DARKO

Dynamic Agile Production Robots That Learn and Optimise Knowledge and Operations

H2020-ICT-2020-2 Grant agreement no: 101017274

DELIVERABLE 5.1

Report on prediction of human motion and intents

Dissemination Level: PUBLIC

Due date: month 43 (July 2024)

Deliverable type: Report

Lead beneficiary: Robert Bosch GmbH (Bosch)

Contents

List of Abbreviations

Abbreviation	Meaning
ADE	Average Displacement Error
API	Application Programming Interface, the public interface provided by a library for use by software developers
ARENA 2036	A large research campus in form of a modern factory hall in Stuttgart-Vaihingen, Germany. Provides an innovation platform for mobility & production of the future and hosts DARKO project demonstrations.
ATC	Large dataset of human motion trajectories, recorded by Brscic et al. in 2013
CLiFF-Map	Circular Linear Flow Field map, a specific type of Maps of Dynamics
CNN	Convolutional Neural Network
CVM	Constant Velocity Model, a popular short-term prediction baseline
FDE	Final Displacement Error
GPU	Graphics Processing Unit, used as a deep learning accelerator to train and run inference on neural networks
GRU	Gated Recurrent Unit, a gating mechanism in RNNs
HRI	Human-Robot Interaction
ILIAD	EU Horizon 2020 project (2016–2020) which deployed a heterogeneous fleet of mobile service robots in intralogistics environments.
LHMP	Long-term Human Motion Prediction
LiDAR	Light Detection And Ranging, a time-of-flight-based sensor that produces point clouds. Also spelled “lidar”.
MoD	Map of Dynamics
MPC	Model Predictive Control, a popular collision avoidance technique where predictions of human motion and robot controls can be integrated as costs over a short outlook horizon
NLL	Negative Log-Likelihood
NN	Neural Network
ORU	Örebro University, member of the DARKO consortium
RGB(-D)	Red, Green, Blue(, Depth)
RNN	Recurrent Neural network
ROS	Robot Operating System, see www.ros.org
SDK	Software Development Kit
SLAM	Simultaneous Localization and Mapping
SPENCER	EU FP7 project (2013–2016) which deployed a mildly humanized service robot in a busy airport terminal at Amsterdam Schiphol Airport.
SVM	Support Vector Machine, a machine learning classifier
THÖR	A context-rich dataset of human and robot motion trajectories, recorded by some of the DARKO consortium members in 2019 as part of the ILIAD project
THÖR-MAGNI	A large-scale extension to the THÖR dataset, recorded as part of the DARKO project
TUM	Technical University of Munich, member of the DARKO consortium
WP	Work package in DARKO
YOLO	A series of 2D object detectors developed by J. Redmon

1 Introduction

The deliverable reports on the system for prediction of human motion and intents developed in the EU H2020 task T5.1, including its scientific results and an initial software prototype that will be providing input to other components in DARKO (e.g. motion planning and control in WP6).

Project partners, contributing to this deliverable, are: Robert Bosch GmbH (BOSCH, lead responsible), Örebro University (ORU), and Technical University of Munich (TUM).

1.1 Motivation

We advocate for the notion that the success of intralogistics robots is deeply intertwined with their collaboration with human labor. The objective should not be to erect entirely new, fully automated warehouses; rather, businesses of varying scales should aim for intelligent automation systems. These systems should not only integrate smoothly into existing warehouse operations but also facilitate effective and safe cooperation between robots and human workers. Therefore, the DARKO project sets an important objective of "Predictive Safety and Efficiency in Human-Robot Coordination" (O2), which builds on the methods developed in WP5.

1.2 Key achievements

During the work on Task 5.1 of WP5, the consortium has reached several milestones that significantly advance the field of human-robot interaction and autonomous navigation.

We begin with several simple and robust baseline methods for trajectory prediction, introduced in our novel Atlas benchmark for human motion prediction methods. This component establishes the foundation for evaluation and benchmarking, adopted in the remainder of the project. Upon these baselines, we proceed to develop prediction methods capable of increasingly more involved understanding of human motion and its context. This work was presented at the 2022 IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN) [1].

