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 complexity in prior work on human motion forecasting, emphasizing the social properties of multi-person motion, the diversity of motion and social interaction, 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 movements 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 movement and switching the codes' modes of operation at different levels. Our framework is general, and we instantiate it with 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. Code is available at https://github.com/Sirui-Xu/DuMMF.

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 motions are structured with respect to both body physics and social norms, and highly dependent on the surrounding environment and its changes; second, human motion is inherently uncertain and multimodal, especially in the long-term horizon.

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 interaction 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 interaction, *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). However, this task only considers the motion and interaction at the trajectory level. Modeling articulated motion 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 interaction, and the complexity of articulated motion. As shown in Figure 1, this task requires the predicted multiple

* Yu-Xiong Wang and Liang-Yan Gui contributed equally to this work.

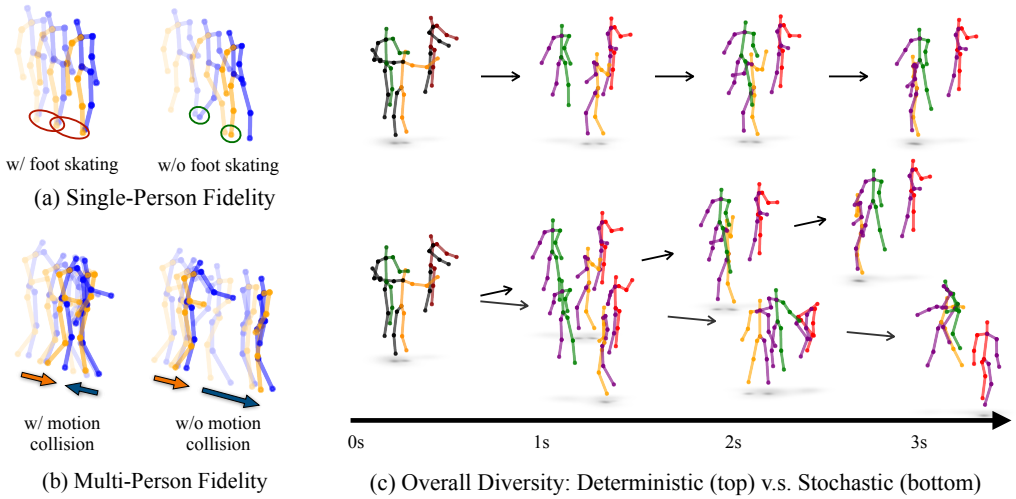

Figure 1: Illustration of the multifaceted challenges in the proposed task of stochastic multi-person 3D motion forecasting. (a) single-person fidelity: for each person, the predicted pose and trajectory should be realistic and consistent with each other, *e.g.*, to avoid foot skating. (b) multi-person 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.

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) above.

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 ***Du**al-level generative modeling framework for **M**ulti-person **M**otion **F**orecasting* (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 tackle 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 and learnable intent codes to improve the realism and diversity of predictions. (d) Our framework is general and can be operationalized with different types of multi-person motion forecasting models. Notably, it can be generalized to unseen data and challenging scenarios with more persons (Sec. G).

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;

Kundu et al., 2019) are based on the improvement and adaptation of deep generative models such as variational autoencoders (VAEs) (Kingma & Welling, 2013), generative adversarial networks (GANs) (Goodfellow et al., 2014), and normalizing flows (NFs) (Rezende & Mohamed, 2015). 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) have put more emphasis on the promotion of diversity. Yuan & Kitani (2020) introduce remapping in the latent space and ensure that the samples in this space are sufficiently diverse. Mao et al. (2021) propose a framework that sequentially generates the different parts of a pose for better controllability of diversity. Although these methods can generate very diverse human motions, most of them are limited to local motion and ignore the global trajectory. Some of the motions they produce are actually unreal; for example, there may be severe foot skating when global trajectories are considered. 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 adapt these two loss functions into our dual-level modeling. Kothari et al. (2021) introduce discrete social anchors and redefine the distribution of trajectories by the predefined anchors. Unlike their hand-crafted anchors, our discrete social intent codes are learnable components obtained through dual-level optimization. Some recent attempts exploit *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 proposed task. As illustrated in Figure 2, our key insight is to *separately model 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 these multi-person predictors used in this paper (Sec. 3.4).

Problem Formulation. We denote the input motion sequence of length T_h for N persons in a scene as $\{\mathbf{X}_n\}_{n=1}^N$, where $\mathbf{X}_i[t]$ is the pose of i-th person at time step t. We aim to predict M future motion sequences of length T_p, denoted as $\{\{\widehat{\mathbf{Y}}_n^m\}_{n=1}^N\}_{m=1}^M$, where $\widehat{\mathbf{Y}}_i^j = [\widehat{\mathbf{Y}}_i^j[T_h+1], \ldots, \widehat{\mathbf{Y}}_i^j[T_h+T_p]]$ is the j-th predicted motion of i-th person. We use the 3D coordinates to represent the absolute joint position of V joints, hence $\forall i, j, t, \mathbf{X}_i[t], \widehat{\mathbf{Y}}_i^j[t] \in \mathbb{R}^{V\times 3}$. We assume to be given the ground truth motion of N persons as $\{\mathbf{Y}_n\}_{n=1}^N$. Our goal is to forecast multiple realistic yet diverse future motions, such that (a) all the M predictions represent human-like motion sequences, 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 $p(\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_N | \mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N)$, where all future movements $\{\mathbf{Y}_n\}_{n=1}^N$ are conditioned on the past sequences $\{\mathbf{X}_n\}_{n=1}^N$ of all per-

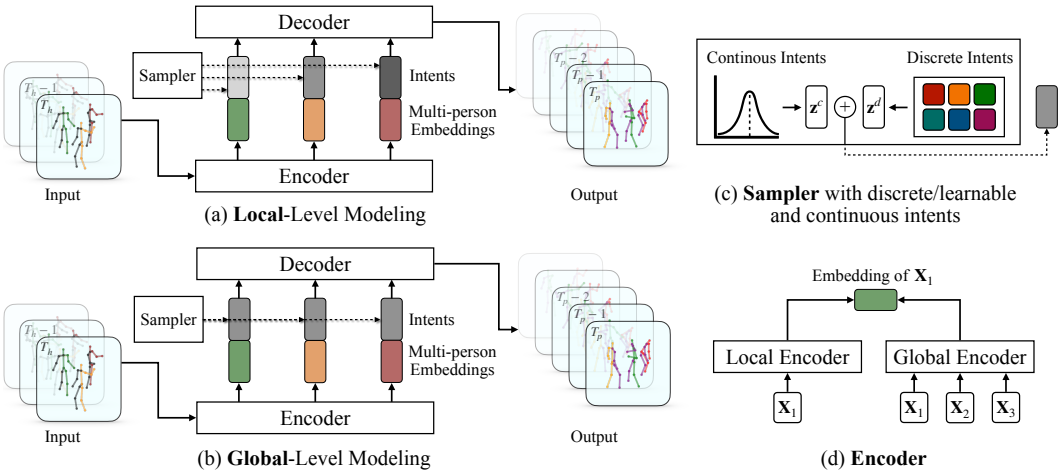

Figure 2: Overview of our proposed dual-level generative modeling with motion inputs of three persons. (a) We combine encoded multi-person embeddings with *independent* intent codes at the local level of modeling individual motion; (b) Differently, the global level of modeling social interactions requires all latent codes to be the *same*; (c) The latent codes consist of discrete intent codes selected from a given set and continuous intent codes; (d) The encoder of the multi-person predictor has a local branch that encodes single-person motions and a global branch that encodes multi-person motions (see Table A for instantiations).

sons. Typically, we can use a latent code $\mathbf{z} \sim p(\mathbf{z})$ to reparameterize this joint distribution as $p(\{\mathbf{Y}_n\}_{n=1}^N|\{\mathbf{X}_n\}_{n=1}^N) = \int p(\{\mathbf{Y}_n\}_{n=1}^N|\{\mathbf{X}_n\}_{n=1}^N, \mathbf{z})p(\mathbf{z})\mathrm{d}\mathbf{z}$. Here, the latent code $\mathbf{z}$ can be seen as a *social intent* that guides the future movements of multiple persons. With this social intent z sampled from the given distribution $p(\mathbf{z})$, a deterministic neural network $\mathcal{G}_\theta$ can map the historical conditions $\{\mathbf{X}_n\}_{n=1}^N$ to a prediction $\{\widehat{\mathbf{Y}}_n\}_{n=1}^N$ under the latent distribution $p_\theta(\{\mathbf{Y}_n\}_{n=1}^N|\{\mathbf{X}_n\}_{n=1}^N)$, which is formulated as

$$\mathbf{z} \sim p(\mathbf{z}),\ \{\widehat{\mathbf{Y}}_n\}_{n=1}^N = \mathcal{G}_\theta(\mathbf{z}, \{\mathbf{X}_n\}_{n=1}^N). \tag{1}$$

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 $\mathbf{z}$ w.r.t. different levels of modeling, without any change to the model architecture $\mathcal{G}$.

