Inferring User Intent using Bayesian Theory of Mind in Shared Avatar-Agent Virtual Environments

Sahil Narang*
University of North Carolina, Chapel-Hill

Andrew Best†
University of North Carolina, Chapel-Hill

Dinesh Manocha‡
University of Maryland, College Park

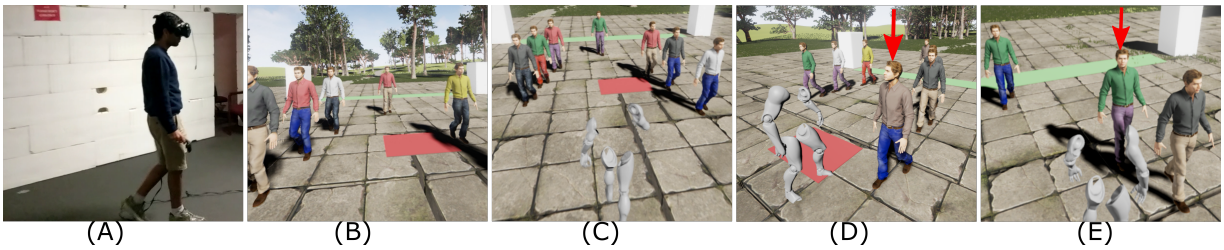


Figure 1: **Agents with Theory of Mind (ATOM) in Immersive Virtual Environments**: (A) A user interacting with the virtual agents using tracked movement and the HTC Vive headset. (B) The user is provided with a first person perspective using the HMD and his/her tracked movement is used to simulate a virtual avatar. (C) The user, by means of his/her avatar, can induce a gaze and/or locomotion-based response from the virtual agents. The user's avatar is visualized from a third person perspective for the sake of visual clarity. (D-E) ATOM agents apply the Bayesian Theory of Mind concept to reason about the observed gaze and motion cues of the user, and to determine the user's intent. In these scenarios, the user's intention is to engage a *target* agent (depicted with red arrow) in a face-to-face interaction. The *target* agent correctly infers the user's intent to interact and responds with gazing behavior. The *other* agents correctly infer the user's lack of interaction intent and do not gaze, producing a more positive experience than prior inference approaches.

Abstract

We present a real-time algorithm to infer the intention of a user's avatar in a virtual environment shared with multiple human-like agents. Our algorithm applies the Bayesian Theory of Mind approach to make inferences about the avatar's hidden intentions based on the observed proxemics and gaze-based cues. Our approach accounts for the potential irrationality in human behavior, as well as the dynamic nature of an individual's intentions. The inferred intent is used to guide the response of the virtual agent and generate locomotion and gaze-based behaviors. Our overall approach allows the user to actively interact with tens of virtual agents from a first-person perspective in an immersive setting. We systematically evaluate our inference algorithm in controlled multi-agent simulation environments and highlight its ability to reliably and efficiently infer the hidden intent of a user's avatar even under noisy conditions. We quantitatively demonstrate the performance benefits of our approach in terms of reducing false inferences, as compared to a prior method. The results of our user evaluation show that 68.18% of participants reported feeling more comfortable in sharing the virtual environment with agents simulated with our algorithm as compared to a prior inference method, likely as a direct result of significantly fewer false inferences and more plausible responses from the virtual agents.

Keywords: multi-agent simulation, virtual reality, avatars, human agents, interactive navigation

Index Terms: Human-centered computing—User studies; Human-centered computing—Virtual reality; Computing methodologies—Artificial intelligence; Computing methodologies—Motion path planning; Computing methodologies—Modeling and simulation;

*e-mail: sahil@cs.unc.edu
†e-mail: best@cs.unc.edu
‡e-mail: dm@cs.umd.edu

1 Introduction

Many VR applications use human-like virtual agents that can elicit social responses from a user [7]. These applications typically allow a user to embody a virtual avatar, i.e. a digital representation of the user in the virtual world, and frequently arise in training tasks, architectural walkthroughs, games, and VR based therapy for crowds phobias, social anxiety, PTSD treatments, etc. Moreover, in social VR applications, it is important to develop capabilities that allow a user to actively interact with other users and virtual agents in shared virtual environments [10, 42].

Modeling cognition and endowing the virtual agents with social capabilities to facilitate natural avatar-agent interactions remains a key challenge in such shared virtual environments. Prior studies in human behavior and social psychology have established the role of non-verbal *social cues* such as proxemics and gazing and the meaningful interpretations thereof, or *social signaling* [15] in conveying one's intent [11, 13, 20, 40]. Moreover, psychologists have argued that humans use a *Theory of Mind* through which they attribute mental states such as beliefs, intents, desires etc. to others, and rely on these attributes to explain and predict their behaviors [43]. To facilitate natural interactions with the user's avatar, virtual agents must also be capable of applying such socio-cognitive reasoning to reliably infer a human's intent, and be able to communicate their own intent using social signals.

In addition to inferring the intent of the user, the virtual agent must respond using appropriate social signals. Research in human-agent interactions has shown that agents incapable of effectively exhibiting human-like behaviors can negatively impact the user's perception and overall task performance [52]. However, current avatar-agent interaction algorithms are often unable to generate appropriate social signals due to limitations in intent inference. Most prior work in avatar-agent or multi-agent interactions mostly relies on instantaneous social cues such as relative distance, velocity and orientation [10, 37, 44, 47]. These methods do not account for the causal relationship between mental states and observed actions. As a result, they are limited in their ability to reason and infer the

intent of the human's avatar. More complex models that account for mental state attribution and higher order reasoning have also been proposed [4]. However, the theoretical foundation of these models necessitates significant domain and task knowledge for practical implementations, and thereby limits their application to avatar-agent virtual environments. A recent approach, called *Bayesian Theory of Mind* (BToM) [6], proposes a causal, probabilistic model that integrates observed social cues with statistical priors over the agent's mental states. BToM models the Theory of Mind concept and inverts the planning for a rational agent to draw Bayesian inferences over the agent's hidden mental states. BToM has been successfully applied to study one-to-one human-agent interactions in simple discretized 2D environments [6, 31]. One of our goals is extend the BToM concept so that it be applied to simulate interactions with a user's avatar in a multi-agent environment, wherein each agent is an independent entity capable of making its own decisions and executing actions such as locomotion to accomplish its goals [21, 26, 36, 58].

Main Results: In this paper, we address the problem of facilitating natural interactions between a user's avatar and multiple human-like virtual agents in complex virtual environments. To that end, we present *Agents with Theory of Mind* (ATOM), a real-time algorithm that enables virtual agents to perceive proxemics and gaze-based social cues and reliably infer the underlying hidden intention of the human. We assume that the intent of the user is either to engage the virtual agent in a stationary, face-to-face interaction, or to simply avoid it. Our algorithm (ATOM) is grounded in previous studies in human behavior and social psychology and offers the following benefits:

- ATOM applies the Bayesian Theory of Mind approach to infer the underlying intent of the user (Section 3). We maintain priors over the user's intentions and compute the posterior probability of intentions based on observed social cues. Our approach uses a novel gaze-based and proxemics-based human prediction model (Section 4). We also account for the dynamic nature of human intentions.

- ATOM accounts for the potential irrationality in human behavior, and is robust against noise.

- ATOM uses the inferred intent of the user to regulate the gaze of the agent. Our approach also enables the agents to accomplish their individual goals using locomotion.

- Our algorithm allows for the presence of a tracked real user in an immersive virtual environment, and can facilitate plausible avatar-agent interactions (Section 6).

We quantitatively evaluate ATOM's capability to accurately infer the underlying intent of the human by simulating a *procedurally controlled avatar*, and comparing against a prior inference method. Our results show that ATOM significantly reduces the rate of false inferences. We further evaluate the robustness of ATOM by varying the initial conditions and hidden intent, and adding noise to the actions of the controlled avtar (Section 5). Our results demonstrate the ability of ATOM to correctly infer the hidden intent of the simulated avatar even under noisy conditions. Combining motion and gaze cues reduces the average inference time by 2.12 seconds compared to motion cues only and increases robustness compared to gaze-based cues only. We also conduct a user evaluation in immersive settings and compare ATOM with a prior motion-based inference algorithm. Participants found that ATOM agents produce responses which are less intrusive and provide greater comfort in 68.18% of all responses, with a mean response of $2.95 \pm +1.759$ for comfort on a 7-point Likert scale.

2 Related Work

In this section, we highlight prior work in simulating plausible movements of multiple human-like virtual agents, as well as generating social interactions between agents and avatars.

2.1 Simulating Plausible Movements of Multiple Virtual Agents

Prior interactive simulation algorithms often decompose the problem of generating realistic movement and behavior of multiple human-like agents into a two-step process. The first step involves computing 2D collision-free trajectories using a simple representation for each agent, such as a bounding disc. This is then followed by generating full-body animation for each agent along its 2D trajectory.