Considering the spatious and topologically complex intralogistic and manufacturing environments where the DARKO platform is expected to operate, we addressed the Long-term Human Motion Prediction using Maps of Dynamics (MoD-LHMP). Our method, which uses the specific CLiFF-Map of human motion fynamics as input from WP3, is capable of accurate, multimodal environment-aware forecasts in a very long-term perspective (CLiFF-LHMP). This work was presented at the 5th Workshop on Long-term Human Motion Prediction[1] as part of the 2023 International Conference on Robotics and Automation (ICRA) [2] and disseminated at the 2023 International Conference on Intelligent Robots and Systems (IROS) [3].

We further improved MoD-LHMP with better decomposition of uncertainty in the training data into laminar (regular) flows and turbulent (outlier) flows. To allow more accurate prediction in realistic manufacturing environments, shared by diverse agents (workers, forklifts, other robots), we extend the MoD-LHMP method with class or activity token of the moving agent. We also propose to use time-conditioned Maps of Dynamics to achieve more accurate predictions in presence of timed everts during the day, such as the morning and evening rish hours, delivery and lunch schedules, etc. This line of research was disseminated at the 2024 International Conference on Robotics and Automation (ICRA) [4], 4th Workshop on Visual Perception for Navigation in Human Environments[2]

[1] https://motionpredictionicra2023.github.io/

[2] https://jrdb.erc.monash.edu/workshops/iccv2023

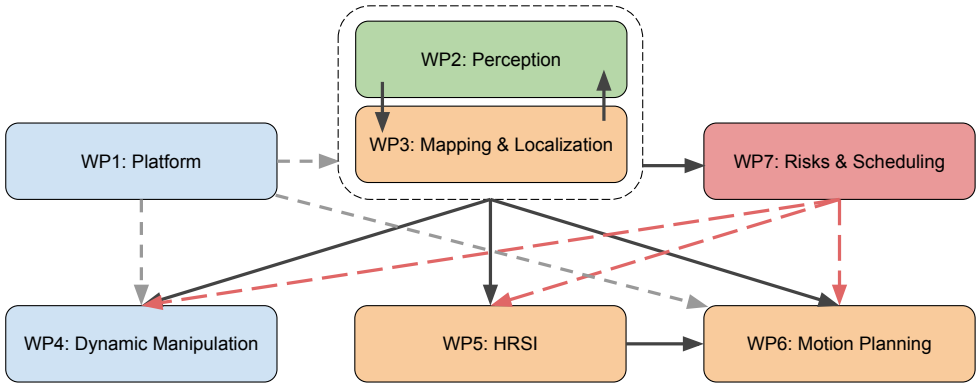


Figure 1: Relation of WP5, which this deliverable reports on, to other work packages in DARKO. Black arrows denote data flow during operation, dashed red arrows indicate constraints and orchestration, and dashed grey arrows indicate hardware dependencies.

as part of the 2023 International Conference on Computer Vision (ICCV) [5], and in the Robotics and Automation Letters (RA-L) journal [6].

We also propose to use full-body poses for more accurate trajectory prediction. Jointly with WP2 T2.5, we propose a joint non-autoregressive transformer method with novel motion transformation technique that allows training directly in 3D coordinates. In addition to the full-body as an important descriptor of human motion, we also investigate significance that gaze plays during navigation. We discover the relation between the head orientation and gaze direction to allow more informed gaze estimation from the robot's on-board sensors. This work was presented at the 2024 IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN) [7].

Development of these advanced trajectory prediction methods would not have been possible relying solely on the prior art datasets, often limited in the context of motion and interaction. To this end, a major milestone of our work in WP5 was recording the THÖR-MAGNI dataset with diverse contextual cues, numerous participants and several elaborate scenarios. This work was presented at the Workshop Towards Socially Intelligent Robots In Real World Applications[3] (SIRRW 2022) as part of the 2022 IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN) [8], and in the International Journal of Robotics Research (IJRR) [9].