Local-Level Modeling: Individual Motion. At this level, the generative model $\mathcal{G}_\theta$ 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 static motion and foot skating. Here, the joint distribution of future human motions can be rewritten in the form of all single-person marginal distributions, *i.e.*, $p(\{\mathbf{Y}_n\}_{n=1}^N|\{\mathbf{X}_n\}_{n=1}^N) = \prod_{n=1}^N p(\mathbf{Y}_n|\mathbf{X}_n)$. To this end, we consider leveraging N different *individual intents* $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N$ independently drawn from $p(\mathbf{z})$ to generate independent future movements, denoted as

$$\mathbf{z}_1, \ldots, \mathbf{z}_n \sim p(\mathbf{z}),\ \{\widehat{\mathbf{Y}}_n\}_{n=1}^N = \mathcal{G}_\theta(\{\mathbf{z}_n\}_{n=1}^N, \{\mathbf{X}_n\}_{n=1}^N). \tag{2}$$

Global-Level Modeling: Social Interactions. Going beyond the local individual level, the generative model $\mathcal{G}_\theta$ 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*, mitigating issues such as human collisions. As illustrated in Figure 2(b), to maintain the network architecture $\mathcal{G}$ 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

$$\mathbf{z} \sim p(\mathbf{z}),\ \{\widehat{\mathbf{Y}}_n\}_{n=1}^N = \mathcal{G}_\theta(\{\mathbf{z}\}_{n=1}^N, \{\mathbf{X}_n\}_{n=1}^N). \tag{3}$$

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 $\mathbf{z}$ (Sec. 3.2), jointly optimize the codes $\mathbf{z}$ and the forecasting model $\mathcal{G}$ 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 (DCM) (Aguirregabiria & 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. However, the intent options for social interaction are more subtle and difficult to define manually than the intent in other applications such as for trajectories (Kothari et al., 2021). Therefore, we use a discrete set of learnable codes $\mathbf{z}^d \in \{\mathbf{Z}^i\}_{i=1}^M$ to represent social intents, similar to Xu et al. (2022b), while having different training strategies specific to this new task, as we will introduce in 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 $\mathbf{z}^c \sim p(\mathbf{z})$ of the generative model $\mathcal{G}_\theta$ 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

$$\mathbf{z}^c \sim p(z),\ \mathbf{z}^d \in \{\mathbf{Z}^i\}_{i=1}^M,\ \{\widehat{\mathbf{Y}}_n\}_{n=1}^N = \mathcal{G}_\theta(\{\mathbf{z}^c + \mathbf{z}^d\}_{n=1}^N, \{\mathbf{X}_n\}_{n=1}^N). \tag{4}$$

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

$$\mathbf{z}_1^c, \dots, \mathbf{z}_n^c \sim p(z),\ \mathbf{z}_1^d, \dots, \mathbf{z}_n^d \in \{\mathbf{Z}^i\}_{i=1}^M,\ \{\widehat{\mathbf{Y}}_n\}_{n=1}^N = \mathcal{G}_\theta(\{\mathbf{z}_n^c + \mathbf{z}_n^d\}_{n=1}^N, \{\mathbf{X}_n\}_{n=1}^N). \tag{5}$$

3.3 Training Under the Guidance of Level-Specific Objectives

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 $\mathbf{z}^d$ are optimized separately at different levels, while the parameters θ of the forecasting model $\mathcal{G}$ are updated based on the fused losses from the two levels.

Local-Level Training. At the local level, we train the model without considering the social interactions. Given the multi-person motion data $(\{\mathbf{X}_n\}_{n=1}^N, \{\mathbf{Y}_n\}_{n=1}^N)$, we first randomly sample the discrete intent codes and merge them with independently sampled random intent codes into $M \times N$ different latent codes $\{\{\mathbf{z}_n^m\}_{n=1}^N\}_{m=1}^M$. We then use each intent and the final historical pose of each person to query M future motion residuals for each person $\{\{\Delta\widehat{\mathbf{Y}}_n^m\}_{n=1}^N\}_{m=1}^M$, and recover M future motion sequences $\{\{\widehat{\mathbf{Y}}_n^m\}_{n=1}^N\}_{m=1}^M$. Local-level objectives fall into two categories, considering single-person fidelity and overall diversity respectively. We define the following objectives for constraining single-person fidelity: (a) local reconstruction loss $\mathcal{L}_{\mathrm{lR}}$; (b) limb loss $\mathcal{L}_{\mathrm{L}}$; (c) local GAN loss $\mathcal{L}_{\mathrm{lGAN}}$. To promote overall diversity, we leverage the following objectives: (a) multi-modal reconstruction loss $\mathcal{L}_{\mathrm{mmR}}$; (b) diversity-promoting loss $\mathcal{L}_{\mathrm{D}}$.

Global-Level Training. We train the model to facilitate the modeling of social interactions at the global level. The difference from the local setting is that now we incorporate the discrete and continuous codes into only M distinct latent codes $\{\{\mathbf{z}^m\}_{n=1}^N\}_{m=1}^M$, which means that the latent codes $\mathbf{z}^m$ of all N individuals are the same for a certain m-th prediction. Correspondingly, we introduce the following objectives to facilitate multi-person fidelity: (a) global reconstruction loss $\mathcal{L}_{\mathrm{gR}}$; (b) global GAN loss $\mathcal{L}_{\mathrm{gGAN}}$.

Please refer to the Appendix (Sec. C) for more detail on the above level-specific learning objectives. We demonstrate that these learning objectives are important, and in some cases critical, to the fidelity and diversity of multi-person motion (see Sec. G of the Appendix).

Table 1: Quantitative results (w/ error bar) of DuMMF and the readily available baseline MRT (Wang et al., 2021b). The baseline and our models are trained only on CMU-Mocap, and are tested on CMU-Mocap (top) and MuPoTS-3D (bottom). With the same backbone, our DuMMF framework significantly outperforms MRT on the deterministic prediction, and provides more accurate and diverse predictions with more intents and predictions.

Method, # of Intents	@$t = 1s$			@$t = 2s$			@$t = 3s$		
	FPD ↑	ADE ↓	FDE ↓	FPD ↑	ADE ↓	FDE ↓	FPD ↑	ADE ↓	FDE ↓
MRT (Wang et al., 2021b), N/A	N/A	0.682 ± 0.005	1.127 ± 0.006	N/A	1.082 ± 0.006	1.765 ± 0.021	N/A	1.435 ± 0.014	2.449 ± 0.057
+ DuMMF (Ours), 1	N/A	$\mathbf{0.670 \pm 0.005}$	$\mathbf{1.117 \pm 0.014}$	N/A	$\mathbf{1.061 \pm 0.009}$	$\mathbf{1.709 \pm 0.028}$	N/A	$\mathbf{1.380 \pm 0.017}$	$\mathbf{2.293 \pm 0.062}$
+ DuMMF (Ours), 2	0.435 ± 0.045	0.668 ± 0.008	1.098 ± 0.018	0.831 ± 0.097	1.044 ± 0.019	1.645 ± 0.045	1.717 ± 0.139	1.342 ± 0.033	2.117 ± 0.085
+ DuMMF (Ours), 3	0.569 ± 0.050	0.664 ± 0.007	1.077 ± 0.009	1.106 ± 0.101	1.019 ± 0.009	1.566 ± 0.029	1.717 ± 0.170	1.286 ± 0.020	1.999 ± 0.073
+ DuMMF (Ours), 5	0.633 ± 0.042	0.658 ± 0.011	1.060 ± 0.020	1.316 ± 0.111	0.992 ± 0.020	1.475 ± 0.033	2.112 ± 0.274	1.226 ± 0.025	1.809 ± 0.049
MRT (Wang et al., 2021b), N/A	N/A	0.427 ± 0.008	0.749 ± 0.012	N/A	0.750 ± 0.011	1.377 ± 0.033	N/A	1.082 ± 0.020	2.059 ± 0.062
+ DuMMF (Ours), 1	N/A	$\mathbf{0.411 \pm 0.008}$	$\mathbf{0.727 \pm 0.014}$	N/A	$\mathbf{0.713 \pm 0.012}$	$\mathbf{1.270 \pm 0.035}$	N/A	$\mathbf{0.996 \pm 0.027}$	$\mathbf{1.820 \pm 0.069}$
+ DuMMF (Ours), 2	0.334 ± 0.110	0.416 ± 0.003	0.729 ± 0.011	0.639 ± 0.233	0.715 ± 0.015	1.253 ± 0.049	1.000 ± 0.377	0.990 ± 0.036	1.783 ± 0.104
+ DuMMF (Ours), 3	0.472 ± 0.055	0.410 ± 0.007	0.712 ± 0.013	0.986 ± 0.121	0.694 ± 0.019	1.175 ± 0.068	1.623 ± 0.185	0.938 ± 0.041	1.586 ± 0.143
+ DuMMF (Ours), 5	0.513 ± 0.061	0.405 ± 0.007	0.703 ± 0.014	1.122 ± 0.150	0.678 ± 0.018	1.139 ± 0.046	1.887 ± 0.277	0.905 ± 0.034	1.509 ± 0.083

Inference. We only use the global level where we sample the same intent for all people in the scene during inference. Note that the uncertainty of human motion substantially increases with the time horizons of the forecasting. So instead of predicting a fixed number of diverse outputs (Mao et al., 2021; Yuan & Kitani, 2020), the number here should grow with the length of the prediction. To this end, we employ autoregressive inference (Wang et al., 2021b) and propose *progressive diversity*. Given the historical motion sequence $\{\mathbf{X}_n\}_{n=1}^N$ and M sampled intent codes, the model outputs M predictions $\{\{\widehat{\mathbf{Y}}_n^m\}_{n=1}^N\}_{m=1}^M$. And then, given a combination of history, the previous predictions $\{\{\mathbf{X}_n, \mathbf{Y}_n^m\}_{n=1}^N\}_{m=1}^M$, and M new intent codes as input, the model outputs M^2 predictions $\{\{\widehat{\mathbf{Y}}_n^m\}_{n=1}^N\}_{m=1}^{M^2}$, etc. (see Figure 1(c)).