2.1.1 2D Trajectory Computation

Most prior 2D crowd simulation techniques can be broadly classified as macroscopic models and microscopic models. Macroscopic models [57] tend to compute the aggregate motion of the crowd by generating fields based on continuum theories of flows. Microscopic models based on multi-agent methods compute trajectories for each individual agent. These use a combination of global [53] and local navigation methods [21, 26, 48, 58] to compute trajectories for each agent such that it avoids collisions with other agents and obstacles. Most of these methods only compute the trajectories of the agents in a 2D plane.

2.1.2 Human-like Motion Synthesis

There is extensive literature in computer graphics and animation on generating human-like motion [59]. We limit our discussion to data-driven, procedural, and physics-based methods. Data-driven methods such as motion graphs [14, 28] create a parameterized graph of blendable motions and apply traversal algorithms to generate trajectories. Such motion databases are often created through motion capture yielding human-like results. Procedural methods apply kinematic principles to generate motions adhering to biomechanic constraints [9]. Physics-based models seek to generate physically-feasible motions by computing actuator forces for each joint to generate the desired motion [23]. These methods generate physically correct motions but may not generate natural motions. Recently, several approaches have been proposed to blend data-driven and physics-based [29] or data-driven and procedural [24] methods.

2.1.3 Coupled 2D Navigation & Motion Synthesis

There are few methods that combine crowd simulation and motion synthesis into one framework. Shao et al. [49] propose an animation system that combines perceptual, behavioral, and cognitive control components to generate rich behaviours for virtual agents. Shapiro et al. [50] present a character animation framework that utilizes a 2D steering algorithm and a motion blending-based technique to generate visually appealing motion. ADAPT [25] combines an open-source navigation mesh and steering algorithm with a set of animation controllers. Multiple techniques have also been proposed for footstep-driven walk synthesis [51]. There is work in the robotics domain that addresses bi-pedal locomotion for multiple robots [41], though they are not fast enough for interactive applications.

2.2 Simulating Social Interactions

In this section, we provide a brief overview of social cues observed in human interactions, and their application in human-robot and avatar-agent interactions.

2.2.1 Social Cues in Human-Human Interaction

In addition to verbal communication, non-verbal social cues such as proxemics, eye gaze, gestures, body postures etc. play a crucial role in human-human interactions. Hall et al. [20] study proxemics in human interactions and postulate that the distance of approach can

convey different social meanings. Emery et al. [13] investigate the neurological processing of eye gaze and highlight its crucial role in non-verbal communication and social signaling. The role of gaze in social interactions has also been studied in the context of human behavior and social psychology. Recent studies have investigated the phenomena of *gaze following* i.e. the tendency of individuals to redirect their visual attention by following the gaze of others [55,56]. Goffman [17] propose a theory of *civil inattention* which states that people owe one another an initial glance followed by a withdrawal of attention. Moreover, one may indicate his or her intention to interact with another individual by violating the rule of civil inattention [11]. The role of eye gaze in communicating intent has also been studied in the specific case of locomotion. Nummenmaa et al. [40] show that gaze can reliably indicate an oncoming individual's intended movement. Our proposed algorithm, ATOM, builds on many of these theories to infer and respond to the human in mixed human-agent virtual environments.

2.2.2 Human & Robot Interaction (HRI)

There is a significant body of work on inferring intent and generating a response for a single robot interacting with a human. Most algorithms rely on perceived sensor information to learn and recognize a human's plan [39, 47] and often do not account for the causal relationship between mental states and observed actions. On the other hand, Breazeal et al. [8] propose a socio-cognitive architecture that enables an anthropomorphic robot to attribute beliefs, intents and desires to its human partner. Similarly, other socio-cognitive architectures have been proposed and are widely used in computational AI [4]. Many existing approaches have also explored the role of eye gaze in both inferring human actions and generating appropriate social signals for the robot [2, 15].

2.2.3 Avatar & Virtual Agent Interaction

Virtual reality-based platforms are increasingly being used to systematically study interactions between multiple humans (with or without avatars) [35, 60], as well as interactions between a human and one or many virtual agent(s). Moussaid et al. [34] demonstrate that a immersive multi-user virtual environment can be used to effectively study and analyze crowd behavior in high stress evacuation scenarios. There is also extensive prior work on embodied conversation agents [12], in which a *single* animated anthropomorphic agent interacts with a user. Recent studies have shown that anthropomorphic agents can engage the user in social interactions in immersive settings [7,52]. Aravena et al. [5] studied interactions between a user and an embodied agent in a virtual classroom and concluded that students who seek to cheat in the classroom may apply the concepts of the Theory of Mind to infer the hidden intent of the teacher. There is also recent work in simulating shared avatar and multi-agent virtual environments. These include algorithms to generate locomotion-based interactions between agents and avatars [36], studies on the impact of human-agent collision avoidance [3, 30, 54], dynamic group behaviors [45], and navigational decision making [10]. Studies have also investigated proxemics in immersive settings in relation to agents [32] and virtual obstacles [46]. Some algorithms seek to simulate agents capable of exhibiting social signals. These include algorithms that simulate gaze-based behaviors using pre-defined *interest points* [19], or gazing location, as well as dynamic interest points generated using a attention model [18]. Recent methods have also sought to simulate face-to-face avatar-agent interactions using gaze [37] and even head movements [44]. However, most of these methods tend to rely on the instantaneous relative velocity and the instantaneous gaze of the human subject to infer his/her hidden intent. The lack of mental state attribution and higher order reasoning makes them susceptible to noise and false interpretations, which can in turn affect the human-agent interaction.

3 Overview

In this section, we introduce the notation and terminology used in the rest of the paper and give an overview of our approach.

3.1 Notation and Assumptions

We denote a scalar variable n with lowercase letters, a vector $\mathbf{x}$ with a bold faced lower case letter, and a set $\mathcal{C}$ of entities with an uppercase calligraphic letter. We use the subscript V-i to denote the ith virtual agent and the subscript H to denote a user. Our approach is designed for multi-agent algorithms, where each virtual agent is modeled as an independent discrete entity, capable of planning and acting on its own. It also has an associated bounding disk of radius r_{V-i} in $\mathbb{R}^2$ space which is equal to half of the shoulder width of the skeletal mesh and used for 2D multi-agent navigation. At any time t, the instantaneous 2D position and velocity of the agent are given as $\mathbf{p}^t_{V-i}$ and velocity $\mathbf{v}^t_{V-i}$, respectively. Similarly, the time-varying trajectory is represented using $\mathbf{p}^{0:t}_{V-i}$ and velocity $\mathbf{v}^{0:t}_{V-i}$. Each virtual agent is capable of full-body locomotion and gazing. The gaze of the agent at time t is represented by a 2D unit vector $\mathbf{h}^t_{V-i}$.

We also assume the presence of a user avatar, represented by a skeletal mesh and associated 2D disk. The time varying position, velocity, and gaze directions of the user are denoted as $\mathbf{p}^{0:t}_U$,$\mathbf{v}^{0:t}_U$ and $\mathbf{h}^{0:t}_U$, respectively. The environment consists of complex obstacles projected to $\mathbb{R}^2$ space and represented as 2D obstacles for multi-agent planning. The union of all entities in the scene, including obstacles, virtual agents, and the user, comprises the simulator state $\mathcal{S}$.

3.2 User Interaction

Our approach allows for the presence of a user in the multi-agent virtual environment. The user is provided with a first person view in an immersive setting, and his/her tracked movements are used to animate a skeletal mesh in the shared virtual environment. Our approach is general and agnostic of the specific method used to track the user.

3.3 Modeling Human-Agent Interactions using ATOM

Our algorithm ATOM facilitates plausible interactions between the user and virtual agents. We assume that the underlying intent of the user is to either engage the agent in stationary, face-to-face interactions or to simply avoid the agent. These comprise the set of intentions, $\mathcal{I} = \{\text{INT}, \text{AVD}\}$. Each virtual agent independently perceives social cues from the user, and applies the following theories in human behavior and social psychology to infer the hidden intent of the user:

- **Theory of Mind:** Humans tend to attribute mental states such as beliefs, desires, intent etc. to others, and rely on this attribution to reason, infer and predict the behavior of others [43].
- **Role of gaze in conveying intent:** Individuals tend to follow the rule of civil inattention [17] i.e. they withdraw visual attention away from the pedestrian and instead direct it towards their intended travel direction [40]. Violation of civil inattention can indicate the individual's intent to engage the user in a face-face interaction [11, 47].
- **Role of motion & proxemics in conveying intent:** The individual's distance of approach can also convey an intent to interact [20].

More details on how ATOM models these concepts to infer the intent of the user and guide the response of the virtual agents can be found in Section 4.

3.4 Shared Avatar & Multi-agent Simulation

At every time-step, we account for the tracked movement of the user to animate his/her skeletal mesh and synchronize the corresponding 2D disk. We also synchronize the skeletal mesh of each virtual agent with its corresponding 2D bounding disk. We then use a Behavioral Finite State Machine to map the simulator state $\mathcal{S}^t$ at time t to a goal position $\mathbf{g}^t_{V-i}$ for each virtual agent i in the simulation.