1.3 Relation to other work packages

Figure 1 illustrates the relation of work package WP5, which this deliverable reports on, to the other technical work packages in DARKO. WP5 receives input from WP2 (Perception) in the form of human positions, velocities and full-body poses, and from WP3 (Mapping and localization) the map of human dynamics. WP5 provides its output to WP6 (Motion planning).

At the technical level, all inter-component communication in DARKO happens via the Robot Operating System (ROS). More details on the communication between different DARKO components at a task level, including the used ROS message types, can be found in the system architecture deliverable D8.2.

[3] https://sirrw-2022.github.io/proceedings/

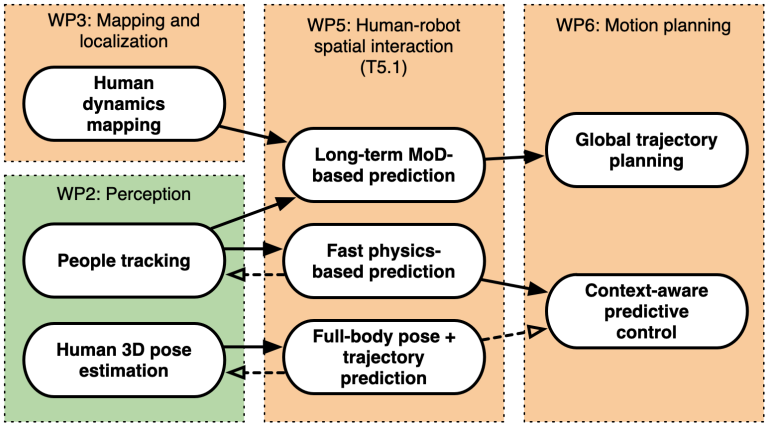


Figure 2: DARKO prediction system architecture, developed in T5.1. Black arrows show the data flow through the prediction modules. Dashed arrows indicate possible or planned output of predictions to the WP2 and WP6 components.

1.4 Repositories

The software developed is hosted in a Gitlab version control system managed by the project coordinator.

Main repositories currently available are:

- The Atlas Benchmark in https://github.com/boschresearch/the-atlas-benchmark. This software is developed by BOSCH members.
- CLiFF-LHMP in https://github.com/test-bai-cpu/CLiFF-LHMP and LaCE-LHMP in https://github.com/test-bai-cpu/LaCE-LHMP. This software is developed by ORU members.
- THÖR-MAGNI Dataset in https://zenodo.org/records/10407223. This software is mainly developed by ORU and TUM members.
- THÖR-MAGNI Human Motion Data Processing and Visualization Tools in https://github.com/tmralmeida/thor-magni-tools. This software is mainly developed by ORU and TUM members.

Other software packages developed by BOSCH members are not yet disclosed to the public.

2 DARKO prediction system architecture

The human motion prediction system on the DARKO robot, developed in WP5 T5.1, hosts an array of methods with increasing levels of context-awareness and high-level cues of human behavior. This system is designed to provide the required levels of prediction accuracy and operation speed, requested by the many downstream components. A more detailed description of the software architecture is provided in the system architecture deliverable D8.2, and its summary is given in Fig. 2.

Below are listed the main components, described in details in the following sections:

- On the most basic level, we deploy fast physics-based methods [11] that can be used by the robot controller in WP6 for fast reaction and collision avoidance. These

methods can also be used to improve the temporal association for people tracking in WP2. These methods are evaluated in Sec. 3.3.

- For the more far-reaching outlook, we deploy long-term trajectory prediction methods which are multi-modal and environment-aware [2, 3, 4]. Our solution is based on the patterns of human dynamics from WP3 (Maps of Dynamics). In topologically complex large indoor spaces with multiple robots or stationary sensors, these predictions can be used by the global path planner in WP6 to re-route the DARKO robot on the path of least disturbance. These methods are described in Sec. 5.

