of the complementary evaluation metrics in the multi-person 3D motion forecasting task, with each focusing on evaluating different aspects of predicted motion. Here for simplicity, we show the metrics without alignment. We also provide ADE, FDE, and FPD with alignment to evaluate the pose and trajectory separately, as explained in Sec. 4 of the main paper.

<table>
<thead>
<tr>
<th>Type</th>
<th>Metrics</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Single-Person Fidelity</strong></td>
<td>Local Average Displacement Error (IADE)</td>
<td>$\frac{1}{N_T} \sum_{n=1}^{N} \min_m | \hat{Y}_n^m - Y_n |_2$</td>
</tr>
<tr>
<td></td>
<td>Local Final Displacement Error (IFDE)</td>
<td>$\frac{1}{N_T} \sum_{n=1}^{N} \min_m | \hat{Y}_n^m[T_f] - Y_n[T_f] |_2$</td>
</tr>
<tr>
<td></td>
<td>Foot Skating Ratio (FSR)</td>
<td>average ratio of frames where both foot joints are close to the ground ($\leq 5$cm) and fast ($\geq 75$mm/s)</td>
</tr>
<tr>
<td><strong>Multi-Person Fidelity</strong></td>
<td>(Global) Average Displacement Error (ADE)</td>
<td>$\min_{m} \frac{1}{N_T} \sum_{n=1}^{N} | \hat{Y}_n^m - Y_n |_2$</td>
</tr>
<tr>
<td></td>
<td>(Global) Final Displacement Error (FDE)</td>
<td>$\min_{m} \frac{1}{N_T} \sum_{n=1}^{N} | \hat{Y}_n^m[T_f] - Y_n[T_f] |_2$</td>
</tr>
<tr>
<td></td>
<td>Trajectory Collision Ratio (TCR)</td>
<td>average ratio of frames where there is a collision between any two trajectories</td>
</tr>
<tr>
<td></td>
<td>Average Human Displacement (AHD)</td>
<td>$\frac{1}{N_T} \sum_{n=1}^{N} \sum_{m=1}^{M} | \hat{Y}_n^m[T_f] - \hat{Y}_n^m[1] |_2$</td>
</tr>
<tr>
<td></td>
<td>Final Pairwise Distance (FPD)</td>
<td>$\frac{1}{N_T} \sum_{n=1}^{N} \sum_{m=1}^{M} \sum_{k=m+1}^{M} | \hat{Y}_n^m[T_f] - \hat{Y}_n^m[T_p] |_2$</td>
</tr>
</tbody>
</table>

For CGAN, a GAN loss (Kocabas et al., 2020) is leveraged to train the model and the local discriminator $D_l$ for individual body realism. Suppose $\{Y_n^N\}_{n=1}^{N}$ is the set of real motion clips sampled from the data, we have the following.

$$L_{GAN} = \frac{1}{N} \sum_{n=1}^{N} \sum_{m=1}^{M} \| D_l(\hat{Y}_n^m) \|_2^2 + \frac{1}{N} \sum_{n=1}^{N} \| D_l(Y_n^N) - 1 \|_2^2. \quad (11)$$

**Global-Level Objectives.** Since in this setting we treat $N$ individuals as a whole, the reconstruction loss is reformulated as

$$L_{GR} = \min_{m=1,...,M} \frac{1}{N} \sum_{n=1}^{N} \| \Delta \hat{Y}_n^m - \Delta Y_n \|_2^2. \quad (12)$$

For CGAN, a global GAN loss is further leveraged to promote the realism of social motion, where the global discriminator $D_g$ takes the motion of all $N$ people as input. Suppose $\{Y_n^{**}\}_{n=1}^{N}$ is the multi-person motion clip sampled from the data, and we have

$$L_{GAN} = \frac{1}{M} \sum_{m=1}^{M} \| D_g(\{\hat{Y}_n^m\}_{n=1}^{N}) \|_2^2 + \| D_g(\{Y_n^{**}\}_{n=1}^{N}) - 1 \|_2^2. \quad (13)$$

Table C: Quantitative results (w/ error bar) of our DuMMF on single-person accuracy.