Next, each virtual agent perceives the user's tracked movement only if the avatar is deemed as *visible* to the agent. Visibility is conditioned on distance, field of view as well as geometric visibility determined by ray casting. For the sake of computational complexity, we do not consider partial visibility, restricting the visibility queries to two dimensional space. Moreover, visibility queries only consider obstacles and do not account for dynamic obstacles or other agents. In case the avatar is determined to be visible, the agent attempts to infer its intent (Section 4.1), and computes a response that can be a combination of locomotion and gazing (Section 4.2). We use a 2D navigation algorithm coupled with a full-body animation system to animate the virtual agents.

4 ATOM Algorithm

In this section, we provide details of our algorithm wherein each agent independently perceives the tracked movement of the user, uses social cues such as proxemics and gaze, and applies the Bayesian Theory of Mind approach to infer the hidden intent of the user.

4.1 Inferring Intent of User using Bayesian Theory of Mind

A virtual agent V tracks the user U under the condition that the user is visible with respect to obstacles. For the purpose of discussion, we assume that the user was tracked by the agent starting at time $T=0$ and denote the tracked position, velocity, and gaze direction of the user until time $T=t-1$ as $\mathbf{p}_U^{0:t-1}$,$\mathbf{v}_U^{0:t-1}$ and $\mathbf{h}_U^{0:t-1}$, respectively. We also assume that the hidden intention of the user is to either engage the agent in a stationary face-to-face interaction, or to avoid the user. These intentions are represented as $\mathcal{I}=\{\text{INT},\text{AVD}\}$. Let $P(\text{INT}|\mathbf{p}_U^{0:t-1},\mathbf{h}_U^{0:t-1})$ and $P(\text{AVD}|\mathbf{p}_U^{0:t-1},\mathbf{h}_U^{0:t-1})$ denote the agent's prior beliefs over the user's intentions. We use the Bayesian Theory of Mind approach to compute the posterior beliefs $P(\text{INT}|\mathbf{p}_U^{0:t},\mathbf{h}_U^{0:t})$ and $P(\text{AVD}|\mathbf{p}_U^{0:t},\mathbf{h}_U^{0:t})$ based on new observations $\mathbf{p}^t_U$,$\mathbf{v}^t_U$ and $\mathbf{h}^t_U$ at time $T=t$. Using Baye's rule, the posterior probability can be determined as:

$$\begin{aligned}&P(\text{INT}|\mathbf{p}_U^{0:t},\mathbf{h}_U^{0:t})\\&\propto P(\mathbf{p}^t_U,\mathbf{h}^t_U|\text{INT},\mathbf{p}_U^{0:t-1},\mathbf{h}_U^{0:t-1})P(\text{INT}|\mathbf{p}_U^{0:t-1},\mathbf{h}_U^{0:t-1})\end{aligned}\quad(1)$$

Based on the Markov assumption, the probability of the system state at time t is a function of the state at time $t-1$ and is independent of prior states. The assumption greatly reduces the complexity of the likelihood function as:

$$P(\mathbf{p}^t_U,\mathbf{h}^t_U|\text{INT},\mathbf{p}_U^{0:t-1},\mathbf{h}_U^{0:t-1})=P(\mathbf{p}^t_U,\mathbf{h}^t_U|\text{INT},\mathbf{p}^{t-1}_U,\mathbf{h}^{t-1}_U)\quad(2)$$

Moreover, we assume conditional independence between the user's gaze and trajectory, i.e.

$$P(\mathbf{p}^t_U,\mathbf{h}^t_U|\text{INT},\mathbf{p}^{t-1}_U,\mathbf{h}^{t-1}_U)=P(\mathbf{p}^t_U|\text{INT},\mathbf{p}^{t-1}_U)P(\mathbf{h}^t_U|\text{INT},\mathbf{h}^{t-1}_U)\quad(3)$$

Based on these assumptions, Equation 1 simplifies to:

$$\begin{aligned}&P(\text{INT}|\mathbf{p}_U^{0:t},\mathbf{h}_U^{0:t})\\&\propto P(\mathbf{p}^t_U|\text{INT},\mathbf{p}^{t-1}_U)P(\mathbf{h}^t_U|\text{INT},\mathbf{h}^{t-1}_U)P(\text{INT}|\mathbf{p}_U^{0:t-1},\mathbf{h}_U^{0:t-1}).\end{aligned}\quad(4)$$

Similarly, one can define $P(\text{AVD}|\mathbf{p}_U^{0:t},\mathbf{h}_U^{0:t})$ and normalize the two to compute the actual probabilities.

The term $P(\mathbf{p}^t_U|\text{INT},\mathbf{p}^{t-1}_U)$ in Eq. 4 represents the likelihood of observing the user at position $\mathbf{p}^t_U$ given the user's past position as $\mathbf{p}^{t-1}_U$ and under the assumption that the user seeks to interact with the agent. Essentially, it models an interpretation of the user's motion and proxemics-based cues in conveying his/her intent. Similarly, $P(\mathbf{h}^t_U|\text{INT},\mathbf{h}^{t-1}_U)$ models an interpretation of the user's gaze in conveying his/her intent. Both models are grounded in theories of human behavior and social psychology. We formalize these models as a reward optimization problem and provide details below.

Interpreting motion & proxemics-based cues

We assume that the user, when located at position $\mathbf{p}^{t-1}_U$ at time $T=t-1$, could have assumed any 2D velocity $\mathbf{v}^{t-1}_U$ such that $\|\mathbf{v}^{t-1}_U\|<s^{MAX}_U$, where s^{MAX}_U denotes the maximum permissible speed. Moreover at time $T=t-1$, the user observed the agent at position $\mathbf{v}^{t-1}_V$. Let $\mathbf{v}^{t-1}_{U*}$ denote the optimal approach velocity such that it would have minimized the distance to the agent, i.e. $\|\mathbf{p}^{t-1}_V-\mathbf{p}^{t-1}_U\|$. Given the actual observed velocity $\mathbf{v}^{t-1}_U$, we define a scalar term ψ as:

$$\psi=\max\left(\frac{\mathbf{v}^{t-1}_U.\left(\frac{\mathbf{v}^{t-1}_{U*}}{\|\mathbf{v}^{t-1}_{U*}\|}\right)}{s^{MAX}_U},10^{-6}\right).\quad(5)$$

As in prior work in Bayesian Theory of Mind [31], we use the Boltzmann policy to model movement-based behavior of the human i.e.

$$P(\mathbf{p}^t_U|\text{INT},\mathbf{p}^{t-1}_U)\propto e^{\beta_m\psi},\quad(6)$$

where β_m represents a "motion rationality index". Similarly, we can define the probability of observing the position $\mathbf{p}^t_U$ if the assumed intention was to avoid the agent as:

$$P(\mathbf{p}^t_U|\text{AVD},\mathbf{p}^{t-1}_U)\propto e^{\beta_m(1-\psi)}.\quad(7)$$

In Equations 6 & 7, β_m accounts for the irrational or suboptimal human movement. As $\beta_m\to\infty$, the motion model tends to expect perfectly rational human movement-based behavior. On the other hand, $\beta_m=0$ implies that any random movement is equally likely. Finally, we can compute absolute probabilities in Equations 6 & 7 by normalizing each using the sum of the two probabilities.

Hall [20] in his seminal work on proxemics and human behavior postulates that interpersonal distance can convey the type of social interaction, and he identified four distance zones: Intimate, Personal, Social, and Public distance. In the public distance zone, defined spatially as the region where interpersonal distance is greater than 3.7m, humans are close enough to notice each other but far enough away to not have face-to-face communication. In this zone, the information gain from motion is rather limited. For example, the user may momentarily deviate from the direct path to the agent with which he or she intends to interact, perhaps to avoid collisions with other agents and obstacle. This is commonly observed in dense scenarios. To avoid penalizing such behavior, we set $P(\mathbf{p}^t_U|\text{INT},\mathbf{p}^{t-1}_U)=P(\mathbf{p}^t_U|\text{AVD},\mathbf{p}^{t-1}_U)=0.5$ when the user is in the public zone with respect to the agent, and allow gaze to play the dominant role in such cases. Moreover, we also rely only on gaze-based cues in cases where the velocity of the user is zero.

Interpreting gaze-based cues

Based on prior studies in human behavior, we assume that a user will follow the rule of civil inattention by directing his/her gaze away from the agent [17,40], and that a violation of this rule indicates a desire to engage in a face-face interaction with the agent [11,47]. Moreover, we assume that the user is capable of altering his/her gaze instantaneously i.e. $P(\mathbf{h}^t_U|\text{INT},\mathbf{h}^{t-1}_U)=P(\mathbf{h}^t_U|\text{INT})$. Let $\mathbf{d}^t_{VU}=$

$\mathbf{p}_V^t - \mathbf{p}_U^t$ denote the displacement of the agent with respect to the user at time $T = t$. We define a function $f_{g-INT}(\mathbf{h}_U^t, \mathbf{d}_{VU}^t)$ such that:

$$f_{g-INT}(\mathbf{h}_U^t, \mathbf{d}_{VU}^t) = \begin{cases} R_g & \text{if } \frac{\mathbf{h}_U^t}{\|\mathbf{h}_U^t\|} \cdot \frac{\mathbf{d}_{VU}^t}{\|\mathbf{d}_{VU}^t\|} > \cos\theta_{GI} \\ 10^{-6} & \text{otherwise} \end{cases}, \quad (8)$$

where R_g and θ_{GI} denote a scalar reward, and angular threshold respectively. As such, the reward function depends only on the current gaze of the user and does not account for the prior gaze. This simplifies the computation but also makes it more prone to instability, or fluctuations, in the presence of noise. Moreover, the discrete reward function equally rewards any gaze within an angular span of $\pm\cos\theta_{GI}$ with respect to the displacement vector $\mathbf{d}_{VU}^t$.