- To move beyond the geometric representation of a moving person as a point on the top-down view plane, as commonly used by the prior art collision avoidance methods, we use the 3D full-body poses from WP2 as input to trajectory prediction to significantly improve trajectory prediction and enable pose prediction in global coordinates [10]. This can in turn improve full-body pose tracking accuracy in WP2 in the events of sensor noise, partial or full occlusions. Furthermore, in WP6 we developed a context-aware MPC-based collision avoidance method which includes static poses and predictions. In the next iterations, this method will be extended with full-body pose predictions. More details are presented in Sec. 6.

- We also use head orientation and gaze direction of a walking person to describe the movement intention [7]. Head orientation is coming from the WP2 perception system, whereas gaze direction prototype is achieved in WP5 using eye-tracking glasses. In a preliminary study, we investigate if the on-board estimation of head orientation can serve as a reasonable proxy for gaze tracking using a dedicated device, and quantify the expected accuracy of this assumption. This study is presented in Sec. 7.

3 The Atlas Benchmark: an Automated Evaluation Framework for Human Motion Prediction

Summary: Human motion trajectory prediction, an essential task for autonomous systems in many domains, has been on the rise in recent years. With a multitude of new methods proposed by different communities, the lack of standardized benchmarks and objective comparisons is increasingly becoming a major limitation to assess progress and guide further research. Existing benchmarks are limited in their scope and flexibility to conduct relevant experiments and to account for contextual cues of agents and environments. To overcome this limitation and set up appropriate benchmarking pipeline, early in the scope of the DARKO project we developed Atlas, a new benchmark to systematically evaluate human motion trajectory prediction algorithms in a unified framework. Atlas offers data preprocessing functions, hyperparameter optimization, comes with popular datasets and has the flexibility to setup and conduct underexplored yet relevant experiments to analyze a method's accuracy and robustness. In order to set up a fast and basic short-term trajectory prediction baseline for the DARKO robot, we compare five popular model- and learning-based predictors. Our findings indicate that, when properly applied, early physics-based approaches are still remarkably competitive. Such results confirm the necessity of benchmarks like Atlas.

3.1 Introduction

Benchmarking motion prediction algorithms is a challenging task. The evaluation outcome can be affected by various factors such as data, parameters, hyperparameters and experiment design. Elaborate and carefully designed experiments are necessary to expose specific abilities or limitations of a method, in particular for complex learning approaches. Influencing factors are, for example, the observation period, i.e. the duration that agents need to be seen to allow for accurate prediction of their motion, or the exact procedure how to set up a testing scenario from sequences of raw person detections. Even when evaluating a simple constant velocity motion model with the same dataset, metrics and prediction horizon, the evaluation results may still vary as reported in [12] and [13] due to differences in testing scenario generation and data pre-processing. The limitations of the protocols commonly used to evaluate new prediction methods have been pointed out by several authors [14, 13, 11].

In this section we describe the Atlas benchmark as a first step towards automated benchmarking of motion prediction methods in a unified framework with systematic variation of key prediction parameters. Atlas includes heterogeneous datasets of human motion trajectories, is capable of automatically extracting testing scenarios, and can deal with varying, missing and noisy agent detections using data interpolation, downsampling and smoothing. Compared to prior art such as TrajNet++ [15], it offers several tunable parameters like the observation period and prediction horizon, is able to import semantic maps and other relevant information such as goal positions in the map, allows to evaluate probabilistic prediction results and to conduct robustness experiments with simulated perception noise. Due to those features, our benchmark works with both short- and long-term predictors. Unlike TrajNet++, it is especially suited for studying how prediction parameters influence the results, in contrast to fixing the main parameters to produce the ranking scores in a specific *challenge*. Furthermore, our benchmark has a direct interface to the hyperparameter estimation framework SMAC3 [16] to calibrate a predictor on a specific dataset. This feature is particularly useful for model-based predictors, which, as we will show in the experiments, can perform still very well compared to recent learning-based ones.

