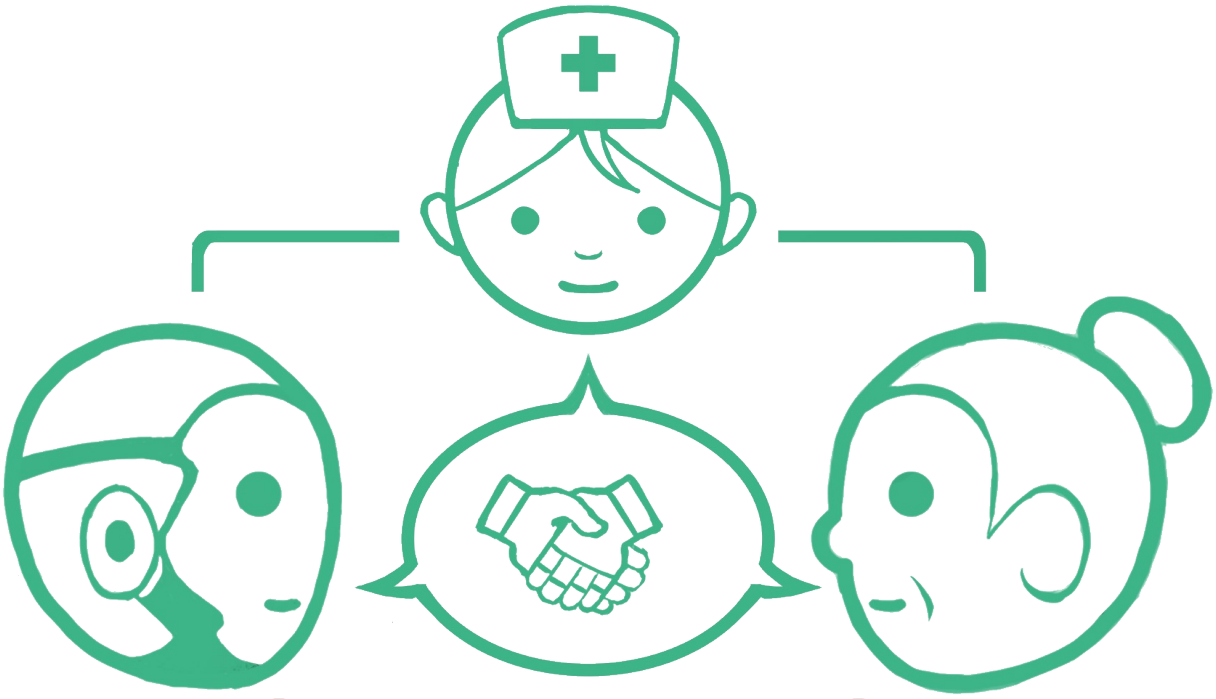

# SPRING

GRANT AGREEMENT N. 871245

## Deliverable D6.2

Specifications of the generator of situated interactions

Due Date: 31/05/2021
Main Author: Timothée Wintz (INRIA)
Contributors: Daniel Hernandez Garcia (HWU), Chris Reinke (INRIA)
Dissemination: Public Deliverable

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 871245.

SPRING

**DOCUMENT FACTSHEET**

| | |
|---|---|
| **Deliverable no.** | D6.2: Specifications of the generator of situated interactions |
| **Responsible Partner** | INRIA |
| **Work Package** | WP6: Learning Robot Behavior |
| **Task** | T6.2: Generation of Multi-party situated interactions |
| **Version & Date** | 2 (28/05/2021) |
| **Dissemination level** | PU (public) |

**CONTRIBUTORS AND HISTORY**

| Version | Editor | Date | Change Log |
|---|---|---|---|
| 1 | INRIA | 25/05/2021 | First Draft |
| 2 | INRIA | 28/05/2021 | Final Draft including all partners comments |
| | | | |

**APPROVALS**

| | |
|---|---|
| **Authors/editors** | INRIA, HWU |
| **Task Leader** | INRIA |
| **WP Leader** | Timothée Wintz (INRIA) |

# Contents

# 1 Task 6.2: Generation of Multi-party Situated Interactions

The goal of this task is to develop software tools to generate data of multi-party situated interactions. The data synthesised by this software modules should be complex enough to usefully contribute to the training of the machine learning architectures described in D6.1. These modules should use the textual and interactive data collected in task T1.2, and can be interfaced to the audio-visual simulator of task T2.2 to produce realistic rendering of scenes. The generator should also include a physical simulation of the ARI robot to produce virtual interaction between the simulated scene and the robot.