<table>
<thead>
<tr>
<th>Method, # of Intents</th>
<th>$@t = 1s$</th>
<th>$@t = 2s$</th>
<th>$@t = 3s$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>IADE ↓</td>
<td>IFDE ↓</td>
<td>IADE ↓</td>
</tr>
<tr>
<td>DuMMF (Ours), 2</td>
<td>0.663 ± 0.007</td>
<td>1.088 ± 0.019</td>
<td>1.029 ± 0.019</td>
</tr>
<tr>
<td>DuMMF (Ours), 3</td>
<td>0.656 ± 0.007</td>
<td>1.056 ± 0.007</td>
<td>0.996 ± 0.009</td>
</tr>
<tr>
<td>DuMMF (Ours), 5</td>
<td>0.648 ± 0.011</td>
<td>1.033 ± 0.022</td>
<td>0.963 ± 0.020</td>
</tr>
</tbody>
</table>

D ADDITIONAL DETAILS OF GENERATIVE MODELS

In this paper, we mainly evaluate the generalizability of our method by incorporating it with a CGAN (Mirza & Osindero, 2014) or a DDPM (Ho et al., 2020) model. For CGAN, we introduce local discriminator and global discriminator for objectives at local and global levels, respectively. For the model incorporated with DDPM, we disregard these two GAN losses.
**Local Discriminator.** To ensure the fidelity of the motion, especially to address the problem of foot skating, where the feet appear to slide in the ground plane, we concatenate the predicted motion of each person $Y^j_t$ with the feet velocities $\Delta F^j_t$ as input to a local discriminator $D_l$. The local discriminator uses a local-range transformer encoder adopted from Wang et al. (2021b).

**Global Discriminator.** To ensure the realism of social interactions and avoid motion collisions, we propose a global discriminator $D_g$ that encodes all motion of $N$ people $(Y^i_{n_f+1})_N$ at the same time and outputs a fidelity score. The global discriminator uses a global-range transformer encoder adopted from Wang et al. (2021b).

### E Additional Implemental Details

Here, we provide more details on the implementation of our method DuMMF. The two hyperparameters ($\alpha$, $\beta$) in diversity promoting loss (Sec. C) are set to $(50, 100)$. The model is trained using a batch size of 32 for 50 epoch, with 6000 training examples per epoch. We use ADAM (Kingma & Ba, 2014) to train the model. The code is based on PyTorch. On one NVIDIA GeForce GTX TITAN X GPU, training an epoch takes approximately 5 minutes. For license, CMU-Mocap (CMU) is free for all users; MuPoTS-3D (Mehta et al., 2018) is for noncommercial purposes. Part of our code is based on AMCParser (MIT license), attention-is-all-you-need-pytorch (MIT license), and MRT (Wang et al., 2021b) (not specified), and XIA (Guo et al., 2022) (GPL license).

Table D: Quantitative comparison on MuPoTS-3D between our DuMMF and deterministic forecasting baselines and their CGAN variants. All models are trained only using skeletal representations on CMU-Mocap and we compare their generalizations on MuPoTS-3D here. The number of intents is set to 5 for stochastic forecasting on 3-person (top) and 2-person (bottom) scenarios. DuMMF significantly improves multi-person accuracy and diversity across various architectures and deterministic predictors.