As with the human motion model, we use the Boltzmann soft-max policy to model the probability of observing gaze $\mathbf{h}_U^t$ as:

$$P(\mathbf{h}_U^t | \text{INT}, \mathbf{h}_U^{t-1}) \propto e^{\beta_g f_{g-INT}(\mathbf{h}_U^t, \mathbf{d}_{VU}^t)}, \quad (9)$$

where β_g denotes the "gaze rationality index" and has a similar interpretation to the motion rationality index.

We also define a function f_{g-AVD}, which rewards avoidance behavior by simply swapping the reward allocation of f_{g-INT}. Similar to Eq. 9, we use f_{g-AVD} to compute $P(\mathbf{h}_U^t | \text{AVD}, \mathbf{h}_U^{t-1})$. Then, we can compute $P(\text{INT} | \mathbf{p}_U^{0:t}, \mathbf{h}_U^{0:t})$ in Equation 4 using Equations 6 and 9. Finally, we infer an intention of the user to engage the agent in a face-to-face interaction if $P(\text{INT} | \mathbf{p}_U^{0:t}, \mathbf{h}_U^{0:t}) \geq \alpha_i$, where α_i is a pre-defined threshold that is experimentally determined.

4.2 Generating response of virtual agent

As described in Section 3, each agent has an independent goal and is capable of locomotion and gazing. The agent uses a velocity-space reasoning algorithm to compute a collision-free velocity with respect to visible entities in its neighborhood, including other agents and the user. It employs the inference algorithm to determine the user's intent. The inferred intention is used to guide the gaze-based response of the agent as:

$$\mathbf{h}_V^t = \begin{cases} \frac{\mathbf{d}_{UV}^t}{\|\mathbf{d}_{UV}^t\|} & \text{if } \mathbf{f}_V^t \frac{\mathbf{d}_{UV}^t}{\|\mathbf{d}_{UV}^t\|} > \cos\theta_{GR} \\ \frac{\mathbf{v}_V^t}{\|\mathbf{v}_V^t\|} & \text{otherwise} \end{cases},$$

where $\mathbf{f}_V^t$ denotes the forward facing unit vector of the virtual agent, and θ_{GR} denotes the maximum angular threshold for a gaze-based response. We set this threshold θ_{GR} such that it is less than the inference threshold θ_{GI} to reflect the fact that humans employ peripheral vision to observe their surroundings.

5 Implementation & Results

In this section, we provide details on the implementation of our algorithm ATOM and quantitatively evaluate its performance over a set of challenging benchmarks, and in the presence of simulated noise.

5.1 Implementation

We have implemented our algorithm in C++ on a desktop PC with an Intel Xeon E5-1620 v3 4-core processor, 16 GB of memory, and Windows 10 OS. We have integrated it with the Unreal Engine to enable real-time user-agent interactions and an animation system for full body motion simulation [50]. We use the HTC Vive headset and controllers to track the movement of the user in a $3.8 \times 3.8m^2$ obstacle-free space, enable object manipulation, haptic feedback and to provide a first person view. Moreover, we rely on the head orientation obtained by tracking the HMD to serve as a substitute for the gaze of the user. This assumption is based on a recent study which showed that both head orientation and gaze offer similar performance in terms of predicting the user's destination in goal-directed travel [16]. Table 1 presents the parameter values used in the ATOM algorithm. Overall, our approach can simulate and render 60+ full-body agents at 60+ fps.

β_m	β_g	$\cos\theta_{GI}$ (deg)	$\cos\theta_{GR}$ (deg)	α_i	R_g	s_U^{MAX} (m/s)
0.1	0.05	60	30	0.8	1	1.5

Table 1: Parameter values used in ATOM. The values defined in this table were used for all the experiments described in this paper.

5.2 Quantitative Evaluation

We demonstrate the performance of our algorithm ATOM in terms of inferring the avatar's intentions on several benchmarks. In each benchmark, we simulate a *procedurally controlled avatar* in order to quantitatively evaluate the simulation under controlled conditions. The avatar is tasked with engaging a pre-determined *target* agent in a face-to-face interaction. As a result, the avatar attempts to move towards the *target* agent while avoiding collisions with other agents. Moreover, we set the initial probabilities such that each agent assumes equal likelihood of interaction and avoidance on part of the avatar.

5.2.1 Metrics

We evaluate the inference capability of ATOM in terms of the following metrics:

- **False Detection Rate:** We compute the number of *other* agents, i.e. all virtual agents except for the *target* agent, that falsely inferred an intention to interact on part of the avatar at any point during the simulation. By definition, a high false detection rate implies inaccurate inference.

- **False Inference Interval:** For each of the *other* agents, we also keep a track of the largest contiguous time interval during which it falsely inferred an intention to interact. At the end of the simulation, we aggregate these time intervals over all agents to compute a 4-dimensional tuple denoting the minimum, maximum, mean and standard deviation of false inference intervals.

- **Target Inference Detection Time:** This denotes the earliest instance in the simulation at which the *target* agent correctly inferred an intention to interact.

- **Target Inference Settled Time:** We measure the settling time as the latest time in the simulation after which the target agent consistently inferred an intention to interact. Ideally, the inference settled time should be equal to the inference detection time i.e. the *target* agent should consistently infer an intention to interact starting from the inference detection time to the end of the simulation. However, the inference may not be robust due to a number of factors, including but not limited to, collision-avoidance behaviors and tracking noise.

5.2.2 Benchmarks

We evaluate ATOM on the following challenging benchmarks and present the results in Table 2.

- **Standing Agents:** The procedurally controlled avatar is initialized between two columns of *other* agents. Each column comprises of 5 stationary agents facing each other. The *target* agent approaches the avatar head-on (Figure 2(A)). As shown in Table 2, ATOM dramatically reduces the false detection rate from 100% to just 10%. Moreover, the only agent that falsely

Figure 2: **Benchmark Environments.** In each scenario, the user, by means of his/her avatar, is tasked with engaging a *target* agent in a face-to-face interaction. For the sake of visual clarity, we render a third person perspective which clearly depicts the user's avatar and also highlight the *target* agent with a red arrow. The figure depicts the following benchmarks: (A) *Standing Agents*, (B) *Anti-podal Circle*, (C) *Crossing Flow*, and (D)*Shibuya Crossing*. More details can be found in Section 5.2.2.

inferred interaction in case of ATOM has a lower false inference interval (0.243 sec) than any of the *other* agents using PedVR. ATOM also has a lower inference detection time for the *target* agent as compared to PedVR. Overall, these metrics collectively highlight the inference accuracy of ATOM as compared to PedVR.

- **Crossing Flow:** The avatar faces a group of seven agents approaching it head-on. The agents are initialized such that the *target* agent is placed directly opposite to the avatar, and is flanked on each side by three *other* agents (Figure 2(C)). As in the prior benchmark, only one agent falsely inferences an intent to interact in case of ATOM while all of the *other* agents falsely infer interaction in case of PedVR (Figure 3). The relative accuracy of inference with ATOM is further evidenced by the significantly low false inference interval.

- **Anti-podal Circle Crossing:** The avatar and seven *other* agents are placed on the circumference of a circle. The *target* agent is situated diametrically opposite to the user and is flanked by three *other* agents on both sides (Figure 2(B)). All agents are initialized with diametrically opposite goals which leads to significant avatar-agent interactions especially near the congested center of the circle. This scenario is especially challenging as the avatar is in the direct path of all other agents as they pass through the center of the circle. As shown in Table 2, three *other* agents falsely infer an intention to interact in case of ATOM as compared to all six *other* agents with PedVR. Moreover, ATOM results in a significantly lower inference detection time for the *target* agent.

- **Shibuya Crossing:** We simulate a busy street crossing, where 48 agents are assigned goal positions randomly and use the pedestrian walking lanes to navigate (Figure 2(D)). The scenario comprises of various obstacles at the end of the pedestrian lanes including subway stations, barriers, traffic light anchors etc. The agents rely on the underlying 2D navigation algorithm to compute collision-free trajectories. Moreover, they use visibility queries to determine if the avatar is *visible*, as described in Section 3. The avatar is initialized at one end of a crosswalk while the *target* agent is the last one of a dense group of *other* agents at the opposite end of the crosswalk. This scenario is also quite challenging as the avatar must navigate through a dense group of oncoming agents, which increases the chance of false positives. Yet, ATOM yields a significantly lower false inference rate as compared to PedVR and even results in lower inference detection time for the *target* agent.