Given the constraints described above, the simulator should provide the following:

- Simulate a scene with several humans and a robot;
- Synthesise realistic behaviours of human agents;
- Include recorded trajectories and conversations.

In the following, we describe more precisely an agent-based simulator satisfying all these needs.

The software can be coarsely separated in two independent but interacting modules: the non-verbal behaviour generator, consisting in abstract interaction between individuals as well as low-level motion generation, and the conversation content generator, producing strings of text during conversation.

We begin by section 2, in which we describe what information about each agent is provided to the simulator. In section 3 we will describe the simulation of non-verbal interaction. In section 4 we will describe the conversation content generator. In both these parts, we present some review of the literature that can help us with the task. Finally, in section 5.1 we present a possible organization of software modules and APIs.

# 2 Agent-based modeling

Agent-based modeling is a common way to implement complex simulations in social scenarios [NH11]. In this framework, the simulation consists of a collection of agents with a number of characteristics and each having their own behaviour. Human agents are therefore autonomous and their behaviour is expected to simulate the behaviour of humans. The ARI Robot is one agent in this world and the simulator must be able to realistically simulate its movement given the inputs of its actuators.

## 2.1 Agent description

In the SPRING project, we consider scenarios where a humanoid robot interacts with various people in a elderly care setting. These people can have several roles: health care worker (nurse, doctor, ...), elderly patient, or accompanying person. These roles will heavily impact their behaviour and the simulator should take this into account. The simulator must take into account each person's ability and physical presence in the scene, in order to accurately produce their motion. Parameters which can influence the way a person behaves include age, sex, weight, disabilities (e.g a person in a wheelchair will not be able to move around exactly as a person walking). Therefore, we propose the following list of attributes for a human agent:

- Identifier: uuid1,uuid2,
- Role: doctor, nurse, patient
- Physical attributes: sex, weight, motion disabilities...

On top of that, the simulator can provide high level input to the agent in the form of goals. These goals can be, for example, a specific destination in the scene, starting a discussion with a given person, or even a conversational goal such as to obtain a specific piece of information from another agent.

To be able to leverage the data acquired in task T1.2, or to more easily incorporate human-produced knowledge, the agents can be locked to a pre-recorded trajectory.