<table>
<thead>
<tr>
<th>Architecture</th>
<th>Predictor</th>
<th>Variants</th>
<th>RNN</th>
<th>Transformer</th>
<th>Transformer</th>
<th>MRT (Wang et al., 2021b)</th>
<th>MRT (Wang et al., 2021b)</th>
<th>XIA (Guo et al., 2022)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNN</td>
<td>SC-MPF (Adell et al., 2020)</td>
<td>N/A</td>
<td>0.467</td>
<td>0.473</td>
<td>0.764</td>
<td>0.840</td>
<td>0.753</td>
<td>1.207</td>
</tr>
<tr>
<td></td>
<td>CGAN</td>
<td>N/A</td>
<td>0.697</td>
<td>0.478</td>
<td>0.793</td>
<td>1.154</td>
<td>0.770</td>
<td>1.258</td>
</tr>
<tr>
<td></td>
<td>CGAN+DuMMF (Ours)</td>
<td>w/o Separation</td>
<td>0.736</td>
<td>0.627</td>
<td>1.092</td>
<td>1.412</td>
<td>0.717</td>
<td>1.846</td>
</tr>
<tr>
<td></td>
<td>w/o Continuous</td>
<td>0.747</td>
<td>0.455</td>
<td>0.744</td>
<td>1.441</td>
<td>0.707</td>
<td>1.103</td>
<td>2.209</td>
</tr>
<tr>
<td>Transformer</td>
<td>SC-MPF (Adell et al., 2020)</td>
<td>N/A</td>
<td>0.704</td>
<td>0.400</td>
<td>0.690</td>
<td>1.099</td>
<td>0.659</td>
<td>1.077</td>
</tr>
<tr>
<td></td>
<td>CGAN</td>
<td>N/A</td>
<td>0.557</td>
<td>0.403</td>
<td>0.693</td>
<td>1.262</td>
<td>0.657</td>
<td>1.075</td>
</tr>
<tr>
<td></td>
<td>CGAN+DuMMF (Ours)</td>
<td>w/o Separation</td>
<td>0.917</td>
<td>0.406</td>
<td>0.707</td>
<td>2.124</td>
<td>0.865</td>
<td>1.163</td>
</tr>
<tr>
<td></td>
<td>w/o Continuous</td>
<td>0.905</td>
<td>0.402</td>
<td>0.701</td>
<td>1.016</td>
<td>0.656</td>
<td>1.071</td>
<td>2.092</td>
</tr>
<tr>
<td>MRT (Wang et al., 2021b)</td>
<td>CGAN</td>
<td>N/A</td>
<td>0.176</td>
<td>0.421</td>
<td>0.741</td>
<td>0.624</td>
<td>0.729</td>
<td>1.288</td>
</tr>
<tr>
<td></td>
<td>CGAN+DuMMF (Ours)</td>
<td>w/o Separation</td>
<td>0.191</td>
<td>0.425</td>
<td>0.744</td>
<td>0.434</td>
<td>0.741</td>
<td>1.264</td>
</tr>
<tr>
<td></td>
<td>w/o Continuous</td>
<td>0.113</td>
<td>0.431</td>
<td>0.745</td>
<td>0.330</td>
<td>0.752</td>
<td>1.286</td>
<td>0.027</td>
</tr>
<tr>
<td></td>
<td>w/o Separation</td>
<td>0.308</td>
<td>0.418</td>
<td>0.730</td>
<td>0.732</td>
<td>0.716</td>
<td>1.242</td>
<td>1.04</td>
</tr>
<tr>
<td></td>
<td>w/o Continuous</td>
<td>0.510</td>
<td>0.408</td>
<td>0.711</td>
<td>1.075</td>
<td>0.683</td>
<td>1.135</td>
<td>1.740</td>
</tr>
</tbody>
</table>

### F Summary of Evaluation Metrics

In the main paper, we mainly focused on evaluating the fidelity and diversity of the predicted multi-person motion based on Average Displacement Error (ADE), Final Displacement Error (FDE), and Final Pairwise Distance (FPD), and briefly discussed other metrics. Here, we explain these and additional metrics in detail and also provide a systematic summary in Table B for better understanding. Additional comparisons based on these metrics are shown in the following sections.

As summarized in Table B, we group the metrics into three types, with each type evaluating different aspects of predicted motion (discussed in Sec. 4 of the main paper) – single-person fidelity, multi-person fidelity, and overall diversity.
Table E: Quantitative comparison on SoMoF Benchmark. Here, we only show the deterministic forecasting results. Our method with MRT predictor significantly outperform two deterministic baselines. We use VIM (Adeli et al., 2020) as the metric. * means we directly report the results from the benchmark leaderboard.

<table>
<thead>
<tr>
<th>Method, # of Intents</th>
<th>Prediction Time</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>100 ms</td>
</tr>
<tr>
<td>SC-MPF* (Adeli et al., 2020), N/A</td>
<td>46.28</td>
</tr>
<tr>
<td>TRiPOD* (Adeli et al., 2021), N/A</td>
<td>30.26</td>
</tr>
<tr>
<td>MRT (Wang et al., 2021b), N/A</td>
<td>22.93</td>
</tr>
</tbody>
</table>

Table F: Quantitative comparisons between our DuMMF and deterministic forecasting baselines and their CGAN variants on SoMoF benchmark. The number of intents is set to 5 for stochastic forecasting. Our DuMMF significantly improves multi-person accuracy and diversity.