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. To demonstrate this, we combine our framework with various types of deterministic multi-person motion predictors, yielding significant improvements across all predictors (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 multi-person predictors used in the paper is given in Table A of the Appendix. And we further detail how we adapt a state-of-the-art deterministic multi-person predictor (Wang et al., 2021b) to our framework in Sec. D 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 2235 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 3-person scenarios, 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 9 people per scene. Please refer to Sec. H of the Appendix for the detail of this generalized setting. We use 15-joint skeleton coordinates with global translation as the representation of the data. Unlike Wang et al. (2021b), we convert all length units to meters as suggested in CMU-Mocap dataset. 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 ℓ_2 distance over time between the

Table 2: Quantitative comparison on CMU-Mocap between our DuMMF and deterministic forecasting baselines and their CGAN stochastic variants. 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.

Architecture	Predictor	Diversifier	Variants	@$t = 1s$			@$t = 2s$			@$t = 3s$		
				FPD ↑	ADE ↓	FDE ↓	FPD ↑	ADE ↓	FDE ↓	FPD ↑	ADE ↓	FDE ↓
RNN	SC-MPF (Adeli et al., 2020)	Deterministic	N/A	N/A	0.767	1.267	N/A	1.274	2.213	N/A	1.779	3.262
		CGAN		0.584	0.732	1.206	1.188	1.170	1.887	1.899	1.518	2.411
		CGAN+DuMMF (Ours)	w/o Separation	0.618	0.739	1.225	1.309	1.188	1.925	2.131	1.538	2.413
			w/o Discrete	0.772	0.735	1.213	1.710	1.170	1.887	3.046	1.513	2.379
			w/o Continuous	**1.051**	0.811	1.368	**2.090**	1.339	2.226	**3.326**	1.791	2.966
			Full	1.041	**0.702**	**1.133**	2.035	**1.091**	**1.692**	3.097	**1.381**	**2.082**
Transformer	SC-MPF (Adeli et al., 2020)	Deterministic	N/A	N/A	0.697	1.167	N/A	1.141	1.908	N/A	1.523	2.637
		CGAN		0.454	0.681	1.111	1.036	1.059	1.649	1.743	1.346	2.094
		CGAN+DuMMF (Ours)	w/o Separation	0.624	0.676	1.099	1.461	1.038	1.592	2.527	1.305	1.954
			w/o Discrete	0.526	0.681	1.110	1.219	1.065	1.665	2.118	1.356	2.102
			w/o Continuous	**0.931**	**0.666**	**1.083**	**2.016**	1.029	1.578	**3.258**	1.298	1.963
			Full	0.888	0.671	1.085	1.797	**1.027**	**1.564**	2.798	**1.285**	**1.911**
	MRT (Wang et al., 2021b)	Deterministic	N/A	N/A	0.681	1.125	N/A	1.082	1.765	N/A	1.427	2.438
		CGAN		0.282	0.662	1.086	0.662	1.023	1.567	1.199	1.287	1.968
		CGAN+DuMMF (Ours)	w/o Separation	0.291	0.677	1.110	0.676	1.056	1.624	1.166	1.328	2.021
			w/o Discrete	0.169	0.669	1.093	0.414	1.049	1.674	0.783	1.353	2.184
			w/o Continuous	0.403	0.673	1.094	0.894	1.035	1.584	1.526	1.306	2.004
			Full	**0.716**	**0.658**	**1.053**	**1.435**	**0.993**	**1.472**	**2.206**	**1.232**	**1.823**
Transformer	MRT (Wang et al., 2021b)	Deterministic	N/A	N/A	0.685	1.138	N/A	1.152	2.024	N/A	1.624	3.018
		CGAN		0.675	0.688	1.442	1.391	1.094	1.750	2.205	1.443	2.398
		CGAN+DuMMF (Ours)	w/o Separation	0.753	0.663	1.069	1.499	1.013	1.533	2.259	1.284	1.970
			w/o Discrete	0.149	0.692	1.161	0.334	1.144	1.937	0.559	1.562	2.770
			w/o Continuous	0.827	**0.623**	**1.010**	1.709	**0.971**	1.485	2.692	**1.226**	1.823
			Full	**1.211**	0.656	1.052	**2.393**	0.992	**1.470**	**3.716**	1.238	**1.796**
	XIA (Guo et al., 2022)	Deterministic	N/A	N/A	0.679	1.136	N/A	1.121	1.910	N/A	1.529	2.727
		CGAN		0.578	0.667	1.088	1.208	1.044	1.631	1.882	1.340	2.111
		CGAN+DuMMF (Ours)	w/o Separation	0.604	0.657	1.068	1.183	1.018	1.589	1.746	1.311	2.071
			w/o Discrete	0.433	0.662	1.080	1.025	1.041	1.663	1.757	1.355	2.183
			w/o Continuous	0.234	0.669	1.113	0.458	1.090	1.814	0.720	1.463	2.507
			Full	**1.143**	**0.634**	**1.010**	**2.264**	**0.957**	**1.423**	**3.477**	**1.197**	**1.764**

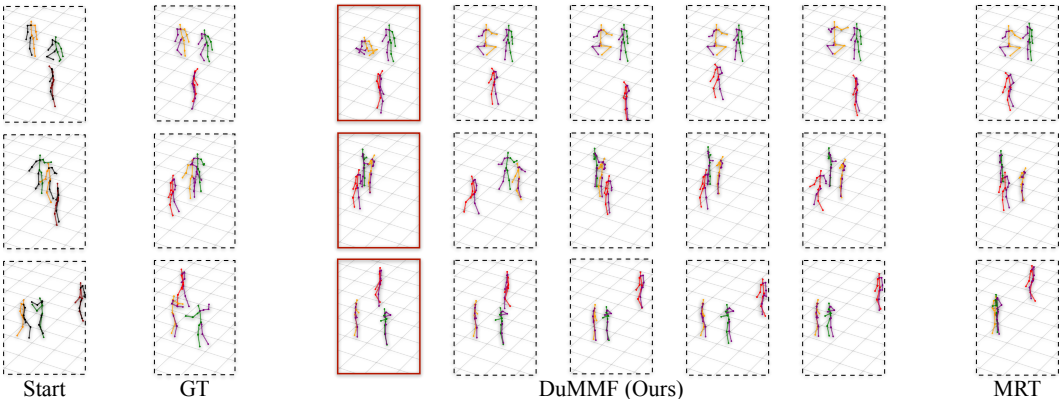

Figure 3: Qualitative comparison on 3-person scenarios. The leftmost column is the ground truth start poses, and the second column is the ground truth poses three seconds later. We show our five sampled 3-second predictions in the middle. Compared with MRT (rightmost column), our model produces diverse predictions, with one closer to the ground truth (highlighted by the red box).

ground truth and the prediction *closest* to the ground truth; (b) **Final Displacement Error (FDE)**: the ℓ_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 ℓ_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.

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 summarized in Table. A). And we 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 their paper and reproduce the results. For the

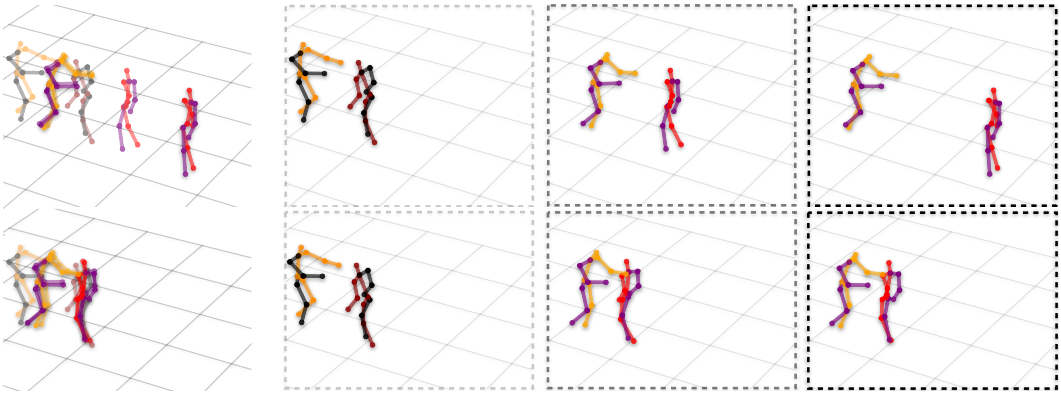

Figure 4: Qualitative comparisons on 2-person scenarios. We visualize the 2 sampled motions at 1, 2, and 3 seconds, stacked in one column on the left and listed sequentially in time in the right three columns. Our method can predict reasonably diverse human-like social interactions.