### 2.2 General architecture

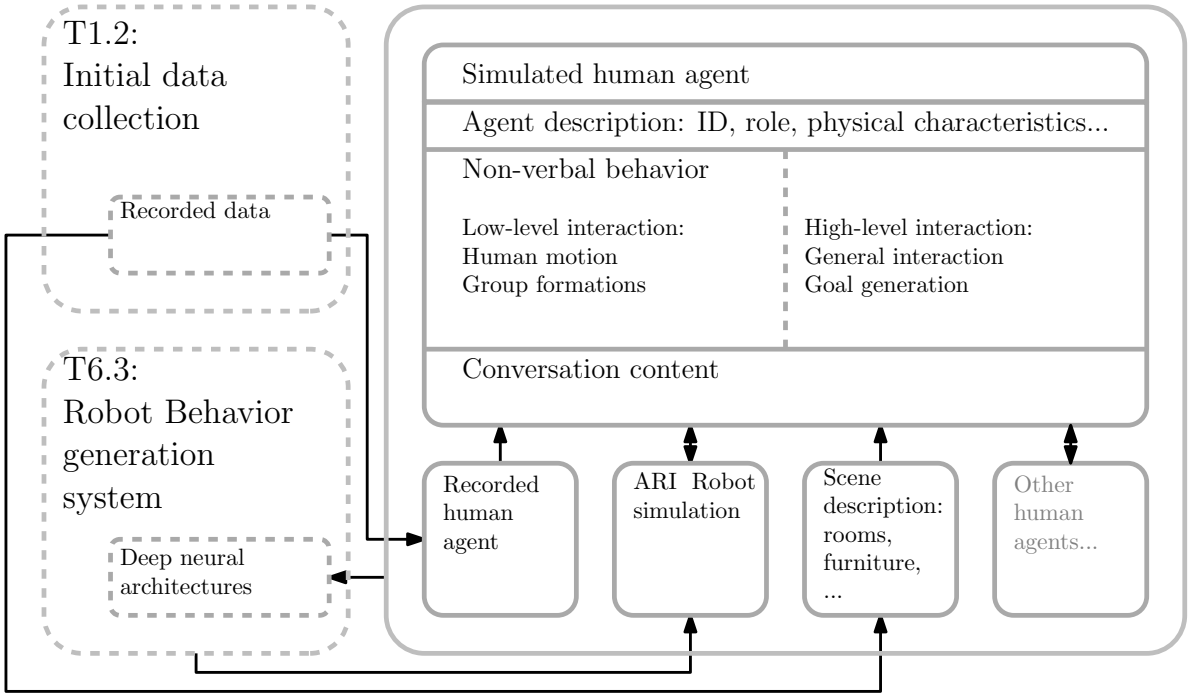


Figure 1: General architecture of the simulator. Dotted modules are other tasks in the SPRING Project that are interacting with the generator of situated interactions.

Figure 1 presents the general architecture for the generator of situated interaction. A scene consists in a general description of the environment, several human agents and the ARI robot. Each agent's behavior is generated by the modules that will be described further below: the non-verbal behavior generator and the conversation content generator. The robot behavior can be interfaced by the controller developed in Work Package 6, the robot behavior manager. The scene content can be retrieved by the deep learning models to train them on simulated data. Agents' behaviors can either be a recorded behavior from the initial data collection dataset, or be controlled by both the non-verbal behavior generator and the conversation content generator.

## 3 Non-verbal behaviour generation

In this section, we describe the non-verbal behaviour generation part of the simulator: this module should be able to simulate the behaviour of agents down to their motion, perception, and interaction with other agents. It can be separated in two sub-parts, the high-level interaction simulator, which simulates the way people interact with each other and the robot, and the motion generation part, which deals with the way agents move about in the scene and perceive their environment.

### 3.1 High-level interaction simulation

The non-verbal behaviour generator must provide high-level interaction behaviour for each of the agents.

#### 3.1.1 Requirements

The interactions in the simulator need to be grounded upon social and behavioural representations of the agents in the environment.

For a situated social simulation it is necessary to generate an agent's low-level physical motions, which produce non-verbal social cues (gaze, position, pose) with corresponding intentional mental states (e.g., beliefs, desires, emotions, actions) that are decoded from them.

The simulator will output, at every timestep, the following states about each agent:

- Intention-belief state: agent's goals and intentions.
- Affective state: agent's emotion or desire.
- Behavioural state: gaze, position, orientation, pose (low-level motion generator).

#### 3.1.2 Intention Modelling

Perception of the social world in terms of agents and their intentional relations and belief attribution is a fundamental cognitive ability that lays the foundations of our social awareness [Woo+09]. This skill enables reasoning about other agents in the environment and performing appropriate decision making. Mentalising, mentalisation, or theory of mind refers to this ability to read the mental states of other agents [FF06].

The psychology, cognitive science and artificial intelligence community has been investigating computational models for artificial intention reading for many years. Mentalising approaches have been used to perform social simulations of multiple artificial agents [PM05]. A recent common approach to model reasoning about other agents' desires and beliefs based on their actions is the Bayesian theory of mind (BToM) model [Bak+17; PK19], that considers the problem of inferring the mental states of agents through probabilistic inverse-planning. In contrast, [Rab+18] designed a neural network which uses meta-learning to build such models of the agents, able to predict the behavior of multiple agents in a false-belief situation given their past and current trajectories.

The generator could make use of some of these methods to generate believable attributions for agents' intentions in the simulated scene.

#### 3.1.3 Emotion Modelling

The ability to recognize and interpret the emotions of others is another important dimension of social interactions and communications. Enabling robots to sense and model emotions will improve their performance across a wide variety of human-robot interaction social interaction applications and can play different roles and have various purposes such as augmenting engagement and the social presence, and give the robot the illusion of life [PLR15].