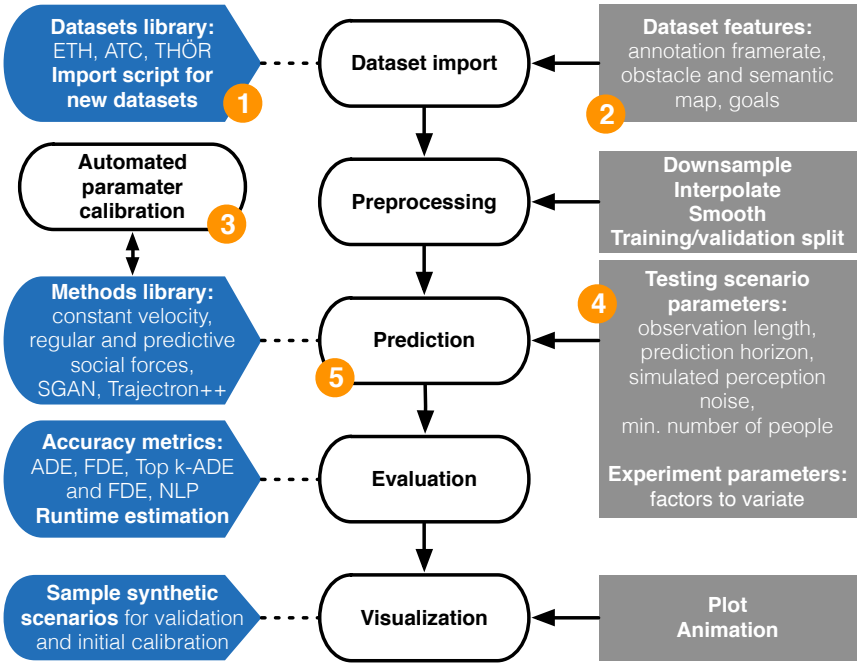


Figure 3: Atlas benchmark overview: (1) Supported import of new datasets (labeled detection streams), (2) Support for contextual cues in the environment, (3) Automated calibration of prediction hyperparameters, (4) Automated parametrized scenario extraction, (5) Direct interface to the prediction methods.

We showcase Atlas by evaluating several popular model- and learning-based methods [17, 18, 19, 20] in terms of their prediction accuracy, ability to predict in new environments, and their robustness to perception noise and limited observations.

3.2 Background

A trajectory prediction method aims to estimate a probability distribution over future positions of a moving agent within a certain time horizon. Typically, a motion predictor uses as input the agent's current or past motion states, possibly augmented by the current or past states of the environment. The environment is represented by the states of other moving agents, a topometric map of static obstacles, and possibly semantic information associated to parts, locations or objects of the map.

For the evaluation of a motion predictor we consider the following elements: datasets (popular examples include [21, 22, 23, 24, 25, 26]), the testing scenario extraction strategy and the evaluation metrics. As testing scenario extraction we denote the conversion of the continuous flow of (agent) detections, where past detections between consecutive frames form the observation history of length $O_s \in \mathbb{R}^+$ seconds (or $O_p \in \mathbb{Z}^+$ positions), and future agent states within horizon $T_s \in \mathbb{R}^+$ seconds (equivalent to $T_p \in \mathbb{Z}^+$ positions) form the predictions to be compared to the ground truth (GT). The metrics used to this end include geometric and probabilistic distance estimates between predicted and GT positions [11].

Based on these insights, the *Atlas benchmark* comes with an automated procedure to extract testing scenarios from datasets with flexible O_p and T_p parameters. Atlas accepts occupancy and semantic maps as input, supports various forms of parametric and non-parametric uncertainty representation, and includes robustness experiments with added

noise to the observed trajectories.

3.3 Atlas design and experiments

Atlas includes five main elements: data import, preprocessing, the actual prediction phase, evaluation and visualization, see Fig. 3. This design allows to interface and parametrize different prediction algorithms for a flexible and highly automated evaluation and analysis.

