Computers & Graphics 106 (2022) 130–140

Contents lists available at ScienceDirect

Computers & Graphics

journal homepage: www.elsevier.com/locate/cag

Special Section on MIG 2021

A2X: An end-to-end framework for assessing agent and environment interactions in multimodal human trajectory prediction

Samuel S. Sohn [a,*], Mihee Lee [a], Seonghyeon Moon [a], Gang Qiao [a], Muhammad Usman [c], Sejong Yoon [b], Vladimir Pavlovic [a], Mubbasir Kapadia [a]

[a] Rutgers University, USA
[b] The College of New Jersey, USA
[c] King Fahd University of Petroleum and Minerals, Saudi Arabia

ARTICLE INFO

Article history:
Received 19 October 2021
Received in revised form 28 April 2022
Accepted 23 May 2022
Available online 7 June 2022

Keywords:
Human trajectory prediction
Datasets
Evaluation metrics

ABSTRACT

In recent years, human trajectory prediction (HTP) has garnered attention in computer vision literature. Although this task has much in common with the longstanding task of crowd simulation, there is little from crowd simulation that has been borrowed, especially in terms of evaluation protocols. The key difference between the two tasks is that HTP is concerned with forecasting multiple steps at a time and capturing the multimodality of real human trajectories. A majority of HTP models are trained on the same few datasets, which feature small, transient interactions between real people and little to no interaction between people and the environment. Unsurprisingly, when tested on crowd egress scenarios, these models produce erroneous trajectories that accelerate too quickly and collide too frequently, but the metrics used in HTP literature cannot convey these particular issues. To address these challenges, we propose (1) the A2X dataset, which has simulated crowd egress and complex navigation scenarios that compensate for the lack of agent-to-environment interaction in existing real datasets, (2) evaluation metrics that convey model performance with more reliability and nuance, and (3) a guideline for future data acquisition in HTP. A subset of the proposed metrics are novel *multiverse metrics*, which are better suited for multimodal models than existing metrics. The dataset is available at: https://mubbasir.github.io/HTP-benchmark.

1. Introduction

The study of human navigation has long been of interest to various research communities such as computer graphics [1], computer vision [2], cognitive science [3], and robotics [4]. Advancements in these areas have seen widespread practical application in pandemic response, architectural design, urban planning, transportation engineering, crowd management, socially compliant robot navigation, and entertainment. Accordingly, the influence of human navigation research has reached countless individuals and will continue to do so in the foreseeable future.

Most applications rely on simulation models [5], which are sufficiently accurate to human behavior and generalizable to unforeseen circumstances. However, the past five years of predictive modeling in computer vision has achieved significantly better accuracy [6], giving it a strong potential to overtake the longstanding models from computer graphics. This is largely due to the transition from using unimodal, discriminative models [2] that predict a single future trajectory to using multimodal, generative models [7–9] that predict a distribution of future trajectories, which captures the inherent uncertainty in human decision-making [10,11]. Despite the evolution of models, however, the accuracy metrics that were introduced with the first unimodal models are still in use today. In order to adapt these fundamentally unimodal metrics to multimodal models, the metrics are computed between each predicted trajectory and the ground truth trajectory, and the minimum error for each metric is reported. This results in a gross overestimation of accuracy that we later show is not consistent with the expected accuracy, which may misguide future research efforts. Furthermore, the minimum value is not actionable, because while it is evident that a state-of-the-art (SOTA) multimodal model can find *an* accurate trajectory, it cannot determine *which* trajectory is most accurate for unseen data. We measure this uncertainty through a decidability metric.

* Corresponding author.
E-mail addresses: samuel.sohn@rutgers.edu (S.S. Sohn), ml1323@rutgers.edu (M. Lee), sm2062@cs.rutgers.edu (S. Moon), gq19@cs.rutgers.edu (G. Qiao), muhammad.usman@kfupm.edu.sa (M. Usman), yoons@tcnj.edu (S. Yoon), vladimir@cs.rutgers.edu (V. Pavlovic), mubbasir.kapadia@rutgers.edu (M. Kapadia).

https://doi.org/10.1016/j.cag.2022.05.010
0097-8493/

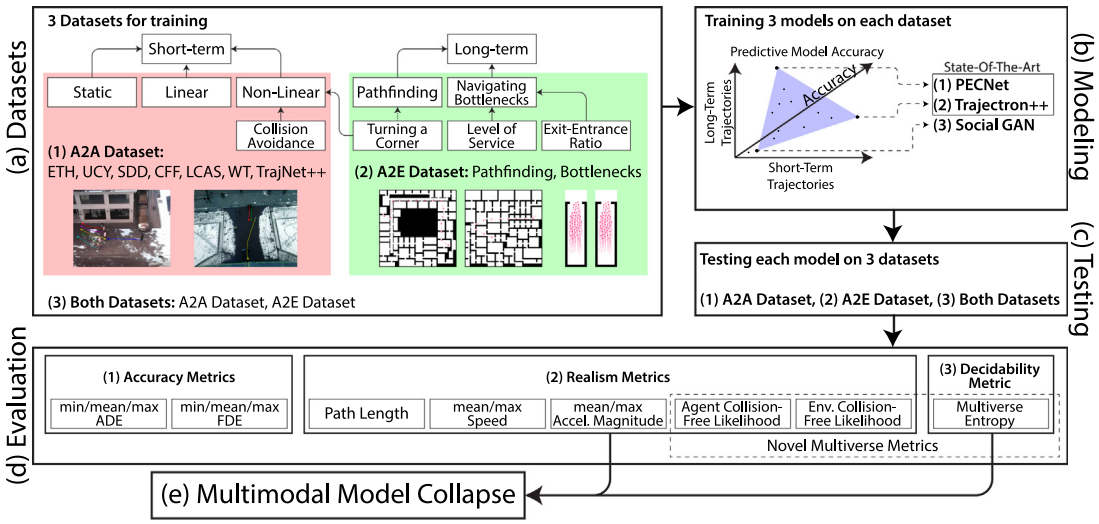

Fig. 1. The above framework image shows (a) the differences between the trajectories of existing datasets (A2A) and the novel dataset (A2E), (b–c) the models trained and tested on combinations of A2A and A2E, (d) the proposed set of metrics for evaluating the accuracy, realism, and decidability of models, and (e) a greedy method for selecting the prediction most realistic movement.

Generalizability cannot be maximized by solely improving upon accuracy metrics. An inaccurate model can be robust by producing realistic trajectories, and an accurate model can fail to be practicable by being undecidable. Models can exist on the continuum between these two extremes, making it critical to consider realism and decidability metrics as well.

Furthermore, there is a stark class imbalance in existing datasets. While datasets are abundant in instances where humans are interacting with each other in open spaces [12–17], they are significantly lacking in both environment information and instances where humans are interacting with their environment. Ultimately, this hinders generalization at a global level and has led to some models being developed without considering environments at all [2,7].

In this work, we provide an augmented human trajectory prediction dataset that compensates for the lack of agent-to-environment interaction in existing datasets with a new simulated dataset. To understand model performance on this new dataset with more reliability and nuance, we propose a comprehensive set of accuracy, realism, and decidability metrics. A subset of these metrics are novel *multiverse metrics*, which are better-suited for multimodal models than existing metrics but are still applicable to unimodal models. The evaluation using these metrics decisively evidences that the new dataset facilitates better robustness and generalization, that current metrics can be misleading, and that there are still remaining challenges to modeling human trajectories. We finally showcase that realism metrics can also be used to decide which prediction to take from an undecidable multimodal model through the process of *Multimodal Model Collapse*. Henceforth, we refer to humans as agents, since our conceptual framework is broadly applicable, e.g. to robotic and vehicular agents.

2. Background

2.1. Models for human trajectory prediction

While crowd simulation has been well-studied in computer graphics literature [18,19], we focus on the use of machine learning techniques for the growing field of human trajectory prediction. Earlier methods such as Social LSTM [2] and Social Attention [20] proposed deterministic models which predict a single future trajectory per agent given the observed trajectories. These unimodal approaches are limited in their ability to represent the inherent uncertainty in an agent's future. Many later studies [7–9,21–24] have assumed the multimodality of future human behavior and learn its distribution to capture the uncertainty. In this paper, we leverage three state-of-the-art (SOTA) methodologies to demonstrate our dataset and evaluation metrics: SocialGAN [7], PECNet [9], and Trajectron++ [8].

SocialGAN [7] adopts the GAN [25] framework to forecast multiple possible future trajectories. The generator creates samples similar to the data distribution while the discriminator distinguishes whether the samples belong to the ground-truth or the generated data. SocialGAN also tackles the problem of potential collisions between the agents in a scenario by introducing a global pooling mechanism. The pooling of all agents' features allows SocialGAN to capture the interactions between agents, which in theory prevents collisions between neighboring agents.

PECNet [9] addresses human trajectory prediction in two steps. First, it predicts a future goal position based on the observed trajectories by modeling the distribution of the goal positions with a Variational Autoencoder (VAE) [26]. After sampling from this learned distribution, PECNet predicts each step of an agent's future trajectory by interpolating between the observed trajectory and the estimated future goal position.