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 to use Cross Interaction Attention (XIA) to encode 2-person information. (d) **MRT** (Wang et al., 2021b): As mentioned in Sec. 3.4, we use this as our main architecture for its state-of-the-art performance. We further extend MRT to 2-person scenarios to compare with **XIA**. For a fair comparison, we use their respective encoders, but we use the *same transformer-based decoder* for the three baselines (b), (c), and (d). For stochastic forecasting baselines, we formulate conditional generative adversarial networks (CGAN) (Mirza & Osindero, 2014) with noise input and all learning objectives that we introduced at the local and global levels, including two GAN losses $\mathcal{L}_{\text{gGAN}}$ and $\mathcal{L}_{\text{lGAN}}$, without dual-level modeling and discrete intents.

Implementation Details. We train the model to predict a 15-frame sequence of 3 people given the real 15, 30, and 45 historical frames at 15Hz, since our encoder (RNN/Transformer) accepts different input lengths. The local and global encoders and the decoder all consist of 3-layer transformers (or RNNs), where we set the dimension of feature to 128. The model is trained with a batch size of 32. As illustrated in Sec. 3.3, the two levels of modeling result in different outputs and have their own learning objectives. For individual modeling, the weight of each loss term $(\lambda_{\text{lR}}, \lambda_{\text{L}}, \lambda_{\text{lGAN}}, \lambda_{\text{mmR}}, \lambda_{\text{D}})$ is $(0.01, 0.0005, 0.0005, 0.0001, 0.0001)$. For social modeling, the weight for each loss term $(\lambda_{\text{gR}}, \lambda_{\text{gGAN}})$ is $(1, 0.0001)$. For evaluation, we use 15 real frames as historical motion and recursively predict the next 15 frames 3 times, as illustrated in Sec. 3.3. Thus, given the number of intents to be M, the model outputs M, M^2, M^3 different predictions in sequence. Additional implementation details are provided in Sec. E of the Appendix.

Quantitative Results. We compare our method with MRT (Wang et al., 2021b) in Table 1. 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, suggesting the *generality* of our approach, which is also helpful for deterministic processes. As the number of intents increases, our method *consistently* provide more accurate and diverse results, especially in long-term predictions. 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, suggesting the *generality* and *superiority* of our approach.

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

Table 3: Ablation study of our DuMMF model with MRT (Wang et al., 2021b) on CMU-Mocap and MuPoTS-3D. We report the accuracy and diversity for root and pose respectively. The results show the effectiveness of our dual-level modeling along with discrete motion intent, and complementariness of local-level and global-level modeling as well as discrete and continuous intent codes.

Method @$t = 3s$	CMU-Mocap (CMU)						Generalized to MuPoTS-3D (Mehta et al., 2018)					
	ROOT			POSE			ROOT			POSE		
	FPD ↑	ADE ↓	FDE ↓	FPD ↑	ADE ↓	FDE ↓	FPD ↑	ADE ↓	FDE ↓	FPD ↑	ADE ↓	FDE ↓
MRT (Wang et al., 2021b)	-	0.753	0.946	-	0.301	0.567	-	0.603	0.835	-	0.220	0.461
Ours, Deterministic	-	**0.738**	**0.916**	-	**0.288**	**0.526**	-	**0.587**	**0.809**	-	**0.203**	**0.411**
w/o Individual	0.178	**0.733**	**0.896**	0.339	0.264	0.453	0.120	**0.584**	0.797	0.253	0.180	0.331
w/o Social	**1.661**	0.751	0.909	**1.077**	0.264	0.413	**2.100**	0.605	0.818	**1.283**	0.183	0.318
w/o Separation	0.167	0.747	0.911	0.293	0.271	0.448	0.105	0.596	0.815	0.213	0.202	0.386
Full	0.249	0.734	0.898	0.562	**0.243**	**0.390**	0.187	0.588	**0.796**	0.444	**0.170**	**0.305**
w/o Discrete	0.114	0.761	0.949	0.196	0.279	0.490	0.083	0.620	0.857	0.158	0.212	0.411
w/o Continuous	0.208	0.746	0.905	0.381	0.265	0.439	0.159	0.594	0.801	0.328	0.196	0.378
Full	**0.249**	**0.734**	**0.898**	**0.562**	**0.243**	**0.390**	**0.187**	**0.588**	**0.796**	**0.444**	**0.170**	**0.305**

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, 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 separating two-level 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 better integrate the advantages of the two levels, thus leading to improved accuracy and diversity.

Ablation: Effectiveness of Discrete Human Intent. In Tables 2 and 3, we would like to show 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'). And the improvements of dual-level modeling is pronounced in the case with discrete intent ('Full' vs. 'w/o Separation'). This further suggests that discrete learnable intents are critical to ensure that the advantages of both levels are integrated during training. The performance without continuous intent ('w/o Continuous') is slightly worse than the full method. Still, discrete intents only support a finite number of outputs, which is limited and not what we want. In Sec. G of the Appendix, we further analyze the effect of the number of discrete intents to the 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 3. In Table 3, 'w/o Social' results in extremely high diversity, especially in trajectories. In Figure 4, we qualitatively show that our generated results reflects real-world diversity of social interactions. Furthermore, we provide qualitative results for 6- and 9-person scenarios 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. Our dual-level framework with level-specific losses enables the model to produce high-quality and diverse predictions, but we still observe artifacts such as foot skating in some predicted motion. This is because we are not intrinsically modeling articulated motion, but only constraining it by some loss functions. This is a common problem with learning-based methods. Therefore, we consider exploiting a physical simulator to further improve the plausibility of the predicted motion.

5 Conclusion

We propose a new task, namely, stochastic multi-person 3D motion forecasting. To simultaneously achieve single-person fidelity, social realism, and overall diversity, we introduce 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 motions and interactions.

Ethics Statement. Our proposed technique is useful in many applications, such as self-driving to avoid crowds. The potential negative societal impact includes: (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, and we are concerned that this may expose the identity information and violate people's privacy. However, on the positive side, our model operates on the processed human skeleton representation, which contains much less identification information compared to the raw data.

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-Insper 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 National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296.

REFERENCES

Cmu graphics lab motion capture database. http://mocap.cs.cmu.edu/. 6, 9, 17, 19, 21

Vida Adeli, Ehsan Adeli, Ian Reid, Juan Carlos Niebles, and Hamid Rezatofighi. Socially and contextually aware human motion and pose forecasting. *IEEE Robotics and Automation Letters*, 2020. 1, 2, 3, 6, 7, 8, 14, 15, 17, 18

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. *arXiv preprint arXiv:2104.04029*, 2021. 1, 2, 3, 6, 14, 18

Victor Aguirregabiria and Pedro Mira. Dynamic discrete choice structural models: A survey. *Journal of Econometrics*, 2010. 5

Alexandre Alahi, Vignesh Ramanathan, and Li Fei-Fei. Socially-aware large-scale crowd forecasting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014. 1, 3, 14

Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016. 3, 7

Javad Amirian, Jean-Bernard Hayet, and Julien Pettré. Social ways: Learning multi-modal distributions of pedestrian trajectories with gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019. 3

Emad Barsoum, John R. Kender, and Zicheng Liu. Hp-gan: Probabilistic 3d human motion prediction via gan. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018. 2

Chandra Bhat, Naveen Eluru, and Rachel Copperman. Flexible model structures for discrete choice analysis. *Handbook of Transport Modelling*, 2008. 5

Apratim Bhattacharyya, Mario Fritz, and Bernt Schiele. Accurate and diverse sampling of sequences based on a "best of many" sample objective. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 3

Zhe Cao, Hang Gao, Karttikeya Mangalam, Qi-Zhi Cai, Minh Vo, and Jitendra Malik. Long-term human motion prediction with scene context. In *Proceedings of the European Conference on Computer Vision*, 2020. 3

Nat Dilokthanakul, Pedro A. M. Mediano, Marta Garnelo, M. J. Lee, Hugh Salimbeni, Kai Arulkumaran, and Murray Shanahan. Deep unsupervised clustering with gaussian mixture variational autoencoders. *arXiv preprint arXiv:1611.02648*, 2016. 3

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2014. 3

Tianpei Gu, Guangyi Chen, Junlong Li, Chunze Lin, Yongming Rao, Jie Zhou, and Jiwen Lu. Stochastic trajectory prediction via motion indeterminacy diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17113–17122, 2022. 14

Wen Guo, Xiaoyu Bie, Xavier Alameda-Pineda, and Francesc Moreno-Noguer. Multi-person extreme motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13053–13064, 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 3

Swaminathan Gurumurthy, Ravi Kiran Sarvadevabhatla, and R. Venkatesh Babu. Deligan : Generative adversarial networks for diverse and limited data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 3

Abner Guzmán-rivera, Dhruv Batra, and Pushmeet Kohli. Multiple choice learning: Learning to produce multiple structured outputs. In *Advances in Neural Information Processing Systems*, 2012. 15