Emotion recognition has been widely explored in the broader fields of human-robot interaction and affective computing, various theories to model human emotions, classification strategies, interaction modalities, and emotional models have been discussed [PLR15; McC+16; Cav+18]. In [SPR20] a list of available datasets for emotion recognition, focus on facial expressions, body poses and kinematics, voice, brain activity, and peripheral physiological responses.

### 3.2 Low-level motion generation

The non-verbal behaviour generator must provide low-level motion generation for each of the agents.

#### 3.2.1 Requirements

Social interactions in the simulator will be situated in a scene. This scene will be provided as an input to the simulator, in the form of a list of obstacles and their description (e.g. wall, furniture). The scene description will also include each agent's initial position.

The simulator will provide an accurate physical simulation of the ARI robot. PAL provides the partners with the kinematic model of the robotic base and the joints. The commands for the robotic base are given in linear and angular velocity, respecting the bounds specified by PAL. The simulator will provide feedback for collision between the robot and human agents, as well as collision between the robot and the world.

The simulator will output, at every timestep, the following information about each agent, including the robot:

- Position and orientation
- Pose: walking, standing, seating
- Walking speed
- Head orientation
- Gaze (direction of eyes)

It must also provide some estimation of the quality of audio-visual signals, since this will be used as a reward for the reinforcement learning architectures described in D6.1:

- Quality of speech signals. Either computed based on a simulation of the recorded sound of speakers and noise sources, or by simulating the resulting quality measure, e.g. the word-error-rate.
- Quality of vision signals. How well algorithms can identify posture, emotions, and other features be observed.

#### 3.2.2 Social navigation

Trajectory generation is a necessary part of generation of human behaviour in a scene. Given its goal provided by the social interaction generator, each agent must be able to navigate to its target destination. This navigation plan must take into account the other agents, and avoid unrealistic trajectories. To this end, the simulator provides a social navigation plan. Identification of human group dynamics and use of social convention is needed to provide realistic approach behavior and social avoidance of collisions [GT00].

Existing works that concentrate on the anticipation of human behaviour can be used to produce new trajectories in a simulated environment. Neural network architectures like Long Short-Term Memory (LSTM) networks [Ala+16; Zha+19] have already been demonstrated to be capable of generating credible human motions. Attention Networks have proved to be able to tackle a lot of different problems in data generation, and have also successfully been applied to human motion generation [VMO18; Fer+18b]. Generative Adversarial Networks (GAN) are another possibility for this task [Fer+18a; Gup+18; Sad+19].

The generator will make use of some of these methods to generate believable motion actions for people in the simulated scene.

#### 3.2.3 Simulation of position of people in groups

In social scene, people tend to form small groups where social interaction takes place. Proxemics is the study of human use of interpersonal space. We will leverage the existing literature on group formations to provide a realistic simulation of groups.

Proxemics have been used to create a simulation of human formation in specific scenarios, for example in a hallway after university class [PV18], or in more general settings [Caf+16; RE09].

In SPRING, the generation of group behavior will possibly more complex than these scenarios, because of the different roles of agents in the scene and how it affects group formations. This will be taken into account in the simulator.

#### 3.2.4 Gesture control

Another aspect of the robot simulation is the ability of ARI to use its motor arms to provide gestures helping with the communication with agents. The simulator should be able to generate meaningful gestures in different situations: to invite a person to follow the robot, to indicate a destination to a person or simply acknowledge the presence of a person.

These gestures should be generated in an abstracted manner: there is no need to simulate the exact position of every joint.

# 4 Conversational content generator (CCG)

The purpose of the CCG is to simulate the utterances that the different agents produce in a social context, so as to be able to test and train policies for the robot's dialogue management decisions (for example with Reinforcement Learning). For example a patient and carer may enter the waiting room together, the patient may ask the carer where they should sit, and the robot (overhearing this) could decide to offer to help find a seat. Or the patient could directly ask the robot questions such as where to find various locations or items (water, coffee etc) based on their goals. There are many possible interactions, some of which have been collected in the initial data collection of WP1.