Trajectron++ [8] proposes a graph-structured recurrent model based on a conditional VAE [27] to predict future trajectories. During training, it encodes both past and future trajectories to obtain the latent factor z from the posterior distribution, while during inference, it is sampled from the prior distribution based only on the past trajectories. Trajectron++ then leverages the graph structure by using edge encodings to model the interaction between the nodes (i.e., agents) in a scene. To model the interaction between agents and the environment, it encodes a local map to avoid obstacle collisions.

We investigate these three models as the representatives of various SOTA works. PECNet [9] exhibits outstanding performance for long-term trajectories while Trajectron++ [8] is highly performant for short-term trajectories. Meanwhile, SocialGAN [7] is one of the earliest and most frequently referred models. Together, these models envelop numerous existing models in terms of both short- and long-term human trajectory prediction accuracy, which Fig. 1.b illustrates. We differentiate between predictive models of short-term and long-term trajectories on the

basis of goal conditioning. A model that is not goal-conditioned will inherently increase in error as the predicted path length increases, sometimes at an exponential rate [8], whereas goal-conditioned models are expected to predict long paths without the same trade-off between path length and error.

2.2. Datasets for human trajectory prediction

Several human trajectory prediction datasets have been collected by the computer vision and computer graphics communities.

ETH [28] and UCY [12] are the most commonly used datasets, which feature five outdoor scenes and over 1600 total trajectories recorded at 2.5 Hz. These datasets contain instances of collision avoidance and group movement, but the recorded trajectories are relatively short due to a small viewing window.

Stanford drone dataset (SDD) [13] consists of eight outdoor scenes in Stanford campus collected from a drone. The dataset contains more than 19,000 targets including not only pedestrians, but also bicyclists, skateboarders, cars and buses. The coordinates of the trajectories are in the image coordinate system from the bird's eye view, instead of physical word coordinate system. In contrast to ETH/UCY, the larger viewing window results in longer recorded trajectories.

Stanford crowd dataset (CFF) [14] consists of pedestrian trajectories collected within a train station building of size 25 m $\times$ 100 m by a set of distributed cameras. Unlike the outdoor environments of ETH/UCY and SDD, the indoor environment of CFF poses stricter constraints on pedestrian motion and the mass transportation setting features significantly higher pedestrian density than ETH/UCY and SDD. However, the dataset is quite noisy due to detection, tracking, and localization error, and the difficulty to measure the accurate positions of the non-navigable areas.

L-CAS 3D Point Cloud People Dataset (LCAS) [15] consists of 28,002 scan frames collected within a university building by a 3D LiDAR mounted on a robot that is either stationary or moving. A scan frame contains around 30,000 3D points, based on which pedestrians are labeled with 3D bounding boxes and marked as either visible or partially visible. While both CFF and LCAS consider indoor environments, LCAS is characterized by low pedestrian density. Both low and high densities are essential to consider, because collective behaviors are known to emerge at high pedestrian densities, but not at low densities [29].

WILDTRACK (WT) [16] was collected with seven static HD cameras in a public square and captured dense groups of pedestrians for approximately 60 min. The seven cameras' fields of view in large part overlap, allowing precise joint calibrations of image sequences, which ensure high-precision trajectory data. Unlike other datasets in which pedestrians are primarily focused on navigation, WT features groups of pedestrians that are occasionally static while engaged in conversation.

Some datasets, such as TrajNet++ [17] and OpenTraj [30] are composed of multiple constituent datasets. TrajNet++ combines the ETH/UCY, CFF, LCAS, and WT datasets, as well as a synthetic dataset generated by the ORCA model [31].

Existing human trajectory datasets have limitations in the sense of embodying interactions. They either do not contain agent-to-environment (A2E) interactions [16], or exhibit limited agent-to-agent (A2A) interactions at small scales in simplistic environments. We speculate that many self-centered pedestrians are prone to avoid or mitigate, consciously or unconsciously, the influence of the environments and other pedestrians during their navigation. In this work, we are proposing datasets that augment A2E and A2A interactions, which may bring benefits for enhancing learning models by encoding more complex trajectory dynamics.

2.3. Evaluation for human trajectory prediction

In computer graphics literature, trajectories are generally measured by motion statistics such as the number of collisions, average speed, average acceleration, and total distance traveled [32]. On the other hand, in machine learning literature [2,7,17], the most commonly used evaluation metrics for trajectory forecasting models are Average Displacement Error (ADE) and Final Displacement Error (FDE). ADE is the average L_2 distance between the ground truth and the predicted trajectories across all future time steps. FDE is the L_2 distance between the final positions of the ground truth and the predicted trajectories.

However, many trajectory forecasting models assume multi-modality in the future behavior, which makes their models generate more than one prediction of the future trajectory given one past trajectory. The current strategy used in the prior works is reporting the minimum ADE/FDE results across randomly sampled k predictions where $k = 20$ in most cases.

In order to evaluate the multi-modal models, Trajectron [22] introduces Negative LogLikelihood (NLL) which is used also in [8, 17]. Given a future time step to predict, they compute the average NLL of the ground truth trajectory under a distribution generated by a kernel density estimate on trajectory prediction samples.

Trajnet++ [17] tackles the issue of various human trajectory prediction models demonstrating their methods on different subsets of benchmark datasets. To evaluate them on the same set of trajectory data, Trajnet++ introduces their own benchmark. Trajnet++ is especially focused on generating data with sufficient human interaction in order to evaluate the capacity for each model to predict plausible trajectories without collisions with other pedestrians. To measure the collisions, they suggest new metrics; Collision1 and Collision2. Collision1 computes the collision rate between a pedestrian's predicted trajectory and its neighbors' predicted trajectories. Collision2 computes the collision rate between a pedestrian's predicted trajectory and its neighbors' ground truth trajectories.

ADE and FDE are applicable to unimodal methods which predict only one future sequence that can be compared with the ground truth future sequence. However, as aforementioned in this section, many multimodal trajectory forecasting models assuming uncertainty and multimodality in pedestrians' future behaviors predict k future sequences (usually $k = 20$). Most of these models report the minimum ADE/FDE results among all k predictions, which, in our view, is overly optimistic. Not only is this a significant underestimation of the error, but it is also an impossible standard in that these models are incapable of choosing the prediction with the minimum error. In Section 4 of this work, we propose new metrics that can tackle this issue.

3. Agent-to-agent and agent-to-environment interaction dataset

We propose a comprehensive trajectory prediction dataset **A2X** that consists of a representative set of trajectories, which will enable better generalization under realistic circumstances that are either complex or unsafe and out-of-distribution (OOD) with respect to current datasets.

In order to understand what the shortcomings of current datasets are (Section 2), we first taxonomize the characteristics of human trajectories. The TrajNet++ benchmark [17] proposed an initial taxonomy that only considers short-term characteristics, e.g., standing still, moving linearly, or avoiding collisions (Fig. 1.a). While the original taxonomy is sufficient for describing the trajectories in many real datasets and their agent-to-agent (A2A) interactions, models that learn exclusively from these types are insufficient for most applications, which consider environments

with obstacles and time frames longer than 5 s, which is the practical limit for most models before they become exponentially erroneous [8]. We have improved upon this by considering long-term characteristics (Fig. 1.a), i.e., pathfinding alone and navigating through crowded bottlenecks. These types of trajectories emerge from agent-to-environment (A2E) interactions, which unfold over a longer time frame than A2A interactions and are essential for navigation within any environment [33].

3.1. Agent-to-agent interactions

For representing A2A interactions, we make use of each prior dataset described in Section 2.2: ETH [28], UCY [12], SDD [13], CFF [14], LCAS [15], WT [16], and TrajNet++ [17]. These datasets feature transient interactions between agents and little interaction with the environment, which is made difficult to measure by the frequent unavailability of environment information. Therefore, we approximate environment information based on the principle of stigmergy [34,35], which observes the self-organization of human navigation along trails. For each position that agents have traveled through in either the training or testing sets of the ground truth, a 1-meter radius around the position is considered to be navigable. This guarantees that predictions with less than 1 meter of displacement from the ground truth at all times will never intersect with the environment. Although this technique has been applied to all datasets for consistency, we recommend using environment information from datasets whenever possible. Additionally, in order to compensate for the imbalance between A2A and A2E interactions in prior datasets, we propose the generation of synthetic data in addition to that of TrajNet++. While real datasets are valuable for their veridicality, there are logistical limitations that prevent the acquisition of real data in OOD scenarios that are unsafe for human participants or prohibitively expensive from an organizational standpoint.

3.2. Agent-to-environment interactions

Two such scenarios are used to sample trajectories exhibiting A2E interactions: (1) pathfinding alone in a large, complex environment, which has prohibitive logistical cost and (2) navigating through bottlenecks of varied width with a dense crowd, which can be unsafe. Though simulation models are normally less accurate than predictive models in predicting human trajectories [2], the prevalent Social Force model [1] currently outperforms predictive models in terms of robustness, has been used in several application domains [4,36,37], and has adequate ecological validity in these A2E scenarios, which lack sufficient real data for training predictive models until **A2X**. Although there are differences between synthetic and real data, the Social Force Model operates on a much simpler set of dynamics to navigate than real humans regardless of its parameterization. In fact, all visual [38], auditory [39], and olfactory [40] stimuli used by humans for navigation are abstracted away in HTP datasets. Therefore, synthetic data makes it uniquely feasible to learn a perfect model using trajectory data alone. However, in light of the differences, we recommend that during analysis, A2A-trained predictive models are primarily evaluated on A2E not with accuracy metrics, but with realism and decidability metrics instead.