Mohamed Hassan, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, and Michael Black. Stochastic scene-aware motion prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, 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 Zhaoqi Wang. Stgat: Modeling spatial-temporal interactions for human trajectory prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6272–6281, 2019. 1

Boris Ivanovic and Marco Pavone. The trajectron: Probabilistic multi-agent trajectory modeling with dynamic spatiotemporal graphs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2375–2384, 2019. 1

Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. 2014. 17

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *arXiv preprint arXiv:1312.6114*, 2013. 3

Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 15

Vineet Kosaraju, Amir Sadeghian, Roberto Martín-Martín, Ian Reid, S Hamid Rezatofighi, and Silvio Savarese. Social-bigat: Multimodal trajectory forecasting using bicycle-gan and graph attention networks. *arXiv preprint arXiv:1907.03395*, 2019. 1, 3

Parth Kothari, Sven Kreiss, and Alexandre Alahi. Human trajectory forecasting in crowds: A deep learning perspective. In *arXiv preprint arXiv:2007.03639*, 2020. 14, 18, 21

Parth Kothari, Brian Sifringer, and Alexandre Alahi. Interpretable social anchors for human trajectory forecasting in crowds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 3, 5, 6, 14

Jogendra Kundu, Maharshi Gor, and R. Babu. Bihmp-gan: Bidirectional 3d human motion prediction gan. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019. 3

Namhoon Lee, Wongun Choi, Paul Vernaza, Christopher B Choy, Philip HS Torr, and Manmohan Chandraker. Desire: Distant future prediction in dynamic scenes with interacting agents. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 3

Giorgio Leonardi. The stucture of random utility models in the light of the asymptotic theory of extremes. In *Theory and decision*, 1984. 5

Jiachen Li, Fan Yang, Masayoshi Tomizuka, and Chiho Choi. Evolvegraph: Multi-agent trajectory prediction with dynamic relational reasoning. *Advances in neural information processing systems*, 33:19783–19794, 2020. 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 *Proceedings of the European Conference on Computer Vision*, 2020. 3

Wei Mao, Miaomiao Liu, and Mathieu Salzmann. Generating smooth pose sequences for diverse human motion prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 1, 3, 6, 15, 18

Ramin Mehran, Alexis Oyama, and Mubarak Shah. Abnormal crowd behavior detection using social force model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009. 3

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 *Proceedings of the International Conference on 3D Vision*, 2018. 6, 9, 17, 19

Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets, 2014. 8

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

Behnam Parsaeifard, Saeed Saadatnejad, Yuejiang Liu, Taylor Mordan, and Alexandre Alahi. Learning decoupled representations for human pose forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2294–2303, 2021. 14

Stefano Pellegrini, Andreas Ess, Konrad Schindler, and Luc Van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *Proceedings of the IEEE International Conference on Computer Vision*, 2009. 3

Mathis Petrovich, Michael J. Black, and Gül Varol. TEMOS: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision (ECCV)*, 2022. 3

Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *Proceedings of the International Conference on Learning Representations International Conference on Machine Learning*, 2015. 3

Andrey Rudenko, Luigi Palmieri, and Kai Oliver Arras. Joint long-term prediction of human motion using a planning-based social force approach. *IEEE International Conference on Robotics and Automation (ICRA)*, 2018. 1

Christian Rupprecht, Iro Laina, Robert DiPietro, Maximilian Baust, Federico Tombari, Nassir Navab, and Gregory D Hager. Learning in an uncertain world: Representing ambiguity through multiple hypotheses. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017. 3

Mandy Ryan and Karen Gerard. Using discrete choice experiments to value health care programmes: current practice and future research reflections. *Applied health economics and health policy*, 2003. 5

Armin Saadat, Nima Fathi, and Saeed Saadatnejad. Towards human pose prediction using the encoder-decoder lstm. 14

Amir Sadeghian, Vineet Kosaraju, Ali Sadeghian, Noriaki Hirose, Hamid Rezatofighi, and Silvio Savarese. Sophie: An attentive gan for predicting paths compliant to social and physical constraints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 3

Tim Salzmann, Boris Ivanovic, Punarjay Chakravarty, and Marco Pavone. Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In *European Conference on Computer Vision*, pp. 683–700. Springer, 2020. 1, 6, 14

Jianhua Sun, Qinhong Jiang, and Cewu Lu. Recursive social behavior graph for trajectory prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 3

Jianhua Sun, Yuxuan Li, Hao-Shu Fang, and Cewu Lu. Three steps to multimodal trajectory prediction: Modality clustering, classification and synthesis. *arXiv preprint arXiv:2103.07854*, 2021. 3

Li-Wu Tsao, Yan-Kai Wang, Hao-Siang Lin, Shuai Hong-Han, Lai-Kuan Wong, and Wen-Huang Cheng. Social-ssl: Self-supervised cross-sequence representation learning based on transformers for multi-agent trajectory prediction. In *European Conference on Computer Vision (ECCV)*, 2022. 3

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017. 16, 17

Anirudh Vemula, Katharina Muelling, and Jean Oh. Social attention: Modeling attention in human crowds. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4601–4607. IEEE, 2018. 1

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 *Proceedings of the European Conference on Computer Vision*, 2018. 18

Jacob Walker, Kenneth Marino, Abhinav Kumar Gupta, and Martial Hebert. The pose knows: Video forecasting by generating pose futures. 2017. 2

Chenxi Wang, Yunfeng Wang, Zixuan Huang, and Zhiwen Chen. Simple baseline for single human motion forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2260–2265, 2021a. 14

Jiashun Wang, Huazhe Xu, Medhini Narasimhan, and Xiaolong Wang. Multi-person 3d motion prediction with multi-range transformers. In *Advances in Neural Information Processing Systems*, 2021b. 1, 2, 3, 6, 7, 8, 9, 14, 15, 16, 17, 18, 19, 20, 21

Chenxin Xu, Maosen Li, Zhenyang Ni, Ya Zhang, and Siheng Chen. Groupnet: Multiscale hypergraph neural networks for trajectory prediction with relational reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6498–6507, 2022a. 14

Sirui Xu, Yu-Xiong Wang, and Liang-Yan Gui. Diverse human motion prediction guided by multi-level spatial-temporal anchors. In *European Conference on Computer Vision (ECCV)*, 2022b. 3, 5

Yi Xu, Dongchun Ren, Mingxia Li, Yuehai Chen, Mingyu Fan, and Huaxia Xia. Tra2tra: Trajectory-to-trajectory prediction with a global social spatial-temporal attentive neural network. *IEEE Robotics and Automation Letters*, 6(2):1574–1581, 2021. 14

Yi Xu, Lichen Wang, Yizhou Wang, and Yun Fu. Adaptive trajectory prediction via transferable gnn. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6520–6531, 2022c. 14

Kota Yamaguchi, Alexander C Berg, Luis E Ortiz, and Tamara L Berg. Who are you with and where are you going? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2011. 3

Xinchen Yan, Akash Rastogi, Ruben Villegas, Kalyan Sunkavalli, Eli Shechtman, Sunil Hadap, Ersin Yumer, and Honglak Lee. Mt-vae: Learning motion transformations to generate multimodal human dynamics. In *Proceedings of the European Conference on Computer Vision*, 2018. 2

Cunjun Yu, Xiao Ma, Jiawei Ren, Haiyu Zhao, and Shuai Yi. Spatio-temporal graph transformer networks for pedestrian trajectory prediction. In *Proceedings of the European Conference on Computer Vision*, pp. 507–523. Springer, 2020. 1

Ye Yuan and Kris Kitani. Diverse trajectory forecasting with determinantal point processes. *arXiv preprint arXiv:1907.04967*, 2019. 3, 6, 18

Ye Yuan and Kris Kitani. Dlow: Diversifying latent flows for diverse human motion prediction. In *Proceedings of the European Conference on Computer Vision*, 2020. 1, 3, 6, 15, 18

Pu Zhang, Wanli Ouyang, Pengfei Zhang, Jianru Xue, and Nanning Zheng. Sr-lstm: State refinement for lstm towards pedestrian trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12085–12094, 2019. 1

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

Yan Zhang, Michael J Black, and Siyu Tang. We are more than our joints: Predicting how 3d bodies move. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 3, 18

Bolei Zhou, Xiaogang Wang, and Xiaoou Tang. Understanding collective crowd behaviors: Learning a mixture model of dynamic pedestrian-agents. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2012. 3

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 the details of the backbone and how it is used 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. 6) 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 3 and Figure 4 in the main paper and more qualitative results in this Appendix (Figure 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 motions, 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. 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) introduces a social pooling module to integrate social information. Adeli et al. (2021) proposes 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) proposes 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.) 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 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.

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.