These scenarios were designed to maximize the potential for the avatar's interactions with the *other* agents, thereby increasing the probability of false positives. Yet, ATOM results in a significantly lower false positive rate in all scenarios. Moreover, the *other* agents that falsely infer interaction tend to do so for a shorter time interval in case of ATOM as compared to PedVR. ATOM also results in a lower target inference detection time for the *target* agent i.e. when using ATOM, the *target* agent takes less time to correctly infer the avatar's intentions. This is largely a by-product of PedVR's limited inference capability which forces it to narrow the thresholds defined in [37]. In contrast, ATOM applies more nuanced reasoning which makes it less prone to false positives even with higher thresholds for distance and field of view. It should also be noted that in case of both ATOM and PedVR, the inference settled time for the *target* agent was the same as the inference detection time. While this was expected in case of PedVR owing to its design, it does lend more credence to the robustness of ATOM.

5.3 Robustness

We evaluate the performance of ATOM in reliably inferring the user's hidden intent in noisy conditions by simulating a *procedurally controlled avatar* and adding artificial noise. Moreover, we evaluate the isolated impact of gaze and motion-based cues and compare them with the coupled inference model (Eq. 4).

Scenario: Figure 4(A) illustrates the virtual scene. The simulated avatar and an agent ('VA-1') are initially placed on the horizontal axis $12m$ away from each other, with the origin located in the middle. Two other agents, 'VA-2' and 'VA-3' are initialized at a ± 30 degree angular offset from VA-1. The initial probabilities are set such that VA-1 infers an intent to interact whereas the other agents infer an intent to avoid. We set the actual goal of the avatar to be VA-2. We also add varying levels of noise to the movement and gaze of the agent, denoted by α_m and α_g respectively. In case of gaze-based noise, $\alpha_g = 1$ implies that the simulated avatar can randomly assume any gaze direction instantaneously, whereas $\alpha_g = 0$ assumes that the avatar consistently gazes in the direction of travel. Similarly, speed-based noise $\alpha_m = 1$ implies that the simulated avatar can randomly assume any speed between 0 and its desired speed, whereas $\alpha_m = 1$ implies the avatar always attempts to move at its desired speed subject to collision-avoidance constraints.

Intention Inference Time: We measure the intention inference time t_d as the earliest instance at which both: VA-1 infers that the controlled-avatar intends to avoid it; and VA-2 infers the controlled-avatar intends to interact with it (Figure 4(B)). Thus, a smaller t_d implies more efficient and accurate inference for VA-1 and VA-2. However, it does not reflect the inference made by other neighboring agents (e.g. VA-3).

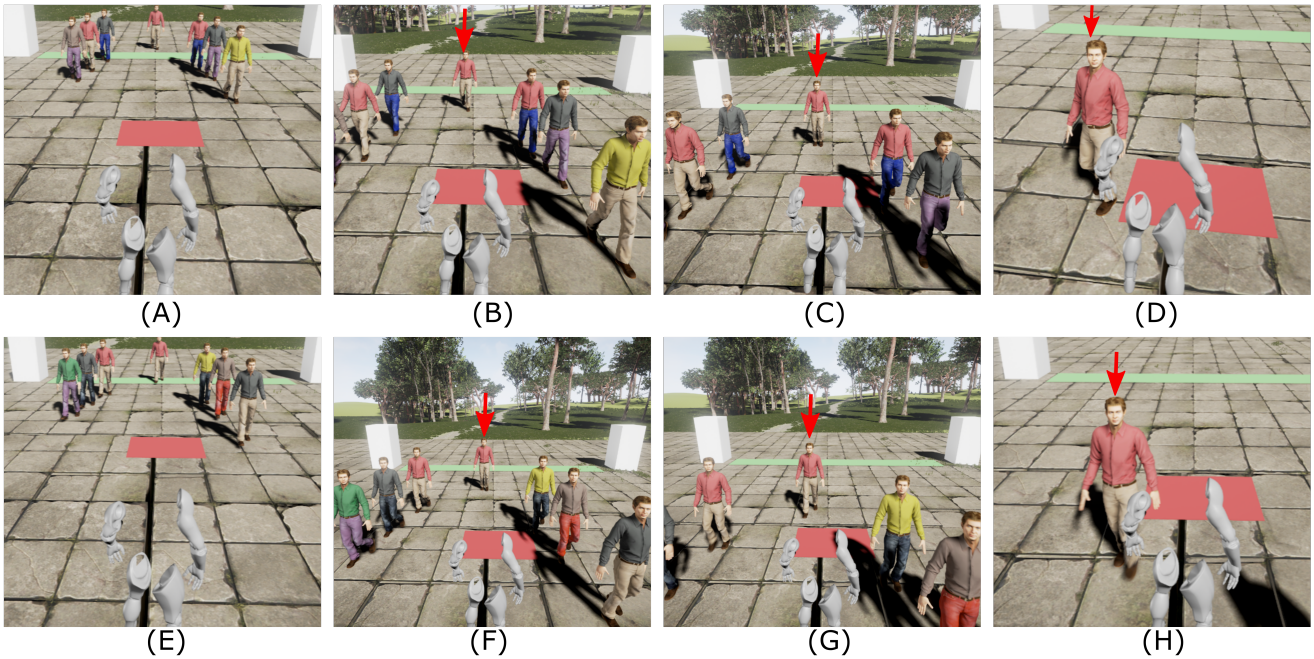


Figure 3: **Comparison of ATOM with PedVR on Crossing Flow benchmark.** In this scenario, the user, by means of his/her avatar seeks to interact with the *target* agent. The *target* agent is rendered with a red arrow for visual clarity. The *target* agent (red shirt) is flanked by three *other* agents on each side and is initialized directly opposite to the avatar. The avatar is depicted using a third person perspective for visual clarity. (A-D) Agents are simulated with our novel Bayesian Inference-based algorithm to infer the hidden intent of the user. (B-C) The *other* agents correctly infer that the user does not intend to interact with them. (D) The *target* agent infers an intent to interact and gazes at the avatar. (E-H) Agents simulated with a prior method, InstantGaze. Due to limitations in its inference capability, the *other* agents incorrectly assume an intention to interact and gaze at the avatar. Our user evaluation suggests that such false inferences and the subsequent behavior can induce a feeling of discomfort in the users.

Figure 5 depicts the intention inference time under varying conditions of noise and rationality factor. As described in Section 4.1, a rationality index of zero implies that the controlled-avatar is equally likely to assume any velocity and gaze direction, and its actions are independent of the underlying intent. On the other hand, a high rationality index implies that the controlled-avatar's actions are highly correlated with its underlying intent. Thus, increasing the rationality index decreases the intention inference time for a well behaved avatar. However, it also makes the inference algorithm more prone to noise and can lead to unstable and inaccurate inferences. Figure 5(A) depicts the impact of just gaze-based cues on the inference time. As expected, the inference time decreases with increasing gaze rationality index β_{GI}. Moreover, the inference time generally tends to increase with an increase in noise. Next, we evaluate the impact of just motion-based cues and again find a inverse relationship between inference time and motion rationality index (Figure 5(B)). However, the fall off is not as rapid compared to gaze-based cues. This is likely because our model for interpreting motion cues (Eq. 5) does not differentiate between motions that lead the controlled-agent to move in the direction opposite to the agent. Moreover, we find that the average detection time for motion-based cues ($t_d = 6.453s$) is much higher than the average for gaze-based cues ($t_d = 1.2324s$). Finally, we couple both gaze and motion-based cues and analyze the results under varying motion-based parameters and fixed gaze-based parameters ($\beta_{GI} = 0.06, \alpha_G = 0.05$)(Figure 5(C)). We find that coupling the two cues reduces the average detection time $t_d = 4.333$ as compared to just motion-based cues. Moreover, it was observed that coupling the cues reduced the instability in inference caused by rapid changes in gaze.

5.4 Comparison with Prior Methods

Our approach to intent inference is comparable to prior works that have applied the Bayesian Theory of Mind approach infer a human user's hidden intent. However, these approaches are largely restricted to 2D domains with discretized actions [6, 31]. In contrast, our work focuses on avatar and multi-agent interactions in more complex environments with a continuous action space.

There is also existing work on simulating gaze based interactions between agents and avatars. Recently, Lynch et al. [33] conducted a user study to evaluate the impact of gaze in collision-avoidance and found that a virtual agent's gaze behavior did not have any significant impact on a user's trajectory in a head-on approach. They suggest that in such a scenario, body motion cues may be sufficient to regulate and coordinate the interaction. Our work is complimentary since we focus on the role of the user's gaze and motion in terms of conveying his/her intent. As part of our future work, we can incorporate their findings to augment the response of the virtual agents. Grillion et al. [18, 19] propose multiple techniques to simulate gaze based behaviors in virtual crowds. Broadly, their approach focuses on regulating the gaze of the virtual agent by identifying *interest points* or gazing locations, as well as applying a IK-based solver to animate the gaze subject to spatial and temporal constraints. Owing to their objective of generating diverse crowd behaviors, their approach incorporates a number of factors to determine the gazing location for each agent but is not particularly suited for interpreting the user's intentions. In contrast, ATOM is user-centric in its design and is most effective at inferring the user's intentions. In the future, we would like to augment ATOM with the gaze regulation mechanism proposed in [18] to enable gazing at other virtual agents. We would also like to augment the gaze regulation model to account for *gaze following* [55, 56].