We leverage the Social Force model to simulate 236 scenarios of a single agent navigating between random points in complex 112×112 m^2 environments from [33] (Fig. 2). This produces long-term isolated interactions between single agents and the environment. We then use the same model to simulate well-studied bottleneck scenarios [41,42] in a 25×7 m^2 room that vary in terms of (a) the density of agents (Level of Service) from $\{0.2, 0.4, 0.6, 0.8, 1.0\}$ agents/m^2 and (b) the ratio between the width of the bottleneck and the width of the room (Exit–Entrance Ratio) from $\{0.2, 0.3, 0.4, 0.6, 0.7\}$ (Fig. 2). A total of 398 scenarios have been generated across all combinations of Level of Service and Exit–Entrance Ratio. This produces long-term interactions between agents as a result of the constricting environment. Exact environment information has been provided for both types of scenarios. We later show that current models trained on existing A2A datasets are unable to generalize to these critical scenarios, but with the addition of training data on these scenarios, the accuracy of predictions significantly improves.

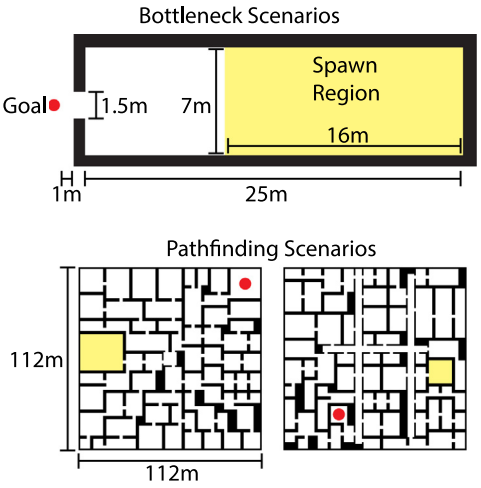

Fig. 2. The above images show the exact dimensions of environments from the bottleneck and pathfinding scenarios in A2E.

4. Accuracy, realism, and decidability of human trajectory prediction

We propose a total of 15 accuracy, realism, and decidability metrics (Fig. 1.d). These metrics are either borrowed from computer vision and computer graphics literature [2,28,32,43] or newly developed *multiverse metrics*, which assess the A2A and A2E interactions of both multimodal models with $k > 1$ and unimodal models with $k = 1$.

4.1. Preliminaries

In accordance with both unimodal and multimodal predictive models, we utilize the following notation for their predictions. A prediction scenario is defined by a set of n agents present in an environment $\mathbf{E}$ at the same time. Each agent a has t_p frames of past position data as input and t_f frames of future position data for ground truth $\mathbf{Y}_{a,0} \in \mathbb{R}^{t_f \times 2}$ and for each prediction $\widehat{\mathbf{Y}}_{a,j} \in \mathbb{R}^{t_f \times 2}$, where $0 \leq j < k$. All position data is in meters and has a frame rate of $1/\Delta t$ Hz based on the dataset. The position at the tth frame is $\mathbf{Y}_{a,0,t} \in \mathbb{R}^2$ for the ground truth and $\widehat{\mathbf{Y}}_{a,j,t} \in \mathbb{R}^2$ for prediction j, where $0 \leq t < t_f$. We then compute the velocities corresponding to the ground truth $\mathbf{V}_{a,0} \in \mathbb{R}^{(t_f-1)\times 2}$ and each prediction $\widehat{\mathbf{V}}_{a,j} \in \mathbb{R}^{(t_f-1)\times 2}$.

Many of the following metrics make use of aggregate functions. For any d-dimensional vector $\mathbf{v} \in \mathbb{R}^d$, we denote the minimum value by $\Omega(\mathbf{v})$, the mean value by $\Theta(\mathbf{v})$, and the maximum value by $\mathrm{O}(\mathbf{v})$. For a matrix of d-many 2D vectors $\mathbf{D} \in \mathbb{R}^{d\times 2}$, function $\Xi(\mathbf{D}, b)$ transforms the 2D vectors into a probability distribution $\mathbf{p} \in \mathbb{R}^b$ over b-many equiangular bins, which radiate from the origin (Fig. 3). Finally, we denote the L_2 norm by $\|\cdot\|$.

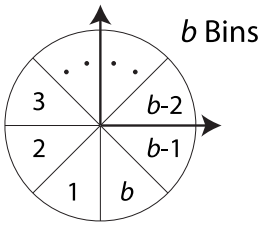

Fig. 3. This images shows how $b = 8$ bins would be arranged in 2D space.

4.2. Accuracy metrics: Comparison to ground truth

Accuracy metrics from computer vision literature are responsible for comparing the ground truth with the predictions based on the displacement error.

Average Displacement Error (ADE). ADE is computed for each prediction j as $\mathbf{a}_j$, the average distance between a position in the ground truth and a position in the prediction across t_f frames (Eq. (1)) [28]. It is then aggregated across the k predictions in three ways: minimum, mean, and maximum, which offers a more reliable expectation of a model's accuracy than the minimum alone.

Final Displacement Error (FDE). FDE is computed for each prediction j as $\mathbf{b}_j$, the distance between the final positions of the ground truth and the prediction (Eq. (2)) [2]. It is aggregated across the k predictions in the same ways as ADE for better reliability.

$$\mathrm{ADE}\left(\mathbf{Y}_a, \widehat{\mathbf{Y}}_a\right) = \left[\Omega(\mathbf{a}), \Theta(\mathbf{a}), \mathrm{O}(\mathbf{a})\right] \quad s.t.\ \ \mathbf{a}_j = \frac{1}{t_f}\sum_{t=0}^{t_f-1}\left\|\mathbf{Y}_{a,0,t} - \widehat{\mathbf{Y}}_{a,j,t}\right\|,\ 0 \le j < k \tag{1}$$

$$\mathrm{FDE}\left(\mathbf{Y}_a, \widehat{\mathbf{Y}}_a\right) = \left[\Omega(\mathbf{b}), \Theta(\mathbf{b}), \mathrm{O}(\mathbf{b})\right] \quad s.t.\ \ \mathbf{b}_j = \left\|\mathbf{Y}_{a,0,t_f-1} - \widehat{\mathbf{Y}}_{a,j,t_f-1}\right\|,\ 0 \le j < k \tag{2}$$

4.3. Realism metrics: Motion and interaction statistics

Realism metrics are used to describe the movement and interactions within the ground truth and the predictions separately. These metrics can then be used to uncover more nuanced differences between the ground truth and predictions. While they cannot ensure that predictions are accurate, they can ensure that predictions are realistic in their movement and plausible. Every realism metric is computed in the same way for both the ground truth and predictions, so $\mathbf{Y}$ is interchangeable with $\widehat{\mathbf{Y}}$ and $\mathbf{V}$ with $\widehat{\mathbf{V}}$. For generality, we consider the ground truth as a unimodal model with $k = 1$, but we refer to it as having k paths instead of predictions.

The following motion statistics are used to describe the movement of agent a in either the ground truth or averaged across the k predictions. They have been used to evaluate crowd simulations in computer graphics research [32], but have not yet been used to evaluate predictive models in computer vision.

$$\mathrm{L}(\mathbf{Y}_a) = \left[\frac{1}{k}\sum_{j=0}^{k-1}\sum_{t=0}^{t_f-2}\left\|\mathbf{Y}_{a,j,t+1} - \mathbf{Y}_{a,j,t}\right\|\right] \tag{3}$$

$$\mathrm{S}(\mathbf{V}_a) = \left[\frac{1}{k}\sum_{j=0}^{k-1}\Theta\left(\mathbf{S}_j\right),\ \frac{1}{k}\sum_{j=0}^{k-1}\mathrm{O}\left(\mathbf{S}_j\right)\right] \quad s.t.\ \ \mathbf{S}_{j,t} = \left\|\mathbf{V}_{a,j,t}\right\|,\ 0 \le t < t_f - 1 \tag{4}$$

$$\mathrm{A}(\mathbf{V}_a) = \left[\frac{1}{k}\sum_{j=0}^{k-1}\Theta\left(\mathbf{A}_j\right),\ \frac{1}{k}\sum_{j=0}^{k-1}\mathrm{O}\left(\mathbf{A}_j\right)\right] \quad s.t.\ \ \mathbf{A}_{j,t} = \left\|\left(\mathbf{V}_{a,j,t+1} - \mathbf{V}_{a,j,t}\right)/\Delta t\right\|,\ 0 \le t < t_f - 2 \tag{5}$$

Path Length. The average path length (m) for an agent a is computed by first finding the length of each path j and then averaging the values across all k paths (Eq. (3)).

Speed. In order to report the speed (m/s), the magnitudes $\mathbf{S} \in \mathbb{R}^{k\times(t_f-1)}$ of velocities in $\mathbf{V}_a$ are first computed for each agent a. Next, two values are reported for speed: the mean speed averaged across k paths and the maximum speed averaged across k paths. For each path j of agent a, the mean and maximum speed are computed across $t_f - 1$ frames (Eq. (5)).