## 4.1 Requirements

1. Simulate humans with conversational goals (e.g. check in, then find the bathroom. After waiting for 4 minutes, adopt a new goal to get some water.... )
2. Simulate utterances in response to utterances from other agents (e.g. robot or other humans) - for example "Robot: How can I help you"; Human1: "I have an appointment at 10am"
3. Simulate multiple humans, for example a patient and their carer (Patient to carer: "Where do we need to go now?", Carer to patient: "I don't know")

## 4.2 Dialogue Simulation

There has been a variety of prior work on user simulation for dialogue systems, most of which has considered only a single user within task-based settings, for example a user who wants to fly from London to New York on a particular date. These simulations have been used to train dialogue managers using Reinforcement Learning. One useful such approach is the "Agenda-based" user simulation, where a user is modelled as having a list of constraints or goals (the Agenda) which they wish to convey in a conversation (e.g. destination = New York; depart-date= 12/4/22). User utterances can then be simulated via probability distributions over user dialogue moves, given a prior system/robot dialogue move. For example, when asked "How may I help you" a user might give only one item from their agenda (provide-info(destination=New York)) with probability 0.6 or they might provide 2 items (provide-info(destination=New York, depart-date= 12/4/22)) with probability 0.2, or perhaps no answer with probability 0.05. These probabilities are estimated from collected data.

The challenges for the SPRING project are to extend such approaches to multi-agent settings with different roles (patient, carer, doctor...) and more complex agendas consisting of various different goals.

The CCG for the SPRING project will develop approaches such as the Multi-user Simulation Environment (MUSE) model of [Kei+13] using up-to-date tools such as [Shi+19][1].

# 5 Implementation

## 5.1 Software tools and code availability

The software will be composed of two main modules, as described above: one for the non-verbal behaviour generation, and the other for conversation content generation. ROS [Sta18] will be used for communication between the modules and for interaction between the learning architectures and the simulator. The ROS interface will possibly be implemented as a lightweight overlay on the rest of the software, so that it can be used independently of the ROS platform more easily, for example when computations on cluster nodes are necessary.

The low-level physical simulation will be implemented in python in a two-dimensional environment. A possible framework is to use the Box2D physics engine with a 2D rendering engine like PyGame. This allows to produce a fast and simple simulation environment, and should be realistic enough for the simple physics of the ARI robot.

The software will be made available on the SPRING project Gitlab repositories: the Generator of Multi-party Situated Interactions modules will be located on the Work Package 6 [SPR] repository. These will be available to the public for the duration specified in the SPRING project proposal.

[1] https://github.com/wyshi/user-simulator

# References


[Ala+16] Alexandre Alahi et al. "Social lstm: Human trajectory prediction in crowded spaces". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 961–971.

[Bak+17] Chris L. Baker et al. "Rational quantitative attribution of beliefs, desires and percepts in human mentalizing". In: *Nature Human Behaviour* (Mar. 2017). DOI: 10.1038/s41562-017-0064. URL: http://www.nature.com/articles/s41562-017-0064.

[Caf+16] Angelo Cafaro et al. "The effects of interpersonal attitude of a group of agents on user's presence and proxemics behavior". In: *ACM Transactions on Interactive Intelligent Systems (TiiS)* 6.2 (2016), pp. 1–33.

[Cav+18] F. Cavallo et al. "Emotion Modelling for Social Robotics Applications: A Review". In: *Journal of Bionic Engineering* 15 (2018), pp. 185–203.

[Fer+18a] Tharindu Fernando et al. "Gd-gan: Generative adversarial networks for trajectory prediction and group detection in crowds". In: *Asian Conference on Computer Vision*. Springer. 2018, pp. 314–330.

[Fer+18b] Tharindu Fernando et al. "Soft+ hardwired attention: An lstm framework for human trajectory prediction and abnormal event detection". In: *Neural networks* 108 (2018), pp. 466–478.

[FF06] Chris D. Frith and Uta Frith. "The Neural Basis of Mentalizing". In: *Neuron* 50.4 (2006), pp. 531–534. ISSN: 0896-6273. DOI: https://doi.org/10.1016/j.neuron.2006.05.001. URL: http://www.sciencedirect.com/science/article/pii/S0896627306003448.

[GT00] Anthony Guye-Vuilleme and Daniel Thalmann. "A high-level architecture for believable social agents". In: *Virtual Reality* 5.2 (2000), pp. 95–106.

[Gup+18] Agrim Gupta et al. "Social gan: Socially acceptable trajectories with generative adversarial networks". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 2255–2264.