Scenario	Method	Num. Agents	Num. Agents w. False Inference(%)	False Inference Interval (sec) min	max	mean	std dev	Target Inference Detection Time (sec)
Standing Agents	PedVR	11	10(100%)	4.208	7.849	5.627	1.477	7.355
	ATOM		1(10%)	0.243	0.243	0.243	0	4.243
Crossing Flow	PedVR	7	6(100%)	2.255	2.560	2.419	0.105	3.665
	ATOM		1(16.66%)	0.738	0.738	0.738	0	3.783
Antipodal Circle	PedVR	7	6(100%)	1.535	3.087	2.481	0.478	4.255
	ATOM		3(50%)	0.936	3.160	2.419	1.048	3.472
Shibuya	PedVR	48	13(27.696%)	2.516	7.679	4.376	2.103	14.315
	ATOM		8(17.02%)	0.075	7.238	3.154	2.315	12.567

Table 2: **Quantitative evaluation of ATOM.** We evaluate ATOM's capability to correctly infer the user's intent in a number of challenging benchmarks, and contrast it with a prior method, PedVR [37]. The above table lists the number and fraction of *other* agents that falsely detected inference, an aggregate of these false inference intervals, and the inference detection time for the *target* agent. As can be seen, ATOM significantly reduces false inferences and in most cases, even lowers the inference detection time for the *target* agent. More details on the metrics and the results can be found in Section 5.2.

Figure 4: **Robust inference in the presence of noise.** To evaluate ATOM's ability to infer intent, we performed a repeated experiment using a procedurally controlled avatar under varying levels of motion and gaze noise. (A) The avatar faces a group of agents. The agent in the center assumes the avatar intends to interact with it. However, the hidden intent of the agent is to interact with the agent on the left, rendered with a red arrow for visual clarity. (B) As the avatar moves towards the agent on the left, all three agents dynamically update their inference. Details on the results are provided in Section 5.3 and Figure 5.

6 User Evaluation

In this section, we provide details of a within-subjects user study conducted to evaluate the impact of our algorithm, ATOM, compared to a prior velocity-based inference approach.

6.1 Experiment Goals & Expectations

ATOM is designed to reduce false inferences i.e. to reduce the false detection rate as well as false inference intervals. In doing so, it should lead to a more positive experience for a user immersed in a virtual environment that is shared with virtual agents. Thus, we hypothesized that the user will:

- notice the qualitative differences in the behavioral response of agents simulated with ATOM in comparison to those simulated with a baseline method,
- feel more "comfortable" in the virtual environment if the neighboring agents do not overtly stare at the user, unless he/she intends to interact with them, and
- feel more as part of the virtual crowd.

6.2 Experimental Design

The study was conducted based on a within-subjects, paired-comparison design. For each scenario, participants interacted with a pair of simulations in random order with a similar exposure time, conditioned on their behavior. Each pair of simulations comprised of ATOM and the baseline method. The participant's movement was tracked and used to simulate his/her virtual avatar. There was a one-to-one mapping between the physical space and the virtual world to ensure congruent proprioceptive cues. During exposure, participants were tasked with approaching a pre-determined *target* agent in an attempt to engage it in a face-to-face interaction. By design, the direct path from the participant to the *target* agent lay within the tracked physical space. During this approach, the participant would potentially encounter *other* virtual agents. After each pair of simulations, participants answered a set of questions before moving on to the next pair of simulations.

6.3 Procedure

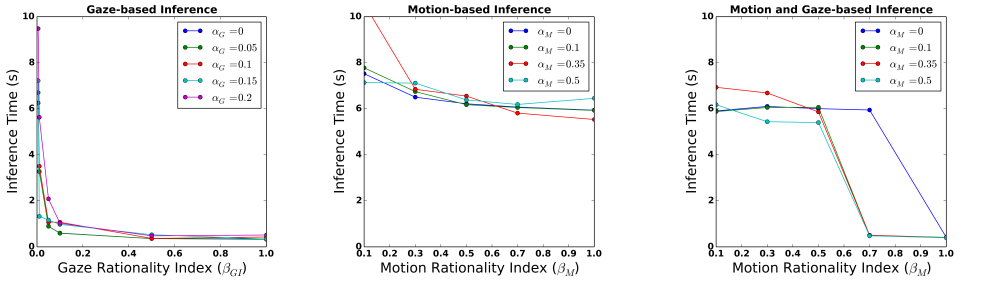
Participants were greeted and provided with informed consent documentation. They were then exposed to a training virtual environment to calibrate their height, and a second to familiarize themselves with the task. The second environment featured a *target* agent but no *other* virtual agents. The *target* agent was initially stationary. The participant was asked to move to an initial position, identified by as red colored tile on the virtual floor, and face the *target* agent. Once at the initial position, the participant was informed that his/her task was to approach the *target* agent and attempt to engage it in a face-to-face interaction. The participant was then instructed to start the simulation whenever he/she felt ready by pressing a trigger button on the hand-held controller. Unbeknown to the participant, the *target* agent would also start approaching his/her avatar once the simulation starts. The simulation was designed to end when the *target* agent was within 1 m of the participant.

At the end of the training simulations, the participants were informed that the testing scenarios may comprise of more than one virtual agent. However, their task remained the same i.e. to approach and engage the target agent. The participants were instructed that the *target* agent would always be centrally located in the group of virtual agents.

6.4 Scenarios

After training, participants experienced the *Standing Agents* and *Crossing Flow* benchmakrs, as described in section 5.2, in random pairwise order. Each of the two scenarios was designed such that:

- It was easy to identify the *target* agent.
- The direct path of the avatar from its initial position to the approaching *target* agent would pass between the two groups of *other* agents, thereby increasing the likelihood of false inferences.



(A) Gaze-based inference *(B) Motion-based inference* *(C) Gaze & Motion-based inference*

Figure 5: We highlight the impact of gaze and motion-based cues on the inference detection time for an agent, under varying conditions of noise. (A) With only gaze-based cues, the agent produces rapid inference even with a small gaze rationality but is also impacted by noise. (B) With only motion-based cues, inference is more robust to noise but is also slower in terms of time even with higher rationality. (C) Combining motion and gaze cues provides efficient and accurate inference even in the presence of noise.

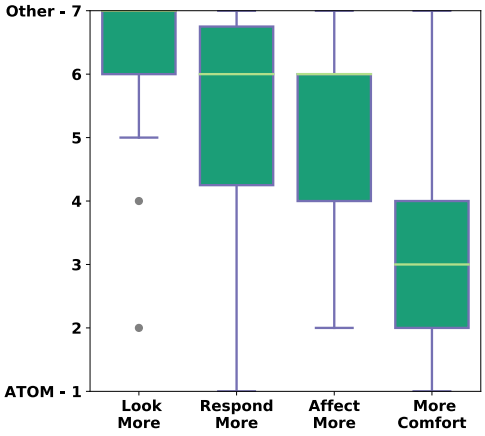


Figure 6: **Participant Preference in User Evaluation**: Participants identified clear differences between PedVR and ATOM in terms of level of agent response and comfort. Participants indicated PedVR agents responded more to them,$\bar{x}$ = 5.36, looked at the participant more, $\bar{x}$ = 6.14, and affected them more in their exploration $\bar{x}$ = 5.32. On the other hand, participants indicated greater comfort in sharing space with ATOM agents, $\bar{x}$ = 2.95. It is likely that the significantly high rate of false positives in case of PedVR causes the user to feel more uncomfortable as he/she is consistently stared at by neighboring agents. In contrast, ATOM agents tend to respond more appropriately, improving the overall experience.

- The direct path of the avatar from its initial position to the approaching *target* agent would lie within the physical tracking space.

6.5 Comparisons

We compared our method to PedVR [37], a recent interaction approach based on velocity-space reasoning. PedVR augments the ORCA [58] collision-avoidance algorithm with gazing behavior based on the instantaneous position and visibility of the avatar with respect to the agent. Both methods were coupled with the same 3D animation system [50] to generate full body motion for virtual agents.

6.6 Metrics

Participants were asked a set of questions designed to capture aspects of interaction with the virtual agents including the level and naturalness of the virtual agents' responses to the participant. They indicated their preference for a simulation using a 7-point Likert scale with 1 indicating strong preference for the simulation presented first, 7 indicating strong preference for simulation presented second, and 4 indicating no preference.