Acceleration Magnitude. Similar to speed, we first compute the magnitudes $\mathbf{A} \in \mathbb{R}^{k\times(t_f-2)}$ of the difference between every pair of consecutive velocities in $\mathbf{V}_a$ for each agent a. The acceleration magnitude (m/s^2) $\mathrm{A}(\mathbf{V}_a)$ is then reported in the same way as speed: the mean acceleration magnitude averaged across k paths and the maximum magnitude averaged across k paths (Eq. (5)).

Traditional measures of collision are unsuitable for multimodal models in which an agent a may be colliding with agent b when it takes the direction of path j, but not when it takes the direction of path $j + 1$. We therefore propose multiverse metrics such as Agent Collision-Free Likelihood (ACFL) and Environment Collision-Free Likelihood (ECFL) to measure the A2A and A2E interactions of multimodal models respectively.

$$\mathrm{ACFL}(\mathbf{Y}, a) = \left[\frac{1}{k}\sum_{j=0}^{k-1}\prod_{b=0}^{n-1}\prod_{i=0}^{k-1}\prod_{t=0}^{t_f-1}\mathbf{1}_{\mathbb{R}>0}\left(\left\|\mathbf{Y}_{a,j,t} - \mathbf{Y}_{b,i,t}\right\| - r\right)\right] \quad s.t.\ \ a \ne b \tag{6}$$

$$\mathrm{ECFL}(\mathbf{Y}_a, \mathbf{E}) = \left[\frac{1}{k}\sum_{j=1}^{k}\prod_{t=0}^{t_f-1}\mathbf{E}\left[\ \left\lfloor s\cdot\mathbf{Y}_{a,j,t,1}\right\rfloor, \left\lfloor s\cdot\mathbf{Y}_{a,j,t,0}\right\rfloor\ \right]\right] \tag{7}$$

$$\mathrm{MVE}(\mathbf{Y}_a) = -\sum_{p\in\mathbf{p}} p\cdot\log_2(p)\ \ s.t.\ \ \mathbf{p} = \Xi\left(\mathbf{D}, 20\right),\quad \mathbf{D}_j = \frac{1}{t_f-1}\left(\sum_{t=1}^{t_f-1}\mathbf{Y}_{a,j,t}\right) - \mathbf{Y}_{a,j,0},\ 0 \le j < k \tag{8}$$

Agent Collision-Free Likelihood (ACFL). In order to assess the quality of A2A interaction under the k^n possible futures for n agents, we propose ACFL, which computes the probability that agent a has a path that is free of collision in all of the $k^{(n-1)}$ possible futures with other agents (Eq. (6)). The indicator function $\mathbf{1}_{\mathbb{R}>0}$ returns 1 when the distance between agents a and b is greater than r meters at time t, and 0 otherwise. This means that if their centers of mass are within r meters of each other, they are considered to be colliding. For analysis, r has been set to 0.3 m (∼1 foot), because it is unlikely that humans passing each other within 0.4 m would not collide [44]. While both real and synthetic data are expected to have an ACFL of ∼1.00 for $r = 0.3$ m, predictive models are not limited by physical rules and can

Table 1
This table showcases the evaluation results of Social GAN (SGAN), PECNet (PECN), and Trajectron++ (T++) after training on either A2A, A2E, or both A2A and A2E and testing on A2A and A2E separately. For every metric in a testing set, the best value has been made bold for each model. Models where minimum accuracy metrics disagree with the averages are red.

Test	Model	Train	Accuracy metrics		Realism metrics						Decidab.
			ADE↓ min/mean/max	FDE↓ min/mean/max	Length	Speed mean/max	Accel. mean/max	ACFL	ECFL	%Diff.↓	MVE↓
Agent-to-Agent Interaction	GT	N/A	0.00/0.00/0.00	0.00/0.00/0.00	4.43	1.01/1.32	0.29/1.04	0.95	1.00	0	0.00
	SGAN	A2A	**0.36**/0.77/1.50	**0.62**/1.61/3.33	**4.22**	**0.96**/1.42	0.09/**0.56**	0.30	**0.98**	**48**	0.90
		A2E	2.21/2.48/2.81	4.02/4.65/5.48	3.15	0.72/1.38	**0.12**/0.40	**0.58**	0.97	51	**0.70**
		Both	0.37/**0.74**/**1.35**	0.65/**1.55**/**2.97**	4.13	0.94/**1.32**	0.06/0.33	0.33	0.98	51	0.84
	PECN	A2A	**0.63**/**0.65**/**0.68**	**1.12**/**1.28**/**1.45**	4.50	1.02/2.15	**0.48**/**3.41**	0.56	0.98	**56**	**0.07**
		A2E	1.25/1.28/1.31	1.83/2.00/2.20	**4.50**	**1.02**/4.16	1.13/8.80	**0.59**	**0.98**	166	0.10
		Both	0.73/0.76/0.79	1.44/1.59/1.74	4.78	1.08/2.61	0.49/4.57	0.57	0.98	85	0.10
	T++	A2A	**0.22**/0.66/1.85	**0.42**/1.51/4.16	**4.38**	**1.00**/2.32	**0.36**/3.09	0.22	**0.98**	47	**1.08**
		A2E	0.56/1.06/1.77	1.13/2.29/**3.90**	4.22	0.96/**1.79**	0.29/**2.18**	**0.25**	0.98	**46**	1.41
		Both	0.23/**0.64**/**1.76**	0.43/**1.48**/4.02	4.35	0.99/2.27	0.35/2.96	0.22	0.98	47	1.13
Agent-to-Env. Interaction	GT	N/A	0.00/0.00/0.00	0.00/0.00/0.00	5.51	1.25/1.40	0.18/0.51	1.00	1.00	0	0.00
	SGAN	A2A	0.28/0.66/1.33	0.50/1.48/3.14	**5.42**	**1.23**/1.70	0.08/**0.45**	0.29	0.90	**47**	0.82
		A2E	**0.19**/**0.41**/**0.96**	**0.27**/**0.86**/**2.17**	4.19	0.95/**1.33**	**0.09**/0.28	**0.35**	**0.94**	48	**0.64**
		Both	0.19/0.56/1.25	0.32/1.28/3.02	5.03	1.14/1.57	0.08/0.40	0.32	0.92	49	0.65
	PECN	A2A	0.47/0.49/0.51	0.98/1.12/1.27	**5.35**	**1.22**/**1.72**	**0.32**/**2.79**	**0.64**	0.92	**117**	**0.03**
		A2E	**0.29**/**0.31**/**0.34**	**0.63**/**0.75**/**0.90**	5.64	1.28/2.44	0.40/3.50	0.60	**0.94**	148	0.04
		Both	0.32/0.34/0.37	0.70/0.81/0.92	5.64	1.28/2.29	0.34/3.41	0.60	0.93	157	0.06
	T++	A2A	0.17/0.81/2.43	0.34/1.86/5.54	5.48	1.25/3.10	0.53/4.41	0.18	0.90	43	1.24
		A2E	**0.10**/**0.29**/**0.64**	**0.19**/**0.69**/**1.61**	**5.41**	**1.23**/**1.63**	0.18/**1.38**	**0.47**	**0.95**	**40**	**0.73**
		Both	0.12/0.37/1.11	0.23/0.87/2.55	5.41	1.23/2.00	**0.27**/2.04	0.42	0.93	40	0.76

achieve lower values of ACFL. Increasing r will decrease the ACFL as the number of A2A interactions increases.

Environment Collision-Free Likelihood (ECFL). ECFL complements ACFL in that it measures the quality of A2E interaction under the k possible futures that agent a can interact with the environment (Eq. (7)). Namely, it reports the probability that agent a has a path that is free of collision with the environment. The environment is represented by a binary matrix $\mathbf{E}$, in which each cell corresponds to a square space and is equal to 1 if that space is navigable and 0 otherwise. $\mathbf{E}[0,0]$ is aligned with the origin of the position data $\mathbf{Y}$, but $\mathbf{E}$ has a scale of $1/s$ meters per unit as opposed to 1 meter per unit like $\mathbf{Y}$. This means that the position $[x, y] = \mathbf{Y}_{a,j,t}$ of agent a taking path j at time t maps to $\mathbf{E}\big[\lfloor s \cdot y \rfloor, \lfloor s \cdot x \rfloor\big]$. For analysis, s has been set to 2 based on the dataset. When agent a's center of mass is intersecting a non-navigable region of the environment like a wall, the agent is considered to be colliding with the environment.

4.4. Decidability metric: Certainty in movement direction

Decidability is a measure of a model's uncertainty in the movement direction of agents, and it is not strictly opposite between unimodal and multimodal models. If a multimodal model has low enough uncertainty in an agent's direction of movement, we consider it to be decidable.

Multiverse Entropy (MVE). We compute MVE to measure the decidability for agent a. We first transform each path j into an average direction vector $\mathbf{D}_j \in \mathbb{R}^2$ as the vector from the initial position $\mathbf{Y}_{a,j,0}$ to the average position of the $t_f - 1$ subsequent points (Eq. (8)). The average direction vectors $\mathbf{D}$ are then transformed into a probability distribution $\mathbf{p} \in \mathbb{R}^b$ over a vector of b-many equiangular bins (Fig. 3). Finally, the entropy of $\mathbf{p}$ is reported as MVE. High values of ACFL and ECFL are contingent on low MVE (high decidability), because high certainty in the direction that an agent will travel along will cause fewer potential collisions with other agents (ACFL) and the environment (ECFL). For experimental purposes, b has been set to k, so that MVE is maximized when every prediction is in a different direction.