<table>
<thead>
<tr>
<th>Architecture</th>
<th>Predictor</th>
<th>Diversifier</th>
<th>Variants</th>
<th>@ t = 1s</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transformer</td>
<td>MRT (Wang et al., 2021b)</td>
<td>Deterministic</td>
<td>N/A</td>
<td>0.504</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CGAN</td>
<td>N/A</td>
<td>0.586</td>
</tr>
<tr>
<td></td>
<td></td>
<td>w/o Separation</td>
<td>1.075</td>
<td>0.567</td>
</tr>
<tr>
<td></td>
<td></td>
<td>w/o Discrete</td>
<td>1.063</td>
<td>0.587</td>
</tr>
<tr>
<td></td>
<td></td>
<td>w/o Continuous</td>
<td>0.969</td>
<td>0.564</td>
</tr>
</tbody>
</table>

Specifically, Local Average Displacement Error (lADE) and Local Final Displacement Error (lFDE) are proposed to further evaluate the single-person fidelity. They compute the average distance between the individual ground truth and the individual prediction closest to the individual ground truth. Note that ADE (or IADE) and FDE (or IFDE) only measure the best predictions in all outputs. While we can average ADE (or IADE) and FDE (or IFDE) by computing the distance between multiple predictions and a single ground truth, this way of assessing the overall prediction quality cannot reflect motion realism. For example, a very realistic but diverse set of outputs may have poor average ADE and FDE. Therefore, we do not use such metrics for our evaluation.

For diversity evaluation, a common metric is Average Pairwise Distance (APD) (Mao et al., 2021; Yuan & Kitani, 2020; 2019) that is the average $\ell_2$ distance between all predicted motion pairs. In this paper, we formulate diverse forecasting as producing more forecasts over time (see Sec. 3.3). This progressive generation is actually closer to reality because the multi-modality of motion should be more pronounced after further time. However, in this case, APD cannot reflect the diversity well, since many predictions will share the same previous segment. Therefore, we only examine the diversity of the last pose (FPD), as the last pose should not be the same.

Moreover, we introduce three tailored metrics to evaluate specific aspects of predicted motion, which correspond to the unique challenges in multi-person motion forecasting as discussed in the main paper. (a) Foot Skating Ratio (Zhang et al., 2021): the average ratio of frames with foot skating. (b) Trajectory Collision Ratio: the average ratio of predictions that is considered to have collision (Kothari et al., 2020) between any two trajectories in the scene. (c) Average Human Displacement: Average displacement of the predicted human body between the last frame and the first frame, reflecting the properties of the predicted motion distribution.

G ADDITIONAL EXPERIMENTAL RESULTS

Additional Quantitative Results. We compare our method with MRT (Wang et al., 2021b) in Table G. We show that the improvement is significant by performing each experiment five times with different random seeds and reporting error bars. When the number of intents is one, equivalent to a deterministic setting, our method marginally outperforms MRT for all metrics on CMU-Mocap and also generalizes better on MuPoTS-3D. This suggests the generality of our approach, which is also advantageous for deterministic processes.

In Table C, we use IADE and IFDE to compare single-person fidelity, and we observe that our model also significantly outperforms the baseline.
We observe a similar performance as on CMU-Mocap (CMU) and MuPoTS-3D (Mehta et al., 2018), we formulate this single-person forecasting architecture (named SRT) with a transformer encoder-

Additional Comparison with Single-Person Forecasting Methods. In Table H, we provide a comparison by applying a stochastic single-person forecasting baseline. For a fair comparison, we formulate this single-person forecasting architecture (named SRT) with a transformer encoder-decoder adapted from MRT (Wang et al., 2021b), where we modified the encoder and decoder to be individually independent. Table H illustrates that (1) the modeling of social interactions is crucial, since the single-person forecasting baseline has much worse performance; (2) the improvement from
Additional Ablation on Impact of Learning Objectives. In Table 1, we evaluate the impact of each loss term within the dual-level framework. In general, using all loss functions yields the best results, since its results are either the best or the second best. We observe that local and global discriminators not only make predictions more accurate, but also more diverse. Note that the reconstruction losses $L_{1R}$ and $L_{3R}$ optimize only the most accurate prediction of all the outputs. It is important to provide supervision for other predictions that are not the best. We observe that limb loss $L_L$ is crucial, as it is the only loss function that provides supervision for all outputs. The multi-modal reconstruction loss also has a large performance impact, since it provides supervision for more than one output.