Method	Local Encoder	Global Encoder
SC-MPF (Adeli et al., 2020)	$\text{GRU}(\mathbf{X_i})$	$\text{MaxPool}(\{\text{GRU}(\mathbf{X_n})\}_{n=1}^{N})$
MRT (Wang et al., 2021b)	$\text{MultiHead}(\mathbf{X_i}, \mathbf{X_i}, \mathbf{X_i})$	$\text{MultiHead}(\{\mathbf{X_n}\}_{n=1}^{N}, \{\mathbf{X_n}\}_{n=1}^{N}, \{\mathbf{X_n}\}_{n=1}^{N})$
XIA (Guo et al., 2022)	N/A	$\{\text{MultiHead}(\mathbf{X_1}, \mathbf{X_2}, \mathbf{X_2}, \text{MultiHead}(\mathbf{X_2}, \mathbf{X_1}, \mathbf{X_1})\}$

C Additional Details of Level-Specific Objectives

We use Δ to represent the residual of the motion sequence. For example, $\Delta\widehat{\mathbf{Y}}_i^j = [\widehat{\mathbf{Y}}_i^j[T_h+1] - \mathbf{X}_i[T_h], \widehat{\mathbf{Y}}_i^j[T_h+2] - \widehat{\mathbf{Y}}_i^j[T_h+1], \ldots, \widehat{\mathbf{Y}}_i^j[T_h+T_p] - \widehat{\mathbf{Y}}_i^j[T_h+T_p-1]]$.

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 $\mathcal{L}_{\text{lR}}$, 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) $\mathcal{L}_{\text{L}}$. Specifically,

$$\mathcal{L}_{\text{lR}} = \frac{1}{N}\sum_{n=1}^{N} \min_{m=1,\ldots,M} \|\Delta\widehat{\mathbf{Y}}_n^m - \Delta\mathbf{Y}_n\|_2^2, \tag{6}$$

$$\mathcal{L}_{\text{L}} = \frac{1}{N*M}\sum_{n=1}^{N}\sum_{m=1}^{M} \|\widehat{\mathbf{L}}_n^m - \mathbf{L}_n\|_2^2, \tag{7}$$

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

A GAN loss (Kocabas et al., 2020) is leveraged to train the model and the local discriminator $\mathcal{D}_l$ for individual body realism. Suppose $\{\mathbf{Y}_n^*\}_{n=1}^N$ is the set of real motion clips sampled from the data, we have the following.

$$\mathcal{L}_{\text{lGAN}} = \frac{1}{N*M}\sum_{n=1}^{N}\sum_{m=1}^{M} \|\mathcal{D}_l(\widehat{\mathbf{Y}}_n^m)\|_2^2 + \frac{1}{N}\sum_{n=1}^{N} \|\mathcal{D}_l(\mathbf{Y}_n^*) - \mathbf{1}\|_2^2. \tag{8}$$

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

$$\{\widetilde{\mathbf{Y}}_i^p\}_{p=1}^P = \{\widetilde{\mathbf{Y}}_i^p | \min_{\mathbf{R},\mathbf{T}} \|\mathbf{R}(\widetilde{\mathbf{X}}_i^p[T_h] - \mathbf{T}) - \mathbf{X}_i[T_h]\|_2 \leq \epsilon\}, \tag{9}$$

$$\mathcal{L}_{\text{mmR}} = \frac{1}{N*P}\sum_{n=1}^{N}\sum_{p=1}^{P} \min_{m=1,\ldots,M} \|\Delta\widehat{\mathbf{Y}}_n^m - \Delta\mathbf{Y}_n^p\|_2^2. \tag{10}$$

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 $\widehat{\mathbf{Y}}_n^m(l)$ and $\widehat{\mathbf{Y}}_n^m(g)$ are the local pose and the global root joint extracted from the global pose $\widehat{\mathbf{Y}}_n^m$, respectively, and α and β are two hyperparameters, this diversity-promoting loss is denoted as

$$\mathcal{L}_{\text{D}} = \frac{1}{N*M(M-1)}\sum_{n=1}^{N}\sum_{m=1}^{M}\sum_{k=m+1}^{M} [\exp(\frac{\|\widehat{\mathbf{Y}}_n^m(g) - \widehat{\mathbf{Y}}_n^k(g)\|_2^2}{\alpha}) + \exp(\frac{\|\widehat{\mathbf{Y}}_n^m(l) - \widehat{\mathbf{Y}}_n^k(l)\|_2^2}{\beta})]. \tag{11}$$

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

$$\mathcal{L}_{\text{gR}} = \min_{m=1,\ldots,M} \frac{1}{N}\sum_{n=1}^{N} \|\Delta\widehat{\mathbf{Y}}_n^m - \Delta\mathbf{Y}_n\|_2^2. \tag{12}$$

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.

Type	Metrics	Definition
Single-Person Fidelity	Local Average Displacement Error (lADE)	$\frac{1}{NT_p}\sum_{n=1}^{N}\min_m \lVert\widehat{\mathbf{Y}}_n^m - \mathbf{Y}_n\rVert_2$
	Local Final Displacement Error (lFDE)	$\frac{1}{N}\sum_{n=1}^{N}\min_m \lVert\widehat{\mathbf{Y}}_n^m[T_p] - \mathbf{Y}_n[T_p]\rVert_2$
	Foot Skating Ratio (FSR)	average ratio of frames where both foot joints are close to the ground ($\leq$ 5cm) and fast ($\geq$ 75mm/s)
Multi-Person Fidelity	(Global) Average Displacement Error (ADE)	$\min_m \frac{1}{NT_p}\sum_{n=1}^{N} \lVert\widehat{\mathbf{Y}}_n^m - \mathbf{Y}_n\rVert_2$
	(Global) Final Displacement Error (FDE)	$\min_m \frac{1}{N}\sum_{n=1}^{N} \lVert\widehat{\mathbf{Y}}_n^m[T_p] - \mathbf{Y}_n[T_p]\rVert_2$
	Trajectory Collision Ratio (TCR)	average ratio of frames where there is a collision between any two trajectories
	Average Human Displacement (AHD)	$\frac{1}{NM}\sum_{n=1}^{N}\sum_{m=1}^{M} \lVert\widehat{\mathbf{Y}}_n^m[T_p] - \widehat{\mathbf{Y}}_n^m[1]\rVert_2$
Overall Diversity	Final Pairwise Distance (FPD)	$\frac{1}{NM(M-1)}\sum_{n=1}^{N}\sum_{m=1}^{M}\sum_{k=m+1}^{M} \lVert\widehat{\mathbf{Y}}_n^m[T_p] - \widehat{\mathbf{Y}}_n^k[T_p]\rVert_2$

The global GAN loss is further leveraged to promote the realism of social motion, where the global discriminator $\mathcal{D}_g$ takes the motion of all N people as input. Suppose $\{\mathbf{Y}_n^{**}\}_{n=1}^N$ is the multi-person motion clip sampled from the data, and we have

$$\mathcal{L}_{\text{gGAN}} = \frac{1}{M}\sum_{m=1}^{M} \lVert\mathcal{D}_g(\{\widehat{\mathbf{Y}}_n^m\}_{n=1}^N)\rVert_2^2 + \lVert\mathcal{D}_g(\{\mathbf{Y}_n^{**}\}_{n=1}^N) - \mathbf{1}\rVert_2^2. \tag{13}$$

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

Method, # of Intents	@$t = 1s$		@$t = 2s$		@$t = 3s$	
	lADE ↓	lFDE ↓	lADE ↓	lFDE ↓	lADE ↓	lFDE ↓
DuMMF (Ours), 2	0.663 ± 0.007	1.088 ± 0.019	1.029 ± 0.019	1.612 ± 0.043	1.308 ± 0.031	2.014 ± 0.086
DuMMF (Ours), 3	0.656 ± 0.007	1.056 ± 0.007	0.996 ± 0.009	1.503 ± 0.032	1.244 ± 0.021	1.883 ± 0.076
DuMMF (Ours), 5	0.648 ± 0.011	1.033 ± 0.022	0.963 ± 0.020	1.400 ± 0.035	1.176 ± 0.025	1.672 ± 0.052

D Additional Details of Network Architecture

Below we detail the main multi-person predictor MRT (Wang et al., 2021b) we use and how it integrates with our DuMMF framework for the proposed task.

Encoder. The encoder consists of a local-range transformer encoder and a global-range transformer encoder. The local encoder takes historical motion residuals $\Delta\mathbf{X}_i$ of a single person as input and extracts a feature for each person. The global encoder takes historical motions $\{\mathbf{X}_n\}_{n=1}^N$ of all persons as input and runs only once to obtain global features.

Decoder. To generate the future motion of the i-th person, we concatenate the local features of the i-th person with the global features as the final encodings, which are used as keys and values for the transformer-based decoder. The decoder takes the last pose of the historical motion $\mathbf{X}_i[T_h]$ and the sampled intent codes $\{\mathbf{z}_n^c + \mathbf{z}_n^d\}_{n=1}^N$ as the query. Note that Wang et al. (2021b) only uses the encoder-decoder attention. On the contrary, we introduce the decoder-decoder attention (Vaswani et al., 2017) to further incorporate motion and intent.

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 $\mathbf{Y}_i^j$ with the feet velocities $\Delta\mathbf{F}_i^j$ as input to a local discriminator $\mathcal{D}_l$. The local discriminator uses an architecture similar to the local encoder, with additional fully connected layers to output a fidelity score for each motion.

Global Discriminator. To ensure the realism of social interaction and avoid motion collisions, we propose a global discriminator $\mathcal{D}_g$ that encodes all motions of N people $\{\mathbf{Y}_n^j\}_{n=1}^N$ at the same time and outputs a fidelity score. The global discriminator uses an architecture similar to the global encoder, with additional fully connected layers.