4.5. Comparing realism metrics

In order to compare realism metrics between the ground truth and predictions for an agent a, we first compute a feature vector for the ground truth $\mathbf{F}_a = \langle \mathrm{L}(\mathbf{Y}_{a,0}), \mathrm{S}(\mathbf{V}_a), \mathrm{A}(\mathbf{V}_a), \mathrm{ACFL}(\mathbf{Y}, a), \mathrm{ECFL}(\mathbf{Y}_a, \mathbf{E}) \rangle$, where $\langle \cdot, \cdot \rangle$ denotes vector concatenation. The same vector concatenation is used to compute the feature vector $\widehat{\mathbf{F}}_{a,j} \in \mathbb{R}^7$ for each prediction j. Eq. (9) returns the percent differences $\widehat{\mathbf{C}}_a \in \mathbb{R}^k$ between the feature vectors of each prediction j and the ground truth of agent a.

$$\widehat{\mathbf{C}}_{a,j} = \frac{100}{7} \sum_{f=0}^{6} \frac{\left| \widehat{\mathbf{F}}_{a,j,f} - \mathbf{F}_{a,0,f} \right|}{\mathbf{F}_{a,0,f}} \quad s.t. \;\; \mathbf{F}_{a,0,f} > 0 \;,\; 0 \leq j < k \tag{9}$$

5. Results

In order to understand the limits of not only the SOTA but also the models that paved the way towards the SOTA, we evaluate three critical multimodal models that are capable of either short-term or long-term trajectory prediction and provide a large coverage over the performance of prior models (Fig. 1.b). In particular, we have selected (1) Social GAN (SGAN) [7], one of the earliest models; (2) Trajectron++ (T++) [8], a SOTA model for short-term trajectory prediction; and (3) PECNet (PECN) [9], a SOTA model for long-term trajectory prediction.

5.1. Training protocol

Each of the three models was trained on 3 combinations from the **A2X** Dataset: A2A interaction, A2E interaction, and both (Fig. 1.b), producing a total of 9 models. We denote that either a model has been trained on a particular combination using a subscript, e.g., SGAN_{Both}. Each trained model was then evaluated on the testing sets of the 3 combinations (Fig. 1.c). The results of the evaluations on A2A and A2E are reported in Table 1, while the results on both A2A and A2E combined and corresponding visualizations are reported in the Supplementary Materials. According to the dataset, the following parameters have been set for the

evaluation: $k = 20$, $t_p = 8$, $t_f = 12$, and $\Delta t = 0.4$, meaning that each agent is receiving 3.2 s of input data and predicting 4.8 s into the future.

Each row of Table 1 reports the accuracy, realism, and decidability metrics of a model averaged across the agents of every testing scenario for a given dataset. The first 5 columns of realism metrics correspond to the dimensions of $\mathbf{F}$ and $\widehat{\mathbf{F}}$, the feature vectors used to compute the percent difference between the ground truth (GT) and predictions. The mean percent difference $\Theta(\widehat{\mathbf{C}}_a)$ of each agent a is averaged across all agents and reported in the final column of the realism metrics. For all accuracy metrics, the realism percent difference, and the decidability metric, a lower value is favorable, while for the remaining realism metrics, a value closer to the ground truth is favorable.

5.2. Analysis

5.2.1. Training on both types of interaction consistently has near-best accuracy

As expected, we find that in terms of all accuracy metrics, models trained on a single type of interaction perform very poorly on test scenarios that feature the other type of interaction. By training any of the three models (SGAN, PECN, or T++) on both types of interactions, we find that the accuracy is consistently near-best among all three training datasets by a small margin. For testing on A2A, a model trained on both types is closer in accuracy to the same model trained on A2A, and for testing on A2E, it is closer to the same model trained on A2E. In fact, when testing on A2A, training SGAN and T++ on both types achieves the best mean/maximum ADE and mean FDE among all training datasets. This makes training on both types of interactions an excellent compromise for balancing accuracy between real-world cases from A2A and critical synthetic cases from A2E.

5.2.2. Existing evaluation metrics can misjudge model accuracy

When testing on A2A, SGAN_{A2A} and T++_{A2A} are misjudged as being better than SGAN_{Both} and T++_{Both} according to minimum ADE and minimum FDE (highlighted in red). Reliance on these overly optimistic existing metrics will lead to choosing models that are less accurate than others on average.

5.2.3. Realism metrics influence model choice based on the use case

We cannot rely only on the accuracy of models to determine which is best, since anything short of perfect accuracy carries risk. The realism metrics allow us to better understand a model's performance in the context of its application. For example, we find that the maximum speed and acceleration for T++_{Both} are significantly higher than the ground truth, which for an application in socially compliant robot navigation can discomfort or potentially harm surrounding humans [45]. In contrast, SGAN_{Both} has lower average accuracy by a small margin, but it boasts higher realism by a large margin in terms of maximum speed, maximum acceleration magnitude, and ACFL. We attribute SGAN_{Both}'s higher ACFL to the tighter spread of its predictions than T++_{Both} according to MVE. Ultimately, the choice of a model depends on the application, but without the joint consideration of the proposed accuracy and realism metrics, a practitioner may be led to choose an unsuitable model.

5.2.4. A2E is essential for learning collision avoidance

Models trained exclusively on A2E interactions tend to have lower likelihoods of A2A collision (higher ACFL) than models trained on A2A interactions alone or on both types of interactions. This highlights the importance of A2E for improving robustness even in real-world scenarios such as A2A.

5.2.5. ECFL indicates that A2A scenarios have trivial A2E interactions

Models trained on A2E achieve the lowest likelihood of A2E collision (highest ECFL) when testing on A2E, but still have some room to improve. In contrast, we find that ECFL is nearly perfect for A2A scenarios, indicating that A2A scenarios do not challenge models with A2E interactions.

5.2.6. Multimodal models can be decidable

Although PECN is a multimodal model, it has a near-zero MVE, which is significantly lower than SGAN and T++. This indicates that PECN has certainty in the direction that agents will travel along (regardless of whether the direction is correct). PECN also achieves the highest ACFL owing to its low MVE, which is low enough to consider PECN as being decidable and likely helps it in performing long-term trajectory prediction.

5.2.7. T++ is a significant indicator of SOTA performance

According to the three latest SOTA works in HTP [46–48], T++ has used to represent the existing SOTA. As it is the best representation of the SOTA, we examine it more closely. In terms of accuracy, T++ achieves the overall best mean ADE on A2A and the overall best minimum/mean ADE and minimum/mean FDE on A2E. In terms of realism, T++ achieves the overall best percent difference in realism metrics on both A2A and A2E, achieving the best percent difference when trained on A2E alone. Regarding decidability, the MVE for T++_{A2A} is marginally better than T++_{Both} when testing on A2A, but on A2E, the MVE for T++_{A2A} is almost twice worse than T++_{Both}. These findings reveal a comprehensive improvement in accuracy, realism, and decidability for T++ when training either jointly on both A2A and A2E or exclusively on the critical synthetic cases in A2E.

5.2.8. Open challenges in HTP

Models trained on both types of interactions do not yet generalize to A2E better than models trained on A2E alone as SGAN_{Both} and T++_{Both} have for A2A, meaning that there is still much room for improvement.

5.3. Multimodal Model Collapse (MMC)

Accuracy metrics cannot be computed on never-before-seen data, because the ground truth is unknown. Consequently, it becomes impossible to find the predicted path with minimum error in accuracy and selecting an arbitrary prediction risks the maximum error. We therefore propose MMC, a baseline greedy method which can make use of the realism metrics to collapse the k predictions of an undecidable multimodal model into the single most socially compliant prediction. In particular, we rely on the proposed comparison of realism metrics (Section 4.5), but instead of computing $\mathbf{F}_a$ from ground truth testing data $\mathbf{Y}_{a,0}$ for each agent a, we compute it as the average across *all* agents in the ground truth *training* data from the same environment. We then replace the k predictions $\widehat{\mathbf{Y}}_a$ with the single prediction j that minimizes the percent difference $\widehat{\mathbf{C}}_{a,j}$ for each agent a. This prediction is the closest in realism to prior ground truth for the same type of scenario (Eq. (9)). Table 2 shows the result of applying this technique to all 9 models. Across all models, we find that the ADE/FDE of the collapsed prediction is only ∼15.76% worse than the mean ADE/FDE of the uncollapsed predictions, and ∼31.63% better than the maximum ADE/FDE. Although the accuracy of the most realistic prediction is lower than the average accuracy over 20 predictions, its performance is consistently much better than the worst-case. Furthermore, the social compliance of models is drastically improved through MMC, making them less likely to produce collisions with other agents.

Table 2
This table reports the results of MMC on each of the 9 trained models. On average, MMC produces predictions that are consistently better than the worse case prediction prior to MMC. Only one value is reported for ADE and FDE, because the minimum, mean, and maximum are equal when $k = 1$. The MVE is always 0 when $k = 1$.