As first step, the datasets and, if available, information on the environment such as goals, obstacles, and semantics are imported into the benchmark. Next, the raw data is preprocessed with downsampling to a user-defined frequency, misdetection interpolation and trajectory smoothing. Once the dataset is ready, we extract the testing scenarios with the user-specified observation and prediction lengths, and the minimum number of observed people. The observed past trajectories of all people in the testing scenario, along with environment data, are explicitly interfaced as input to the prediction algorithm. The returned predictions are evaluated against the ground truth using several metrics. Finally, the prediction results can be visualized with plots or animations. Meta-parameters to control the data processing and benchmark setup are stored in separate yaml files, and the benchmark is accessed via Jupyter notebooks.

Building on the datasets, pre-processing steps and metrics described above, the benchmark enables researchers to set up and conduct several experiments to study prediction performance under varying conditions. Such experiments are not only key for researchers to better understand the algorithm or model at hand, e.g. during ablation studies, but also for practitioners to evaluate a predictor within a system with adjacent up- and downstream tasks and real-world deployments.

3.3.1 Prediction accuracy conditioned on parameters

O_s and T_s are among the main factors, associated with predicting motion. The accuracy naturally degrades for further time instances, while longer observations may improve it overall. In Atlas it is possible to measure the accuracy of prediction conditioned on these two main parameters. Further accuracy breakdown is possible by conditioning the measured values on the number of people in the scenario.

3.3.2 Transfer experiments

A crucial part of evaluating a prediction method is to analyze its generalization ability to new environments not included in the training data. Such experiments are most often overlooked in related work. In Atlas it is possible to script hyperparameter optimization in one dataset, and evaluate the method in another. In the future we plan to extend this functionality for training models.

3.3.3 Robustness experiments

For a system to work in the real world, a predictor must be robust against imperfection in perception such as noisy agent position observations. One possible way to quantify robustness, implemented in Atlas, is by measuring accuracy on the testing scenarios, after artificially adding increasing amounts of white Gaussian noise.

3.4 Evaluation of fast short-term prediction methods

With the Atlas benchmark described above, we now demonstrate its usage in an example evaluation. To this end, we conducted experiments to study and compare the performance

	Prediction horizon				
	Method	1.6 s	3.2 s	4.8 s	8 s
ADE	CVM	0.155 ± .04	0.319 ± .09	0.499 ± .15	0.870 ± .30
	Sof	0.156 ± .05	0.318 ± .09	0.494 ± .15	0.870 ± .30
	Kara	0.164 ± .05	0.324 ± .09	0.508 ± .15	0.872 ± .31
	SGAN	0.240 ± .08	0.500 ± .15	0.785 ± .24	–
	T++	0.152 ± .04	0.340 ± .09	0.549 ± .16	1.006 ± .29
FDE	CVM	0.272 ± .08	0.621 ± .19	1.000 ± .33	1.845 ± .71
	Sof	0.272 ± .08	0.620 ± .19	1.008 ± .33	1.846 ± .71
	Kara	0.279 ± .08	0.623 ± .19	1.000 ± .33	1.845 ± .71
	SGAN	0.418 ± .13	0.977 ± .31	1.592 ± .51	–
	T++	0.273 ± .08	0.689 ± .20	1.149 ± .35	2.187 ± .70