[Kei+13] Simon Keizer et al. "Training and evaluation of an MDP model for social multi-user human-robot interaction". In: *Proceesings of SIGDIAL*. 2013.

[McC+16] Derek McColl et al. "A Survey of Autonomous Human Affect Detection Methods for Social Robots Engaged in Natural HRI". en. In: *Journal of Intelligent & Robotic Systems* 82.1 (Apr. 2016), pp. 101–133. ISSN: 0921-0296, 1573-0409. DOI: 10.1007/s10846-015-0259-2. URL: http://link.springer.com/10.1007/s10846-015-0259-2 (visited on 05/17/2021).

[NH11] Muaz Niazi and Amir Hussain. "Agent-based computing from multi-agent systems to agent-based models: a visual survey". In: *Scientometrics* 89.2 (2011), pp. 479–499.

[PK19] Jan Pöppel and Stefan Kopp. "Satisficing Mentalizing: Bayesian Models of Theory of Mind Reasoning in Scenarios with Different Uncertainties". In: *CoRR* abs/1909.10419 (2019). arXiv: 1909.10419. URL: http://arxiv.org/abs/1909.10419.

[PLR15] Ana Paiva, Iolanda Leite, and Tiago Ribeiro. "Emotion Modeling for Social Robots". In: 2015.

[PM05] David V. Pynadath and Stacy C. Marsella. "PsychSim: Modeling Theory of Mind with Decision-Theoretic Agents". In: *Proceedings of the 19th International Joint Conference on Artificial Intelligence*. IJCAI'05. Edinburgh, Scotland: Morgan Kaufmann Publishers Inc., 2005, pp. 1181–1186.

[PV18] Claudio Pedica and Hannes Högni Vilhjálmsson. "Study of nine people in a hallway: Some simulation challenges". In: *Proceedings of the 18th International Conference on Intelligent Virtual Agents*. 2018, pp. 185–190.

[Rab+18] Neil Rabinowitz et al. "Machine Theory of Mind". In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by Jennifer Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. Stockholmsmässan, Stockholm Sweden: PMLR, Oct. 2018, pp. 4218–4227. URL: http://proceedings.mlr.press/v80/rabinowitz18a.html.

[RE09] Matthias Rehm and Birgit Endrass. "Rapid prototyping of social group dynamics in multiagent systems". In: *AI & society* 24.1 (2009), pp. 13–23.

[Sad+19] Amir Sadeghian et al. "Sophie: An attentive gan for predicting paths compliant to social and physical constraints". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019, pp. 1349–1358.

[Shi+19] Weiyan Shi et al. "How to Build User Simulators to Train RL-based Dialog Systems". In: *Proceedngs of EMNLP*. 2019. arXiv: 1909.01388 [cs.CL].

[SPR] SPRING Project. *WP6: Robot Behavior*. URL: https://gitlab.inria.fr/spring/wp6_robot_behavior.

[SPR20] Matteo Spezialetti, Giuseppe Placidi, and Silvia Rossi. "Emotion Recognition for Human-Robot Interaction: Recent Advances and Future Perspectives". In: *Frontiers in Robotics and AI* 7 (2020), p. 145. ISSN: 2296-9144. DOI: 10.3389/frobt.2020.532279. URL: https://www.frontiersin.org/article/10.3389/frobt.2020.532279.

[Sta18] Stanford Artificial Intelligence Laboratory et al. *Robotic Operating System*. Version ROS Melodic Morenia. May 23, 2018. URL: https://www.ros.org.

[VMO18] Anirudh Vemula, Katharina Muelling, and Jean Oh. "Social attention: Modeling attention in human crowds". In: *2018 IEEE international Conference on Robotics and Automation (ICRA)*. IEEE. 2018, pp. 1–7.

[Woo+09] Amanda L. Woodward et al. "Chapter 6 The Emergence of Intention Attribution in Infancy". In: *The Psychology of Learning and Motivation*. Vol. 51. Psychology of Learning and Motivation. Academic Press, 2009, pp. 187–222. DOI: https://doi.org/10.1016/S0079-7421(09)51006-7. URL: https://www.sciencedirect.com/science/article/pii/S0079742109510067.

[Zha+19] Pu Zhang et al. "Sr-lstm: State refinement for lstm towards pedestrian trajectory prediction". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019, pp. 12085–12094.