Test	Model	Train	Accuracy metrics		Realism metrics						Decidab.
			ADE↓ min = mean = max	FDE↓ min = mean = max	Length	Speed mean/max	Accel. mean/max	ACFL	ECFL	%Diff.↓	MVE↓
Agent-to-Agent Interaction	GT	N/A	0.00	0.00	4.43	1.01/1.32	0.29/1.04	0.95	1.00	0	0.00
	SGAN	A2A	0.91	1.99	**4.28**	**0.97**/1.20	0.16/**0.41**	0.69	**0.99**	**37**	0.00
		A2E	2.57	4.97	3.75	0.85/**1.32**	**0.20**/0.37	**0.79**	0.97	40	0.00
		Both	**0.86**	**1.86**	4.25	0.97/1.15	0.11/0.23	0.70	0.99	41	0.00
	PECN	A2A	**0.65**	**1.27**	**4.44**	**1.01**/**1.56**	**0.33**/**1.79**	0.66	**0.98**	**56**	0.00
		A2E	1.28	2.03	4.33	0.98/3.23	1.02/6.37	**0.68**	0.98	166	0.00
		Both	0.76	1.55	4.70	1.07/2.12	0.44/3.18	0.64	0.98	85	0.00
	T++	A2A	**0.81**	**1.83**	**4.51**	**1.03**/**1.31**	0.44/0.98	**0.66**	**0.99**	**26**	0.00
		A2E	1.05	2.27	4.53	1.03/1.32	**0.42**/0.97	0.63	0.98	30	0.00
		Both	0.81	1.84	4.51	1.03/1.31	0.44/**1.00**	0.65	0.99	26	0.00
Agent-to-Env. Interaction	GT	N/A	0.00	0.00	5.51	1.25/1.40	0.18/0.51	1.00	1.00	0	0.00
	SGAN	A2A	0.76	1.84	**5.00**	**1.14**/1.44	0.15/**0.33**	0.63	0.96	**38**	0.00
		A2E	**0.69**	**1.60**	4.73	1.08/1.30	0.13/0.23	**0.68**	**0.98**	40	0.00
		Both	0.73	1.77	4.55	1.03/**1.36**	**0.16**/0.27	0.66	0.97	40	0.00
	PECN	A2A	0.49	1.11	5.39	1.22/**1.45**	0.25/**1.10**	**0.69**	0.93	**117**	0.00
		A2E	**0.30**	**0.71**	**5.54**	**1.26**/1.71	**0.31**/1.41	0.62	0.93	148	0.00
		Both	0.34	0.78	5.60	1.27/1.97	0.32/1.41	0.64	**0.94**	157	0.00
	T++	A2A	0.90	2.06	4.99	1.13/1.48	0.57/1.27	0.46	0.97	31	0.00
		A2E	**0.34**	**0.86**	**5.36**	**1.22**/**1.44**	**0.29**/0.85	**0.61**	**0.98**	**24**	0.00
		Both	0.52	1.20	5.34	1.21/1.48	0.41/**0.99**	0.57	0.97	28	0.00

6. Temporal resolution of human trajectory prediction datasets

The proposed dataset and evaluation metrics have been made consistent with existing datasets and metrics to ensure compatibility with prior work and applicability to future work. The most prevalent HTP datasets, ETH [28] and UCY [12], were collected at $1/\Delta t = 2.5$ Hz, which has been adopted by many HTP models that have downsampled other datasets to similar temporal resolutions [17,49]. These models are learning the dynamics of human navigation, but only those which can be measured at $1/\Delta t$ Hz or lower. While 2.5 Hz exceeds the preferred human step frequency of ∼2 Hz [50], we show in the following simulation experiment that for scenarios with many A2 A interactions, realism metrics (e.g., speed, acceleration, turning, speed change, and ACFL) can quickly degrade in accuracy as temporal resolutions decreases.

6.1. Simulation experiment protocol

In order to investigate the loss of information as temporal resolution decreases, we considered several additional realism metrics for capturing more dynamics: (1) the average radians turned per second, (2) the maximum radians turned per second, (3) the average number of turning direction changes per second, (4) the average positive or negative change in speed per second, (5) the maximum change in speed per second, and (6) the average number of changes from increasing to decreasing speed (or vice versa) per second [32]. This superset of realism metrics was recorded for two simulations (interactionless and interacting) using the same model and parameterization as A2E, each with 448 agents. In the interactionless simulation, each agent was tasked with linearly navigating to a goal by itself in an obstacle-less environment. In the interacting simulation, agents were initialized with a density of 1 agent per meter^2 in the 1.5-meter bottleneck scenario (Section 3.2, Fig. 2). Interactionless scenarios are characterized as having no A2A or A2E interactions, and interacting scenarios are characterized as having an abundance of these types of interactions. The trajectories of all agents were recorded at 80 Hz and divided into 4.8-second scenarios (Section 5.1) amounting to 23,180 for the interactionless simulation and 34,187 for the interacting simulation. All scenarios were then resampled at 40, 20, 10, 5, and 2.5 Hz.

6.2. Analysis

For each simulation, the average values of realism metrics at each temporal resolution have been reported in Table 3. In both simulations, trajectory length decreased by ∼0.5 m from 80 Hz to 2.5 Hz, indicating nonlinear motion at higher temporal resolutions. Despite the lack of A2A and A2E interactions in interactionless scenarios, which should have produced perfectly linear motions, the quantization of pathfinding instructions resulted in miniscule values for mean and max turning. However, since agents were otherwise uninterrupted, interactionless scenarios had higher values of length, mean speed, max speed, and ACFL than the interacting scenarios. These agents were also better able to maintain their desired speed of 1.5 m/s and direction toward the goal, causing the mean and max speeds to be highly similar and the acceleration, turning, and speed change to be considerably low. The ECFL was maximal in all interactionless and interacting scenarios, because unlike the predictions made by HTP models, it is nearly impossible for a simulated agent's center of mass to intersect an obstacle. Therefore, ECFL has been excluded from this analysis.

For interacting scenarios, the mean and max for acceleration, turning, and speed change were significantly higher at 80 Hz than at 2.5 Hz, indicating that agents were jerking away from other agents or walls after nearing them. Furthermore, the mean speed change notably switched from negative to positive between 5 Hz and 10 Hz, meaning that agents were slowing down at lower temporal resolutions (i.e., larger scales) but speeding up at higher temporal resolutions (i.e., smaller scales). Although a high initial temporal resolution can always be downsampled to analyze data at multiple scales, it is currently not possible to accurately upsample a low initial temporal resolution, such as 2.5 Hz. The disparity between low and high temporal resolutions is also apparent with ACFL. At 2.5 Hz, ACFL is overestimated by

Table 3
This table reports the average value for realism metrics at different temporal resolutions for the interactionless and interacting simulations. On average, all metrics tend to diverge from their measurements at 80 Hz.

Type	Hz	Length	Mean speed	Max speed	Mean Accel.	Max Accel.	Mean turning	Max turning	Turning direction changes	Mean speed change	Max speed change	Speed Inc./Dec. changes	ACFL	ECFL
Interactionless	80	7.1570	1.4949	1.5004	0.0512	0.1751	0.0000	0.0018	0.0010	0.0050	0.0831	0.0338	1.0000	1.0000
	40	7.1383	1.4949	1.5002	0.0193	0.1172	0.0003	0.0036	0.0023	0.0069	0.0903	0.0271	1.0000	1.0000
	20	7.1008	1.4949	1.5001	0.0126	0.0990	0.0004	0.0043	0.0035	0.0080	0.0899	0.0322	1.0000	1.0000
	10	7.0258	1.4948	1.5000	0.0106	0.0864	0.0005	0.0042	0.0044	0.0085	0.0811	0.0336	1.0000	1.0000
	5.0	6.8758	1.4947	1.5000	0.0096	0.0697	0.0005	0.0037	0.0051	0.0086	0.0673	0.0439	1.0000	1.0000
	2.5	6.5758	1.4945	1.5000	0.0087	0.0478	0.0005	0.0028	0.0065	0.0078	0.0460	0.0483	1.0000	1.0000
Interacting	80	5.0899	1.0632	1.3599	1.0175	8.2408	1.6271	32.0681	3.7244	0.0084	3.0623	3.8904	0.9020	1.0000
	40	5.0755	1.0629	1.3578	0.9657	4.8384	1.5736	20.0786	2.9963	0.0071	2.3942	3.7266	0.9116	1.0000
	20	5.0464	1.0624	1.3526	0.9074	3.5220	1.4933	12.1472	2.5194	0.0042	2.0362	3.1515	0.9236	1.0000
	10	4.9864	1.0609	1.3387	0.7723	2.4646	1.2693	6.8014	2.0949	0.0018	1.5361	2.3962	0.9347	1.0000
	5.0	4.8645	1.0575	1.3051	0.5386	1.4400	0.8758	3.2959	1.5445	−0.0001	0.9145	1.6417	0.9442	1.0000
	2.5	4.6279	1.0518	1.2500	0.3105	0.6923	0.4888	1.3383	1.0231	−0.0033	0.4299	1.0067	0.9572	1.0000

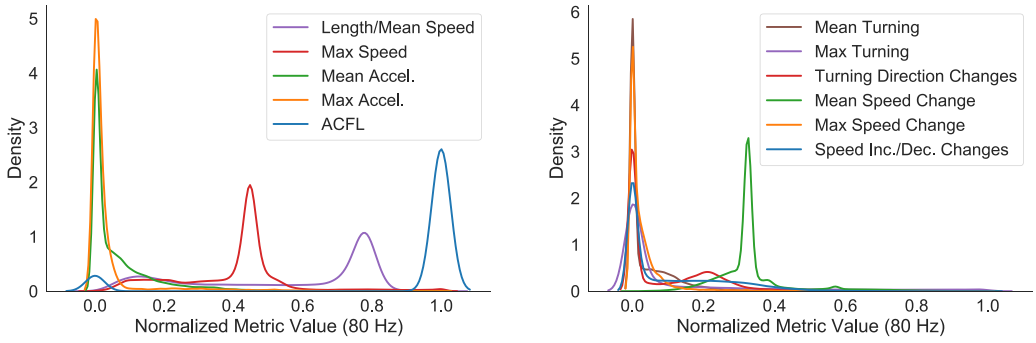

Fig. 4. The above plots show the distribution of each realism metric sampled at 80 Hz from the interacting simulation.