Table 1: ADE in the ATC dataset varying the prediction horizons

	Prediction horizon				
	Method	1.6 s	3.2 s	4.8 s	8 s
ADE	CVM	0.20 ± 0.09	0.50 ± 0.22	0.87 ± 0.39	1.80 ± 0.73
	Sof	0.29 ± 0.13	0.54 ± 0.22	0.82 ± 0.34	1.42 ± 0.52
	Kara	0.32 ± 0.14	0.57 ± 0.23	0.85 ± 0.35	1.44 ± 0.54
	SGAN	0.29 ± 0.11	0.65 ± 0.22	1.08 ± 0.36	–
	T++	0.18 ± 0.07	0.47 ± 0.18	0.84 ± 0.33	1.68 ± 0.62
FDE	CVM	0.38 ± 0.17	1.07 ± 0.48	1.95 ± 0.86	4.10 ± 1.51
	Sof	0.46 ± 0.20	1.02 ± 0.44	1.66 ± 0.74	2.95 ± 1.08
	Kara	0.49 ± 0.21	1.05 ± 0.45	1.69 ± 0.75	2.96 ± 1.08
	SGAN	0.52 ± 0.19	1.36 ± 0.47	2.33 ± 0.27	–
	T++	0.34 ± 0.14	1.04 ± 0.42	1.91 ± 0.75	3.79 ± 1.28

Table 2: ADE in the THÖR3 dataset varying the prediction horizon

of a small range of popular methods for human motion prediction, from simple physics-based baselines [17, 18, 27] to state-of-the-art deep learning methods [19, 20]. More details on the methods are available in [28].

These methods are evaluated on the ETH [29], ATC [24] and THÖR [26] datasets, the latter recorded by the DARKO consortium members in the pevious EU H2020 project ILIAD (2017-2021), and extended as THÖR-MAGNI in the course of DARKO (see Sec. 4).

We present the results in Tables 1–3, Fig. 4–8 and show example predictions in Fig. 9–12. Tables 1 and 2 show the results of evaluating the ADE and FDE on different prediction horizons with the fixed observation period $O_s = 3.2$ s. In the ATC dataset, which contains mostly straight linear motion, even in crowded scenes, the force-based approaches perform on the level of constant velocity. Trajectron++ and SGAN, on the other hand, attempt to predict more variety in motion than what exists in real life, leading to higher displacement errors (see an example scenario in Fig. 10).

In the THÖR3 dataset (Table 2 and Fig. 12), on the contrary, people navigate in a tighter environment across multiple directions, increasing the importance of good interaction modeling. Here Trajectron++ outperforms the CVM, however the best and most stable results are reached by the force-based methods.

In the experiments with different observation horizons we found all methods to perform very robustly even with observation lengths as short as 1.2 s, see Fig. 4 and 5. The ADE/FDE results for increasing amounts of noise in the agent positions added to the ETH and THÖR3 data are shown in Fig. 6 and 7. The performance of all methods, including Trajectron++, degrades considerably when the observations become more unreliable (with $\sigma \geq 0.2$) where SGAN shows an almost linear degradation to the amount of noise, as compared to exponential decrease of other methods. Again, the model-based methods outperform the learning-based ones.

Table 3 summarizes the transfer experiment, where the methods are calibrated on one dataset and tested on another. We observe that the predictive social force approach (Kara)

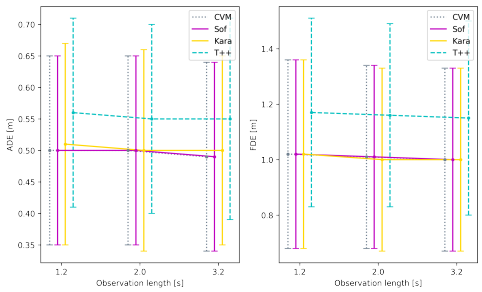


Figure 4: ADE/FDE in the ATC dataset with different observation lengths

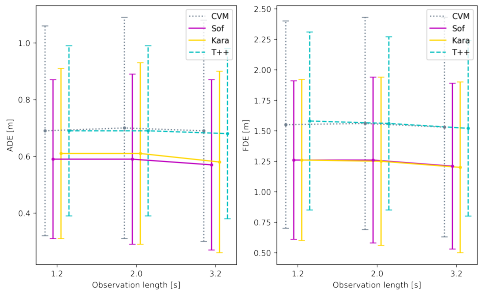


Figure 5: ADE/FDE in the THÖR1 dataset with different observation lengths

delivers more stable transfer performance in all cases as compared to the Sof method.