Incorporate of DuMMF and the Transformer. Our dual-level modeling framework operates primarily in the decoding stage. Specifically, we get the key $\mathbf{K}$ and the value $\mathbf{V}$ of the transformer-based (Vaswani et al., 2017) decoder, from the same concatenated encodings $\mathbf{H}$ obtained by the local and global encoders. The query $\mathbf{Q}$ is from the end pose (at time step T_h) $\mathbf{X}_i[T_h]$ of i-th person's historical sequence $\mathbf{X}_i$. Based on the dual-level modeling mechanism discussed in Sec. 3.1 of the main paper, we use the i-th intent code $\mathbf{z}_i^d$ from the discrete set $\{\mathbf{Z}^m\}_{m=1}^M$ and the continuous intent code $\mathbf{z}_i^c$ to generate the i-th prediction, leading to a total of M predictions. The discrete and continuous intent codes are combined to get the key, value, and query as follows:

$$\mathbf{H}_i = (\mathbf{z}_i^c + \mathbf{z}_i^d)^{\mathrm{T}}\mathbf{W}_z, \tag{14}$$

where $\mathbf{W}_z$ is a linear layer. We then apply linear layers $\mathbf{W}_K, \mathbf{W}_V, \mathbf{W}_Q$ as follows:

$$\mathbf{K}_i = \mathbf{H}_i^{\mathrm{T}}\mathbf{W}_K,\ \mathbf{V}_i = \mathbf{H}_i^{\mathrm{T}}\mathbf{W}_V,\ \mathbf{Q}_i = [\mathbf{X}_i[T_h], \mathbf{H}_i]^{\mathrm{T}}\mathbf{W}_Q. \tag{15}$$

Finally, L stacked standard transformer decoder layers are applied and we get the predicted motion of the i-th person. We also follow Wang et al. (2021b) to adopt Discrete Cosine Transform (DCT) to the inputs and Inverse Discrete Cosine Transform (IDCT) to the outputs. Please refer to Wang et al. (2021b) for more details.

E Additional Implementation Details

Here, we provide more details on the implementation of our method DuMMF. The two hyperparameters (α, β) 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 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.

Architecture	Predictor	Diversifier	Variants	@$t=1s$			@$t=2s$			@$t=3s$		
				FPD ↑	ADE ↓	FDE ↓	FPD ↑	ADE ↓	FDE ↓	FPD ↑	ADE ↓	FDE ↓
RNN	SC-MPF (Adeli et al., 2020)	Deterministic	N/A	N/A	0.514	0.880	N/A	0.902	1.632	N/A	1.319	2.594
		CGAN		0.467	0.473	0.786	0.846	0.753	1.207	1.215	0.967	1.518
		CGAN+DuMMF (Ours)	w/o Separation	0.511	0.481	0.801	0.987	0.772	1.252	1.471	1.001	1.600
			w/o Discrete	0.697	0.478	0.793	1.534	0.770	1.258	2.800	1.008	1.618
			w/o Continuous	0.736	0.627	1.092	1.412	1.074	1.860	2.262	1.476	2.588
			Full	**0.747**	**0.455**	**0.744**	**1.441**	**0.707**	**1.103**	**2.200**	**0.893**	**1.364**
Transformer	SC-MPF (Adeli et al., 2020)	Deterministic	N/A	N/A	0.430	0.731	N/A	0.723	1.258	N/A	0.995	1.781
		CGAN		0.504	0.403	**0.690**	1.099	0.659	1.077	1.834	0.857	1.371
		CGAN+DuMMF (Ours)	w/o Separation	0.787	0.407	0.701	1.791	0.667	1.080	3.047	0.865	1.363
			w/o Discrete	0.557	0.403	0.693	1.202	0.657	1.075	1.975	0.857	1.382
			w/o Continuous	**0.917**	0.406	0.707	**2.124**	0.685	1.163	**3.535**	0.924	1.573
			Full	0.865	**0.402**	0.691	1.816	**0.656**	**1.071**	2.902	**0.852**	**1.352**
	MRT (Wang et al., 2021b)	Deterministic	N/A	N/A	0.427	0.747	N/A	0.747	1.354	N/A	1.071	2.043
		CGAN		0.176	0.421	0.741	0.452	0.729	1.288	0.878	1.017	1.846
		CGAN+DuMMF (Ours)	w/o Separation	0.194	0.423	0.743	0.456	0.731	1.267	0.855	1.000	1.754
			w/o Discrete	0.133	0.421	0.743	0.335	0.732	1.286	0.627	1.012	1.805
			w/o Continuous	0.308	0.418	0.730	0.732	0.716	1.242	1.304	0.982	1.734
			Full	**0.510**	**0.408**	**0.711**	**1.075**	**0.683**	**1.135**	**1.740**	**0.905**	**1.491**
Transformer	MRT (Wang et al., 2021b)	Deterministic	N/A	N/A	0.475	0.822	N/A	0.853	1.605	N/A	1.268	2.539
		CGAN		0.556	0.457	0.791	1.123	0.770	1.300	1.778	1.041	1.784
		CGAN+DuMMF (Ours)	w/o Separation	0.669	0.441	0.752	1.292	0.713	1.146	1.939	0.922	1.469
			w/o Discrete	0.100	0.475	0.846	0.231	0.843	1.523	0.402	1.192	2.207
			w/o Continuous	0.691	**0.413**	**0.710**	1.462	**0.674**	1.109	2.338	0.885	1.433
			Full	**1.133**	0.432	0.734	**2.283**	0.694	**1.104**	**3.575**	**0.891**	**1.384**
	XIA (Guo et al., 2022)	Deterministic	N/A	N/A	0.478	0.840	N/A	0.848	1.555	N/A	1.223	2.338
		CGAN		0.489	0.434	0.743	1.022	0.706	1.145	1.611	0.918	1.474
		CGAN+DuMMF (Ours)	w/o Separation	0.479	0.435	0.741	0.902	0.705	1.140	1.310	0.913	1.464
			w/o Discrete	0.451	**0.428**	0.742	1.065	0.710	1.192	1.804	0.947	1.586
			w/o Continuous	0.215	0.447	0.784	0.440	0.768	1.335	0.693	1.052	1.853
			Full	**0.895**	0.431	**0.731**	**1.902**	**0.690**	**1.092**	**3.057**	**0.879**	**1.333**

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**

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.

Method, # of Intents	Prediction Time				
	100 ms	240 ms	500 ms	640 ms	900 ms
SC-MPF* (Adeli et al., 2020), N/A	46.28	73.88	130.23	160.83	208.44
TRiPOD* (Adeli et al., 2021), N/A	30.26	51.84	85.08	104.78	146.33
MRT (Wang et al., 2021b), N/A	**22.93**	**42.35**	79.41	99.02	137.93

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.

Architecture	Predictor	Diversifier	Variants	$@t = 1s$		
				FPD ↑	ADE ↓	FDE ↓
Transformer	MRT (Wang et al., 2021b)	Deterministic	N/A	N/A	0.575	0.961
		CGAN		0.501	0.582	0.983
		CGAN+DuMMF (Ours)	w/o Separation	0.864	0.590	0.969
			w/o Discrete	**1.075**	0.567	0.957
			w/o Continuous	1.063	0.587	1.018
			Full	0.969	**0.564**	**0.935**

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 thee 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**.

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 lADE) and FDE (or lFDE) only measure the best predictions in all outputs. While we can *average* ADE (or lADE) and FDE (or lFDE) 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 ℓ_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. In Table C, we use lADE and lFDE to compare single-person fidelity, and we observe that our model also significantly outperforms the baseline.

Additional Comparisons on SoMoF Benchmark. We investigate the performance of our framework on the Social Movement Prediction Challenge (SoMoF) (Adeli et al., 2020; 2021). We specifically use the 3DPW dataset (von Marcard et al., 2018), where we use labeled trajectories and poses (13-joint human skeleton), but we do not use videos of given scenes as input. We discard the multi-modal reconstruction loss since the 3DPW provided by SoMoF benchmark is relatively small. In Table E, we provide results on our implemented

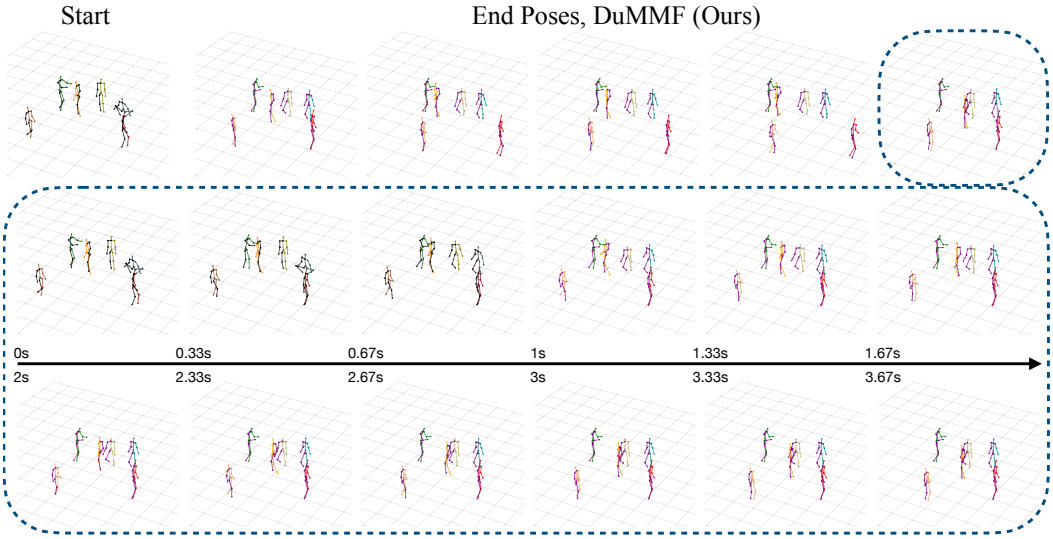

Figure A: Qualitative results on CMU-Mocap. We evaluate the generalizability and scalability of our model to predict 3-second motion on the constructed **6**-person motion test data. The top row is the historical pose, and the five end poses predicted by our model; and we show the sequence of one of the predictions in the bottom two rows (highlighted by the blue dashed box).