Additional Analysis of the Number of Discrete Latent Codes. Note that the number of discrete latent codes is not restricted to the same number as the number of predictions per second. We chose our DuMMF for the multi-person forecasting model is much higher than that for the single-person forecasting model.

Table H: Quantitative comparison on CMU-Mocap between our DuMMF based on MRT (Wang et al., 2021b) and a single-person forecasting baseline adapted from MRT. The number of intents is set to 5 for stochastic forecasting on 3-person. For DuMMF with SRT, we use the variant of “w/o Separation” to skip the social interaction modeling in DuMMF, making it a full single-person prediction baseline.

Figure B: Qualitative results on CMU-Mocap. We evaluate the generalizability and scalability of our model to predict 3-second motion on the constructed 9-person motion test data. The top row is the historical pose, and the five end poses predicted by our model; and we show the sequence of one of the predictions in the bottom two rows (highlighted by the blue dashed box).
Table J: Ablation study of the number of discrete latent codes on CMU-Mocap using skeletal representations. For producing 5 predictions per second, we observe that both the accuracy and diversity of predictions decrease significantly as the number of discrete intents increases.

<table>
<thead>
<tr>
<th># of Intents</th>
<th>@t = 1s</th>
<th>@t = 2s</th>
<th>@t = 3s</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FPD ↑</td>
<td>ADE ↓</td>
<td>FDE ↓</td>
</tr>
<tr>
<td>5</td>
<td>0.716</td>
<td>0.658</td>
<td>1.053</td>
</tr>
<tr>
<td>6</td>
<td>0.498</td>
<td>0.667</td>
<td>1.087</td>
</tr>
<tr>
<td>7</td>
<td>0.465</td>
<td>0.671</td>
<td>1.093</td>
</tr>
<tr>
<td>9</td>
<td>0.263</td>
<td>0.669</td>
<td>1.087</td>
</tr>
<tr>
<td>15</td>
<td>0.231</td>
<td>0.678</td>
<td>1.127</td>
</tr>
</tbody>
</table>

them to be the same for simplicity and a better trade-off between prediction performance and training efficiency. We find that this setup also achieved the best performance. To ensure that all predictions come from different discrete codes, the number of discrete latent codes should not be less than the number of predictions. In Table J, we provide an ablation study when the number of predictions per second is 5. If the number of discrete intents is greater than 5, we randomly select a discrete code without replacement (excluding any of the previously selected intents) to generate a prediction. We observe that both the accuracy and diversity of the predictions decrease as the number of intents increases. An explanation of such behavior could be: When the number of discrete latent codes is the same as the number of predictions, each code is explicitly and fully optimized for a particular prediction, leading to the best prediction accuracy and diversity; whereas, when the number of intents increases, the random selection strategy may hurt performance, as the probability of selecting the best five discrete codes decreases. We hypothesize that a better training and selection strategy might improve performance with more discrete codes.

H Generalizability and Scalability: Evaluation on More-Person Scenarios

Datasets with More People per Scene. In the main paper, we constructed multi-person motion data with 3 people per scene. Here, we construct more challenging datasets with a significantly increased number of people; this also increases the complexity of social interactions. Specifically, we first sample 2-person and 1-person motion data from CMU-Mocap (CMU), and then compose them together. We use handcraft rules to filter out scenes with trajectory collisions. Instead of retraining the model in the novel setting with more people per scene, we directly evaluate the model trained on 3-person motion data, and test if our model is able to scale to scenarios with more people.

Qualitative Results. Similar to Figure 3 of the main paper, we provide visualizations for 6-person and 9-person scenarios in Figure A and Figure B of the end poses of predictions of 3 second. We randomly select a prediction and visualize the corresponding full motion sequence as well. Note that our model is trained on only 3-person data without any fine-tuning on more-person data, but it still performs well on more-person data, suggesting the generalizability and scalability of our approach.