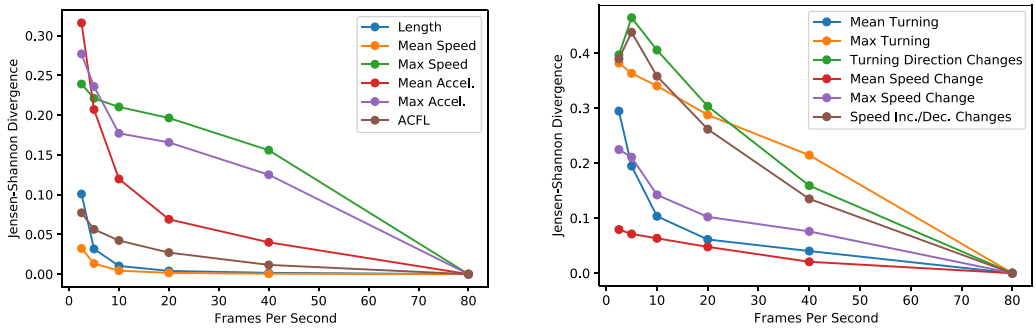

Fig. 5. The above plots show the Jensen–Shannon Divergence of different temporal resolutions with respect to 80 Hz for every realism metric.

almost 6% compared to the ACFL at 80 Hz, which is a notable difference.

It is evident that as the temporal resolution decreases, the average values of all realism metrics diverge in both simulations. We complement this analysis by comparing the distributions of values instead of their averages. These distributions sampled at 80 Hz from the interacting simulation are visualized in Fig. 4. To quantify the changes in a realism metric from 80 Hz to 2.5 Hz, the distribution at each temporal resolution is compared with the 80 Hz distribution using Jensen–Shannon Divergence (JSD) [51] (Table 3). These distributions are represented as histograms, and for each metric, the same bin size is used across both simulations and all temporal resolutions. Since the simplistic agent dynamics in the interactionless simulation are effectively the same between 80 Hz and 2.5 Hz, the bin size of each metric is calibrated as the smallest size from $\{\alpha \cdot 10^{\beta} \mid \alpha \in 1..9, \beta \in -3..1\}$ at which all temporal resolutions have zero JSD with respect to 80 Hz in *interactionless* scenarios. This ensures that the distributions are neither too sensitive to minute differences nor too insensitive to notable differences. Fig. 5 depicts the JSD values of realism metrics at different temporal resolutions for *interacting* scenarios. For all metrics, there is still a clear divergence as temporal resolution decreases. In fact, a large majority of metrics increase monotonically in JSD as temporal resolution decreases and below 20 Hz, they increase faster than a linear rate. Fig. 6 shows the average JSD across the superset of realism metrics at each temporal resolution. The average JSD grows approximately linearly from 80 Hz to 20 Hz and at a significantly faster rate from 20 Hz to 2.5 Hz.

6.3. Discussion

These results convey that particularly in scenarios with many A2A and A2E interactions, having a temporal resolution of 2.5 Hz likely degrades the realism metrics measured on real data given its effect on synthetic data. Since most current datasets do not feature this level of interaction [12,13,13–17,28], we believe there is no immediate issue with their temporal resolutions. Moving

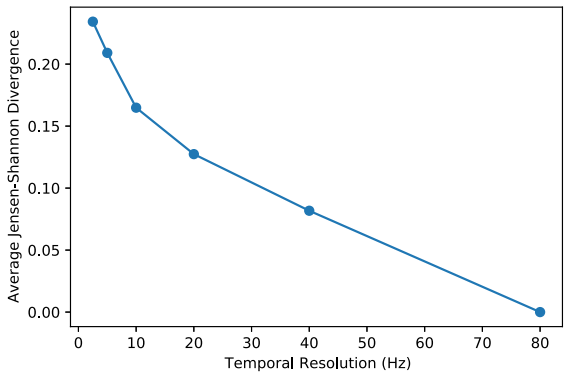

Fig. 6. The above plot shows the average Jensen–Shannon Divergence (JSD) across all realism metrics. Between 80 Hz and 20 Hz, the JSD increases at an approximately linear rate, but below 20 Hz, the JSD increases at a significantly faster rate.

forward, we recommend the acquisition of data using a temporal resolution of at least 20 Hz, i.e., the point at which the average JSD starts to decrease linearly as temporal resolution increases. This would encourage the future modeling of HTP at higher temporal resolutions and thereby allow models to more accurately learn motion statistics and collision avoidance, which current datasets sampled at 2.5 Hz limit (Table 3). Furthermore, the accuracy gained from increasing the temporal resolution of datasets and models would improve the interpretation of realism metrics when making the choice between two models for a specific use case.

The Social Force model used to simulate the synthetic data in both the A2E dataset (Section 3.2) and the simulation experiment (Section 6.1) was parameterized according to the original work [1] to produce realistic results. However, it is known that most realistic parameterizations tend to produce oscillating behavior [52], which is registered by changes in speed from increasing to decreasing (or vice versa) and turning direction changes. In order to prevent insignificant oscillations from influencing these two metrics, instantaneous changes in turning direction under 0.01 rad/s and instantaneous changes in speed from increasing to decreasing (or vice versa) under 0.01 m/s^2 were ignored.

Although synthetic data exhibits deviations from real data, it uniquely captures minute details in motion without any noise and at a high temporal resolution. This allows for the investigation of temporal resolution without considering the effects of spatial resolution. However, for future data acquisition in the real world, this must be carefully considered so as not to negate the benefits of a high temporal resolution with a low spatial resolution.

7. Conclusion

With the growing attention toward human trajectory prediction, it has become more important than ever to unify future research efforts in the right direction in terms of datasets and evaluation. In this work, we have brought to light the shortcomings of existing datasets, which hinder generalization, and existing evaluation metrics, which misrepresent model performance. By augmenting existing datasets with critical scenarios that feature substantial interactions between pedestrian agents and the environment, we have evidenced that models can generalize better. By proposing a comprehensive set of novel and existing evaluation metrics, we have not only proven the unreliability of existing evaluation metrics, but also highlighted the subtle factors that are essential for choosing the best trajectory prediction model for a particular application. Finally, we have proposed a guideline for future data acquisition in HTP to ensure that in the long term, researchers can tackle the learning of dynamics not only at a low temporal resolution, but at high temporal resolutions and *multiple scales* of temporal resolution, which comes with a different set of complexities. Together, these contributions show that there is still much room for improvement even among the SOTA models.

CRediT authorship contribution statement

Samuel S. Sohn: Conceptualization, Methodology, Software, Formal analysis, Writing - Original Draft. **Mihee Lee:** Software, Investigation, Writing - Original Draft. **Seonghyeon Moon:** Software, Investigation. **Gang Qiao:** Software, Investigation. **Muhammad Usman:** Visualization. **Sejong Yoon:** Writing – review & editing, Supervision, Funding acquisition. **Vladimir Pavlovic:** Writing – review & editing, Supervision, Funding acquisition. **Mubbasir Kapadia:** Conceptualization, Writing – review & editing, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The research was supported in part by NSF awards: IIS-1703883, IIS-1955404, IIS-1955365, RETTL-2119265, and EAGER-2122119. The authors acknowledge use of the TCNJ ELSA HPC cluster, funded by NSF grant OAC-1828163, for conducting the research reported in this paper. This material is based upon work supported by the U.S. Department of Homeland Security under Grant Award Number 22STESE00001 01 01. Disclaimer. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.cag.2022.05.010. It provides dataset statistics and additional quantitative and qualitative results related to Section 5.

References

[1] Helbing D, Molnar P. Social force model for pedestrian dynamics. Phys Rev E 1995;51(5):4282.
[2] Alahi A, Goel K, Ramanathan V, Robicquet A, Fei-Fei L, Savarese S. Social LSTM: Human trajectory prediction in crowded spaces. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 961–71.
[3] Wiener JM, Büchner SJ, Hölscher C. Taxonomy of human wayfinding tasks: A knowledge-based approach. Spatial Cogn Comput 2009;9(2):152–65.
[4] Ferrer G, Garrell A, Sanfeliu A. Social-aware robot navigation in urban environments. In: 2013 European conference on mobile robots. IEEE; 2013, p. 331–6.
[5] Pelechano N, Allbeck JM, Kapadia M, Badler NI. Simulating heterogeneous crowds with interactive behaviors. CRC Press; 2016.
[6] Rudenko A, Palmieri L, Herman M, Kitani KM, Gavrila DM, Arras KO. Human motion trajectory prediction: A survey. Int J Robot Res 2020;39(8):895–935.
[7] Gupta A, Johnson J, Fei-Fei L, Savarese S, Alahi A. Social GAN: Socially acceptable trajectories with generative adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2018. p. 2255–64.