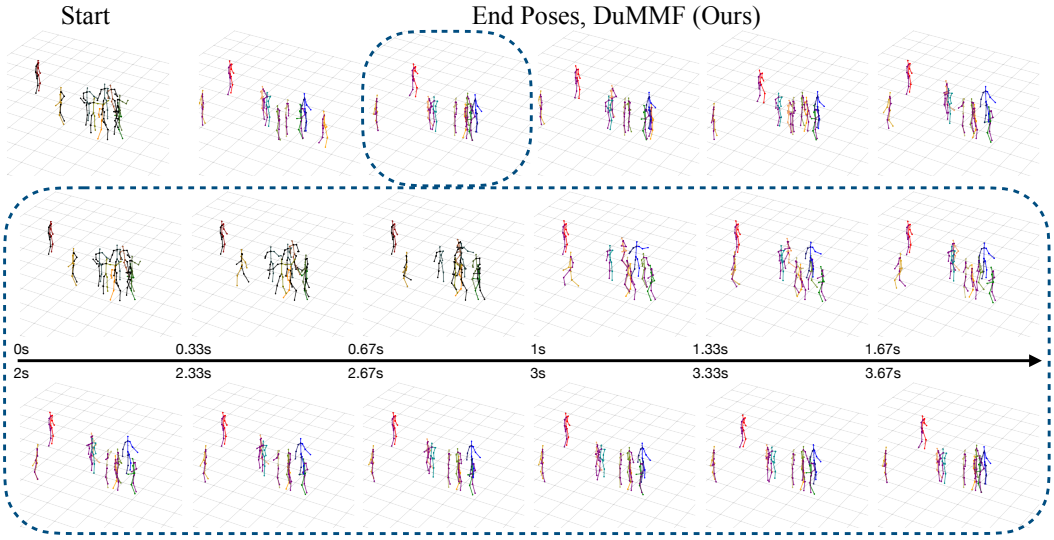

Figure B: Qualitative results on CMU-Mocap. We evaluate the generalizability and scalability of our model to predict 3-second motion on the constructed **9**-person motion test data. The top row is the historical pose, and the five end poses predicted by our model; and we show the sequence of one of the predictions in the bottom two rows (highlighted by the blue dashed box).

deterministic prediction and compare with baselines directly reported from the leaderboard on the SoMoF benchmark. In Table F, we use SoMoF benchmark for stochastic multi-person forecasting. Note that ADE and FDE require access to ground truth data, which is not publicly available from SoMoF benchmark. Thus, we report the results on validation set. We observe a similar performance as on CMU-Mocap (CMU) and MuPoTS-3D (Mehta et al., 2018), specifically, using DuMMF, the model has significantly better accuracy and diversity in stochastic multi-person forecasting.

Additional Comparison with Single-Person Forecasting Methods. In Table G, 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.,

Table G: 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.

Architecture	Predictor	Diversifier	Variants	@$t = 1s$			@$t = 2s$			@$t = 3s$		
				FPD ↑	ADE ↓	FDE ↓	FPD ↑	ADE ↓	FDE ↓	FPD ↑	ADE ↓	FDE ↓
Transformer	SRT	CGAN	N/A	N/A	0.681	1.125	N/A	1.082	1.765	N/A	1.427	2.438
	MRT (Wang et al., 2021b)			0.282	0.662	1.086	0.662	1.023	1.567	1.199	1.287	1.968
	SRT	**CGAN+DuMMF (Ours)**	w/o Separation	0.291	0.677	1.110	0.676	1.056	1.624	1.166	1.328	2.021
	MRT (Wang et al., 2021b)		Full	**0.716**	**0.658**	**1.053**	**1.435**	**0.993**	**1.472**	**2.206**	**1.232**	**1.823**

Table H: Ablation study of our DuMMF model on CMU-Mocap. We report the accuracy and diversity with error bar for root and pose respectively. The results show the impact of different learning objectives. Best results are bolded, and next best results are underlined.

$\mathcal{L}_{\text{lR}}$	$\mathcal{L}_{\text{L}}$	$\mathcal{L}_{\text{lGAN}}$	$\mathcal{L}_{\text{mmR}}$	$\mathcal{L}_{\text{D}}$	$\mathcal{L}_{\text{gGAN}}$	rootFPD ↑	poseFPD ↑	rootADE ↓	rootFDE ↓	poseADE ↓	poseFDE ↓
	✓	✓	✓	✓	✓	$\underline{0.251 \pm 0.025}$	$\mathbf{0.541 \pm 0.047}$	0.748 ± 0.010	0.919 ± 0.017	0.247 ± 0.007	0.401 ± 0.013
✓		✓	✓	✓	✓	0.242 ± 0.069	0.437 ± 0.146	0.784 ± 0.029	0.983 ± 0.075	0.274 ± 0.015	0.459 ± 0.030
✓	✓		✓	✓	✓	0.241 ± 0.026	0.536 ± 0.055	$\mathbf{0.735 \pm 0.005}$	$\mathbf{0.897 \pm 0.005}$	0.250 ± 0.004	0.407 ± 0.014
✓	✓	✓		✓	✓	0.218 ± 0.074	0.468 ± 0.186	0.749 ± 0.011	0.920 ± 0.017	0.254 ± 0.012	0.418 ± 0.038
✓	✓	✓	✓		✓	0.240 ± 0.032	0.537 ± 0.067	$\underline{0.738 \pm 0.008}$	0.900 ± 0.010	$\mathbf{0.241 \pm 0.010}$	$\underline{0.393 \pm 0.021}$
✓	✓	✓	✓	✓		0.232 ± 0.058	0.506 ± 0.111	0.740 ± 0.004	0.905 ± 0.012	0.247 ± 0.008	0.402 ± 0.028
✓	✓	✓	✓	✓	✓	$\mathbf{0.254 \pm 0.037}$	$\underline{0.538 \pm 0.071}$	$\underline{0.738 \pm 0.007}$	$\mathbf{0.897 \pm 0.010}$	$\underline{0.245 \pm 0.008}$	$\mathbf{0.391 \pm 0.015}$

2021b), where we modified the encoder and decoder to be individually independent. Table G illustrates that (1) the modeling of social interaction is crucial, since the single-person forecasting baseline has much worse performance; (2) the improvement from our DuMMF for the multi-person forecasting model is much higher than that for the single-person forecasting model.

Additional Ablation on Impact of Learning Objectives. In Table H, 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 $\mathcal{L}_{lR}$ and $\mathcal{L}_{gR}$ 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 $\mathcal{L}_L$ 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 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 I, 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.

Table I: Ablation study of the number of discrete latent codes. 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.

# of Intents	@$t = 1s$			@$t = 2s$			@$t = 3s$		
	FPD ↑	ADE ↓	FDE ↓	FPD ↑	ADE ↓	FDE ↓	FPD ↑	ADE ↓	FDE ↓
5	**0.716**	**0.658**	**1.053**	**1.435**	**0.993**	**1.472**	**2.206**	**1.232**	**1.823**
6	0.498	0.667	1.087	1.086	1.033	1.578	1.814	1.300	1.971
7	0.465	0.671	1.093	0.960	1.032	1.582	1.502	1.298	1.986
9	0.263	0.669	1.087	0.565	1.027	1.578	0.941	1.295	2.028
15	0.231	0.678	1.127	0.532	1.076	1.704	0.966	1.384	2.209

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 significantly increased number of people – up to 6 or 9 people; this also increases the complexity of social interactions. Specifically, we first sample 3-person motion data from the original 3-person test data in CMU-Mocap (CMU), and then compose them together to generate 6-person or 9-person motion data. We use handcraft rules (Kothari et al., 2020) to filter out scenes with trajectory collisions. Instead of retraining the model in the novel setting with more people per scene (Wang et al., 2021b), 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. Note that here we use MRT (Wang et al., 2021b) as the backbone of DuMMF framework.

Qualitative Results. Similar to Sec. 4 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 6-person or 9-person data, but it still performs well on more-person data, suggesting the *generalizability* and *scalability* of our approach.