6.7 Results

Our study was taken by 11 participants, 8 male, with a mean age of 27.91 ± 2.11 years. To compensate for exposure order, we collapsed responses across scenes. We reserve discussion to the four questions for which significance was observed, and standardize responses such that 1 indicates preference for ATOM. For each question, a one-sample t-test was performed against a hypothetical mean of 4 (no preference). The question "In which simulation did you feel more comfortable in sharing the space with the virtual characters," was shown to be significant, $t(21) = -2.788, p = 0.011$. Responses to the question "In which simulation did the agents respond to you more?" were also significant, $t(21) = 4.101, p = 0.001$. The question "In which simulation did the characters seem to look at you more?" was shown to be significant, $t(21) = 7.808, p =< 0.000$, as well as "In which simulation did the presence of the virtual characters affect you more in the way you explored the space", $t(21) = 4.672, p =< 0.000$. Figure 6 and Table 3 provide further details on participant responses.

6.8 Discussion

As hypothesized. participants observed clear differences between ATOM and PedVR in terms of the behavioral response of the virtual agents. Participants indicated PedVR agents responded more to them, $\bar{x}$ = 5.36, and looked at them more, $\bar{x}$ = 6.14. This result can be attributed to the tendency of PedVR agents to falsely infer and intention to interact, which can cause neighboring agents to overtly gaze at the avatar. In contrast, ATOM relies on a more nuanced approach to infer the user's intent which decreases the likelihood of false positives, as evidenced in Section 5.2.

In addition to observing these differences, participants reported that PedVR agents affected them more in their exploration, $\bar{x}$ = 5.32, and they felt more comfortable sharing space with ATOM agents, $\bar{x}$ = 2.95. This can again be related to the rate of false positives. It is likely that by consistently gazing at the avatar in close proximity, PedVR agents tend to induce a feeling of discomfort. In contrast, ATOM generates a more plausible response wherein only the target agent consistently gazes at the user.

7 Conclusion, Limitations & Future Work

We present a novel approach (ATOM) to improve the interactions between an avatar and virtual agents in a shared virtual environment. Our approach builds on prior work on Bayesian Theory of Mind in

Question (In which simulation did...)	1	2	3	4	5	6	7	mean	SD
you feel more comfortable in sharing the space with the virtual characters?	5	5	5	4	1	0	2	2.95	$\pm$ 1.759
the characters seem to look at you more?	0	1	0	1	3	5	12	6.14	$\pm$ 1.283
the agents respond to you more?	1	0	1	4	4	6	6	5.36	$\pm$ 1.560
the presence of the virtual characters affect you more in the way you explored the space?	0	1	0	6	3	8	4	5.32	$\pm$ 1.323

Table 3: **Participant Response Frequency**: Participants prefer our method (ATOM) over a prior method on several dimensions. Participants reported feeling more comfortable in sharing the space with the virtual agents simulated with our algorithm, ATOM, as compared to a prior method.

psychology literature to make inferences about the avatar's hidden intentions. As a result of the Bayesian interpretation, ATOM's capability to infer the user's intent is robust in dealing with momentary deviations, for e.g. a user may deviate from the direct path to the intended *target* agent in order to avoid regions of high congestion. Likewise, ATOM is robust to sensor noise. Yet, it can still infer a change in the user's intentions. ATOM can be combined with interactive multi-agent and crowd simulation algorithms to generate locomotion and gaze-based behaviors. We quantitatively evaluate the performance of ATOM on a set of challenging benchmarks and show that it considerably improves the inference accuracy as compared to a baseline method. We also conducted an active user study in VR and observed that users tend to feel more comfortable in sharing the virtual environment with ATOM agents than with those simulated with prior methods.

Our approach, ATOM, has some limitations. While the agents are capable of navigating in multi-agent multi-avatar environments, their inference capability, and as a consequence their gaze-response, is limited to a single user. As such, they do not gaze at other agents. ATOM only accounts for a limited set of user intentions i.e. to either avoid the agent or engage it in a face-to-face interaction. ATOM also makes the simplifying assumption that a user's gaze and motion are conditionally independent and thus, fails to reflect the interdependence of the two cues [22]. Moreover, it only considers two cues for inferring the user's intent i.e. gaze and locomotion. The interpretation of gaze-based cues is also limited, as described in Section 4.1. ATOM is also limited in its ability to simulate a appropriate response from the virtual agents. The inferred intent of the user only affects the gazing behavior of the agent, and does not impact their locomotion or goals. Finally, we rely on the head orientation of the user to estimate his/her gaze.

There are many avenues for future work. In addition to overcoming these limitations, we would like to develop algorithms that enhance the inference and response capabilities of the agents. This includes taking into account factors such as the context of the simulation, high level behaviors of the avatars, gestures and other modalities, to draw a meaningful inference of the user's hidden intentions. We would like to extend our visibility queries to 3D space, and account for dynamic obstacles and the presence of other agents. Moreover, we would like to extend our algorithm to allow agents to infer and respond to multiple agents and avatars. We would also like to enhance the response of the agent and possibly include additional modalities such as head movements, gestures, facial expressions etc. We would like to build on findings from prior work on gaze allocation [18, 19, 55, 56] to generalize the gaze-based response of the agents. Khamis et al. [27] study smooth pursuit eye movements as a interaction mechanism and offer insights pertaining to VR than can guide our use case. We would like to use personalized avatars that mimic the user's appearance and walking style [38] which can potentially improve the user's engagement in the virtual environment. Finally, we would like to utilize inside-out trackers [1] to reliably track the user's gaze and conduct more extensive evaluation to study the effectiveness of our approach and use it for different applications.

ACKNOWLEDGMENTS

This work was partially supported by ARO grant W911NF-19-1-0069 and Intel.

REFERENCES

[1] Fove. 2014.
[2] H. Admoni and B. Scassellati. Social eye gaze in human-robot interaction: a review. *Journal of Human-Robot Interaction*, 6(1):25–63, 2017.
[3] J. Ahn, N. Wang, D. Thalmann, and R. Boulic. Within-crowd immersive evaluation of collision avoidance behaviors. In *Proceedings of the 11th ACM SIGGRAPH International Conference on Virtual-Reality Continuum and its Applications in Industry*, pp. 231–238. ACM, 2012.
[4] J. R. Anderson, M. Matessa, and C. Lebiere. Act-r: A theory of higher level cognition and its relation to visual attention. *Human-Computer Interaction*, 12(4):439–462, 1997.
[5] C. Aravena, M. Vo, T. Gao, T. Shiratori, L.-F. Yu, and E. Contributors. Perception meets examination: Studying deceptive behaviors in vr. In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society*, 2017.
[6] C. L. Baker, R. Saxe, and J. B. Tenenbaum. Action understanding as inverse planning. *Cognition*, 113(3):329–349, 2009.
[7] D. Bombari, M. S. Mast, E. Canadas, and M. Bachmann. Studying social interactions through immersive virtual environment technology: Virtues, pitfalls, and future challenges. *Frontiers in psychology*, 6, 2015.
[8] C. Breazeal, D. Buchsbaum, J. Gray, D. Gatenby, and B. Blumberg. Learning from and about others: Towards using imitation to bootstrap the social understanding of others by robots. *Artificial life*, 11(1-2):31–62, 2005.
[9] A. Bruderlin and T. Calvert. Interactive animation of personalized human locomotion. In *Proc. of Graphics Interface*, pp. 17–23, 1993.
[10] J. Bruneau, A. H. Olivier, and J. Pettré. Going through, going around: A study on individual avoidance of groups. *IEEE Transactions on Visualization and Computer Graphics*, 21(4):520–528, April 2015.
[11] M. S. Cary. The role of gaze in the initiation of conversation. *Social Psychology*, pp. 269–271, 1978.
[12] J. Cassell. Embodied conversational agents: representation and intelligence in user interfaces. *AI magazine*, 22(4):67, 2001.
[13] N. J. Emery. The eyes have it: the neuroethology, function and evolution of social gaze. *Neuroscience & Biobehavioral Reviews*, 24(6):581–604, 2000.
[14] A. W. Feng, Y. Xu, and A. Shapiro. An example-based motion synthesis technique for locomotion and object manipulation. In *Proceedings of the ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games*, pp. 95–102. ACM, 2012.
[15] S. M. Fiore, T. J. Wiltshire, E. J. Lobato, F. G. Jentsch, W. H. Huang, and B. Axelrod. Toward understanding social cues and signals in human–robot interaction: effects of robot gaze and proxemic behavior. *Frontiers in psychology*, 4, 2013.
[16] J. Gandrud and V. Interrante. Predicting destination using head orientation and gaze direction during locomotion in vr. In *Proceedings of the ACM Symposium on Applied Perception*, SAP '16, pp. 31–38. ACM, New York, NY, USA, 2016. doi: 10.1145/2931002.2931010
[17] E. Goffman. *Behavior in public places*. Simon and Schuster, 2008.
[18] H. Grillon and D. Thalmann. Simulating gaze attention behaviors for crowds. *Computer Animation and Virtual Worlds*, 20(2-3):111–119, 2009.

[19] H. Grillon, B. Yersin, J. Maïm, and D. Thalmann. Gaze behaviors for virtual crowd characters. In *Cross-Modal Analysis of Speech, Gestures, Gaze and Facial Expressions*, pp. 201–213. Springer, 2009.