[8] Salzmann T, Ivanovic B, Chakravarty P, Pavone M. Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In: European conference on computer vision (ECCV). Springer; 2020, p. 683–700.

[9] Mangalam K, Girase H, Agarwal S, Lee K-H, Adeli E, Malik J, Gaidon A. It is not the journey but the destination: Endpoint conditioned trajectory prediction. 2020, arXiv preprint arXiv:2004.02025.

[10] Scharine AA, McBeath MK. Right-handers and Americans favor turning to the right. Human Factors 2002;44(2):248–56.

[11] Dubey RK, Sohn SS, Hoelscher C, Kapadia M. Fusion-based wayfinding prediction model for multiple information sources. In: 2019 22th international conference on information fusion (FUSION). IEEE; 2019, p. 1–8.

[12] Lerner A, Chrysanthou Y, Lischinski D. Crowds by example. In: Computer graphics forum, vol. 26. (3):Wiley Online Library; 2007, p. 655–64.

[13] Robicquet A, Sadeghian A, Alahi A, Savarese S. Learning social etiquette: Human trajectory understanding in crowded scenes. In: European conference on computer vision (ECCV). Springer; 2016, p. 549–65.

[14] Alahi A, Ramanathan V, Fei-Fei L. Socially-aware large-scale crowd forecasting. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2014. p. 2203–10.

[15] Yan Z, Duckett T, Bellotto N. Online learning for human classification in 3D LiDAR-based tracking. In: Proceedings of the 2017 IEEE/RSJ international conference on intelligent robots and systems; 2017.

[16] Chavdarova T, Baqué P, Bouquet S, Maksai A, Jose C, Bagautdinov T, Lettry L, Fua P, Van Gool L, Fleuret F. Wildtrack: A multi-camera hd dataset for dense unscripted pedestrian detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2018. p. 5030–9.

[17] Kothari P, Kreiss S, Alahi A. Human trajectory forecasting in crowds: A deep learning perspective. IEEE Trans Intell Transp Syst 2021. http://dx.doi.org/10.1109/TITS.2021.3069362.

[18] Thalmann D, Musse SR. Crowd simulation. Springer Science & Business Media; 2012.

[19] Kapadia M, Pelechano N, Allbeck J, Badler N. Virtual crowds: Steps toward behavioral realism. Synth Lect Vis Comput Comput Graph Animation Comput Photogr Imaging 2015;7(4):1–270.

[20] Vemula A, Muelling K, Oh J. Social attention: Modeling attention in human crowds. In: 2018 IEEE international conference on robotics and automation (ICRA). 2018, p. 4601–7. http://dx.doi.org/10.1109/ICRA.2018.8460504.

[21] Zhao T, Xu Y, Monfort M, Choi W, Baker C, Zhao Y, Wang Y, Wu YN. Multi-Agent tensor fusion for contextual trajectory prediction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2019. p. 12118–26.

[22] Ivanovic B, Pavone M. The trajectron: Probabilistic multi-agent trajectory modeling with dynamic spatiotemporal graphs. In: 2019 IEEE/CVF international conference on computer vision (ICCV). 2019, p. 2375–84. http://dx.doi.org/10.1109/ICCV.2019.00246.

[23] Amirian J, Hayet J-B, Pettré J. Social ways: Learning multi-modal distributions of pedestrian trajectories with gans. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops; 2019.

[24] Mangalam K, An Y, Girase H, Malik J. From goals, waypoints & paths to long term human trajectory forecasting. 2020, arXiv preprint arXiv:2012.01526.

[25] Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial nets. In: Proceedings of the 27th international conference on neural information processing systems; 2014. p. 2672–80.

[26] Kingma DP, Welling M. Auto-encoding variational Bayes. In: 2nd International conference on learning representations (ICLR), Banff, AB, Canada, April 14-16, 2014, conference track proceedings. 2014.

[27] Sohn K, Lee H, Yan X. Learning structured output representation using deep conditional generative models. In: Neural information processing systems (NIPS). 2015.

[28] Pellegrini S, Ess A, Schindler K, Van Gool L. You'll never walk alone: Modeling social behavior for multi-target tracking. In: 2009 IEEE 12th international conference on computer vision (CVPR). IEEE; 2009, p. 261–8.

[29] Templeton A, Drury J, Philippides A. From mindless masses to small groups: conceptualizing collective behavior in crowd modeling. Rev General Psychol 2015;19(3):215–29.

[30] Amirian J, Zhang B, Castro FV, Baldelomar JJ, Hayet J-B, Pettré J. Opentraj: Assessing prediction complexity in human trajectories datasets. In: Proceedings of the asian conference on computer vision; 2020.

[31] Van Den Berg J, Guy SJ, Lin M, Manocha D. Reciprocal n-body collision avoidance. In: Robotics research. Springer; 2011, p. 3–19.

[32] Singh S, Kapadia M, Faloutsos P, Reinman G. Steerbench: A benchmark suite for evaluating steering behaviors. Comput Animation Virtual Worlds (CAVW) 2009;20(5–6):533–48.

[33] Sohn SS, Zhou H, Moon S, Yoon S, Pavlovic V, Kapadia M. Laying the foundations of deep long-term crowd flow prediction. In: European conference on computer vision (ECCV). Springer; 2020, p. 711–28.

[34] Parunak HVD. A survey of environments and mechanisms for human-human stigmergy. In: International workshop on environments for multi-agent systems. Springer; 2005, p. 163–86.

[35] Helbing D, Schweitzer F, Keltsch J, Molnar P. Active walker model for the formation of human and animal trail systems. Phys Rev E 1997;56(3):2527.

[36] Wei-Guo S, Yan-Fei Y, Bing-Hong W, Wei-Cheng F. Evacuation behaviors at exit in CA model with force essentials: A comparison with social force model. Physica A 2006;371(2):658–66.

[37] Zeng W, Chen P, Nakamura H, Iryo-Asano M. Application of social force model to pedestrian behavior analysis at signalized crosswalk. Transp Res C 2014;40:143–59.

[38] Nardini M, Jones P, Bedford R, Braddick O. Development of cue integration in human navigation. Curr Biol 2008;18(9):689–93.

[39] Lokki T, Grohn M. Navigation with auditory cues in a virtual environment. IEEE MultiMedia 2005;12(2):80–6.

[40] Jacobs LF, Arter J, Cook A, Sulloway FJ. Olfactory orientation and navigation in humans. PLoS One 2015;10(6):e0129387.

[41] Seyfried A, Boltes M, Kähler J, Klingsch W, Portz A, Rupprecht T, Schadschneider A, Steffen B, Winkens A. Enhanced empirical data for the fundamental diagram and the flow through bottlenecks. Pedestrian Evacuation Dyn 2008 2010;145–56.

[42] Haworth B, Usman M, Berseth G, Kapadia M, Faloutsos P. Evaluating and optimizing level of service for crowd evacuations. In: Proceedings of the 8th ACM SIGGRAPH conference on motion in games; 2015. p. 91–6.

[43] Guy SJ, Van Den Berg J, Liu W, Lau R, Lin MC, Manocha D. A statistical similarity measure for aggregate crowd dynamics. ACM Trans Graph 2012;31(6):1–11.

[44] Knorr AG, Willacker L, Hermsdörfer J, Glasauer S, Krüger M. Influence of person-and situation-specific characteristics on collision avoidance behavior in human locomotion.. J Exp Psychol Human Percept Perform 2016;42(9):1332.

[45] Kruse T, Pandey AK, Alami R, Kirsch A. Human-aware robot navigation: A survey. Robot Auton Syst 2013;61(12):1726–43.

[46] Mangalam K, An Y, Girase H, Malik J. From goals, waypoints & paths to long term human trajectory forecasting. 2020, arXiv preprint arXiv:2012.01526.

[47] Yuan Y, Weng X, Ou Y, Kitani K. AgentFormer: Agent-aware transformers for socio-temporal multi-agent forecasting. 2021, arXiv preprint arXiv:2103.14023.

[48] Wang C, Wang Y, Xu M, Crandall DJ. Stepwise goal-driven networks for trajectory prediction. 2021, arXiv preprint arXiv:2103.14107.

[49] Mangalam K, An Y, Girase H, Malik J. From goals, waypoints & paths to long term human trajectory forecasting. In: Proceedings of the IEEE/CVF international conference on computer vision; 2021, p. 15233–42.

[50] MacDougall HG, Moore ST. Marching to the beat of the same drummer: the spontaneous tempo of human locomotion. J Appl Physiol 2005;99(3):1164–73.

[51] Lin J. Divergence measures based on the Shannon entropy. IEEE Trans Inform Theory 1991;37(1):145–51.

[52] Kretz T. On oscillations in the social force model. Physica A 2015;438:272–85.