[20] E. T. Hall. *The hidden dimension*. Doubleday & Co, 1966.

[21] D. Helbing, I. Farkas, and T. Vicsek. Simulating dynamical features of escape panic. *Nature*, 407:487–490, 2000.

[22] M. A. Hollands, A. E. Patla, and J. N. Vickers. look where youre going!: gaze behaviour associated with maintaining and changing the direction of locomotion. *Experimental brain research*, 143(2):221–230, 2002.

[23] S. Jain, Y. Ye, and C. K. Liu. Optimization-based interactive motion synthesis. *ACM Trans. Graph.*, 28(1):10:1–10:12, Feb. 2009.

[24] R. S. Johansen. *Automated semi-procedural animation for character locomotion*. PhD thesis, Aarhus Universitet, Institut for Informations-og Medievidenskab, 2009.

[25] M. Kapadia, N. Marshak, A. Shoulson, and N. I. Badler. ADAPT: The agent development and prototyping testbed. *IEEE Transactions on Visualization and Computer Graphics*, 20(7):1035–1047, 2014.

[26] I. Karamouzas, B. Skinner, and S. J. Guy. Universal power law governing pedestrian interactions. *Physical Review Letters*, 113(23):238701, 2014.

[27] M. Khamis, C. Oechsner, F. Alt, and A. Bulling. Vrpursuits: interaction in virtual reality using smooth pursuit eye movements. In *Proceedings of the 2018 International Conference on Advanced Visual Interfaces (AVI18). ACM, New York, NY, USA*, vol. 7, 2018.

[28] L. Kovar, M. Gleicher, and F. Pighin. Motion graphs. In *ACM Transactions on Graphics (TOG)*, vol. 21, pp. 473–482, 2002.

[29] T. Kwon and J. Hodgins. Control systems for human running using an inverted pendulum model and a reference motion capture sequence. In *Proceedings of the 2010 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, SCA '10, pp. 129–138. Eurographics Association, Goslar Germany, Germany, 2010.

[30] M. Kyriakou, X. Pan, and Y. Chrysanthou. Interaction with virtual crowd in immersive and semi-immersive virtual reality systems. *Computer Animation and Virtual Worlds*, 2016. CAVW-16-0002.R1.

[31] C. Liu, J. B. Hamrick, J. F. Fisac, A. D. Dragan, J. K. Hedrick, S. S. Sastry, and T. L. Griffiths. Goal inference improves objective and perceived performance in human-robot collaboration. In *Proceedings of the 2016 international conference on autonomous agents & multiagent systems*, pp. 940–948. International Foundation for Autonomous Agents and Multiagent Systems, 2016.

[32] J. Llobera, B. Spanlang, G. Ruffini, and M. Slater. Proxemics with multiple dynamic characters in an immersive virtual environment. *ACM Trans. Appl. Percept.*, 8(1):3:1–3:12, Nov. 2010.

[33] S. D. Lynch, J. Pettré, J. Bruneau, R. Kulpa, A. Crétual, and A.-H. Olivier. Effect of virtual human gaze behaviour during an orthogonal collision avoidance walking task. In *2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 136–142. IEEE, 2018.

[34] M. Moussaïd, M. Kapadia, T. Thrash, R. W. Sumner, M. Gross, D. Helbing, and C. Hölscher. Crowd behaviour during high-stress evacuations in an immersive virtual environment. *Journal of The Royal Society Interface*, 13(122):20160414, 2016.

[35] M. Moussaid, V. R. Schinazi, M. Kapadia, and T. Thrash. Virtual sensing and virtual reality: How new technologies can boost research on crowd dynamics. *Frontiers in Robotics and AI*, 5:82, 2018.

[36] S. Narang, A. Best, and D. Manocha. Simulating movement interactions between avatars & agents in virtual worlds using human motion constraints. In *2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 9–16. IEEE, 2018.

[37] S. Narang, A. Best, T. Randhavane, A. Shapiro, and D. Manocha. Pedvr: Simulating gaze-based interactions between a real user and virtual crowds. In *Proceedings of the 22nd ACM conference on virtual reality software and technology*, pp. 91–100. ACM, 2016.

[38] S. Narang, A. Best, A. Shapiro, and D. Manocha. Generating virtual avatars with personalized walking gaits using commodity hardware. In *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*, pp. 219–227. ACM, 2017.

[39] C. J. Needham, P. E. Santos, D. R. Magee, V. Devin, D. C. Hogg, and A. G. Cohn. Protocols from perceptual observations. *Artificial Intelligence*, 167(1-2):103–136, 2005.

[40] L. Nummenmaa, J. Hyönä, and J. K. Hietanen. I'll walk this way: Eyes reveal the direction of locomotion and make passersby look and go the other way. *Psychological Science*, 20(12):1454–1458, 2009.

[41] C. Park, A. Best, S. Narang, and D. Manocha. Simulating high-dof human-like agents using hierarchical feedback planner. In *Proceedings of the 21st ACM Symposium on Virtual Reality Software and Technology*, pp. 153–162. ACM, 2015.

[42] T. S. Perry. Virtual reality goes social. *IEEE Spectrum*, 53(1):56–57, 2016.

[43] D. Premack and G. Woodruff. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4):515–526, 1978.

[44] T. Randhavane, A. Bera, and D. Manocha. F2fcrowds: Planning agent movements to enable face-to-face interactions. *Presence: Teleoperators and Virtual Environments*, 26, 2017.

[45] F. A. Rojas and H. S. Yang. Immersive human-in-the-loop hmd evaluation of dynamic group behavior in a pedestrian crowd simulation that uses group agent-based steering. In *Proceedings of the 12th ACM SIGGRAPH International Conference on Virtual-Reality Continuum and Its Applications in Industry*, pp. 31–40. ACM, 2013.

[46] F. A. Sanz, A. H. Olivier, G. Bruder, J. Pettré, and A. Lécuyer. Virtual proxemics: Locomotion in the presence of obstacles in large immersive projection environments. In *2015 IEEE Virtual Reality (VR)*, pp. 75–80, March 2015.

[47] S. Satake, T. Kanda, D. F. Glas, M. Imai, H. Ishiguro, and N. Hagita. How to approach humans?: strategies for social robots to initiate interaction. In *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*, pp. 109–116. ACM, 2009.

[48] A. Schadschneider. Cellular automaton approach to pedestrian dynamics - theory. *Pedestrian and Evacuation Dynamics*, pp. 75–86, 2002.

[49] W. Shao and D. Terzopoulos. Autonomous pedestrians. In *Proceedings of the 2005 ACM SIGGRAPH/Eurographics symposium on Computer animation*, pp. 19–28. ACM, 2005.

[50] A. Shapiro. Building a character animation system. In J. Allbeck and P. Faloutsos, eds., *Motion in Games*, vol. 7060 of *Lecture Notes in Computer Science*, pp. 98–109, 2011.

[51] S. Singh, M. Kapadia, G. Reinman, and P. Faloutsos. Footstep navigation for dynamic crowds. *Computer Animation and Virtual Worlds*, 22(2-3):151–158, 2011.

[52] M. Slater, B. Lotto, M. M. Arnold, and M. V. Sanchez-Vives. How we experience immersive virtual environments: the concept of presence and its measurement. *Anuario de psicología*, 40(2), 2009.

[53] G. Snook. Simplified 3D movement and pathfinding using navigation meshes. In *Game Programming Gems*, chap. 3, pp. 288–304. Charles River, Hingham, Mass., 2000.

[54] N. Sohre, C. Mackin, V. Interrante, and S. J. Guy. Evaluating collision avoidance effects on discomfort in virtual environments. In *Virtual Humans and Crowds for Immersive Environments (VHCIE), 2017 IEEE*, pp. 1–5. IEEE, 2017.

[55] Z. Sun, W. Yu, J. Zhou, and M. Shen. Perceiving crowd attention: Gaze following in human crowds with conflicting cues. *Attention, Perception, & Psychophysics*, 79(4):1039–1049, May 2017. doi: 10.3758/s13414-017-1303-z

[56] T. D. Sweeny and D. Whitney. Perceiving crowd attention: Ensemble perception of a crowds gaze. *Psychological science*, 25(10):1903–1913, 2014.

[57] A. Treuille, S. Cooper, and Z. Popović. Continuum crowds. In *Proc. of ACM SIGGRAPH*, pp. 1160–1168, 2006.

[58] J. van den Berg, S. J. Guy, M. Lin, and D. Manocha. Reciprocal n-body collision avoidance. In *Inter. Symp. on Robotics Research*, pp. 3–19, 2011.

[59] H. Welbergen, B. Basten, A. Egges, Z. Ruttkay, and M. Overmars. Real time character animation: A trade-off between naturalness and control. *Computer Graphics Forum*, 29(8), 2010.

[60] H. Zhao, T. Thrash, S. Wehrli, C. Hölscher, M. Kapadia, J. Grübel, R. P. Weibel, and V. R. Schinazi. A networked desktop virtual reality setup for decision science and navigation experiments with multiple participants. *JoVE (Journal of Visualized Experiments)*, (138):e58155, 2018.