An overall conclusion from the experiments, supported by the qualitative analysis in Fig. 9–12, is that the model-based prediction methods, properly calibrated, with velocity filtering and goal projection, offer a surprisingly competitive alternative to the complex state-of-the-art deep learning approaches. This result seems to confirm the recent findings by Schöller et al. [13], once again indicating that learning interactions is an extremely challenging task prone to evaluation pitfalls. That, and the considerable runtime differences in favor of the model-based approaches in Fig. 8, justifies the need for further research into interaction models, both engineered and learned ones. Another conclusion is that, in our experiments, the predictive social force model does not reliably outperform the original method. Finally, the results of the force-based methods calibrated on simpler datasets with a lot of linear motion (such as the ETH and ATC) converge to the CVM model up to the 3rd decimal digit (i.e. less than 1 cm difference).

		Test			
Calibrate	Dataset	ETH	ATC	THÖR1	THÖR3
	ETH	CVM: 0.283 ± .12	0.499 ± .15	0.69 ± .39	0.87 ± .39
		Sof: 0.277 ± .11	0.494 ± .15	0.58 ± .32	0.80 ± .34
		Kara: 0.278 ± .12	0.498 ± .15	0.62 ± .33	0.81 ± .38
		SGAN: 0.787 ± .42	0.785 ± .24	0.94 ± .39	1.08 ± .36
		T++: 0.399 ± .36	0.549 ± .16	0.66 ± .30	0.84 ± .33
	ATC	Sof: 0.29 ± .13	0.497 ± .15	0.67 ± .37	0.85 ± .38
		Kara: 0.34 ± .15	0.501 ± .16	0.58 ± .31	0.81 ± .36
	THÖR1	Sof: 0.31 ± .14	0.491 ± .15	0.57 ± .30	0.76 ± .34
		Kara: 0.28 ± .11	0.493 ± .15	0.58 ± .32	0.81 ± .34
	THÖR3	Sof: 0.29 ± .12	0.50 ± .15	0.60 ± .32	0.82 ± .34
		Kara: .28 ± .11	0.49 ± .15	0.61 ± .33	0.85 ± .35

Table 3: ADE in the transfer experiments on different datasets

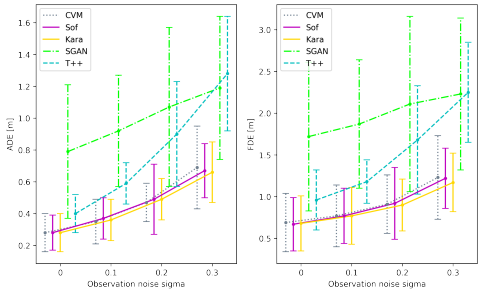


Figure 6: ADE/FDE in the ETH dataset with added noise

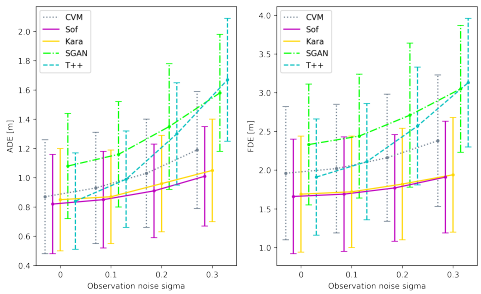


Figure 7: ADE/FDE in the THÖR3 dataset with added noise

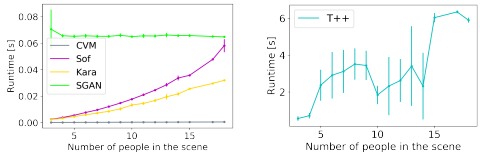


Figure 8: Average runtimes to compute predictions for $O_s = 3.2$ s and $T_s = 4.8$ s in the scenes from the ATC dataset, sorted by the number of people. **Left**: model-based methods and SGAN, **right**: Trajectron++. Despite achieving roughly comparable performance, the model-based methods are two orders of magnitude faster than the Trajectron++. SGAN has constant runtime performance. The irregular shape of the Trajectron++ performance curve is explained by the number of and the irregularities in the scenarios: due to the pooling and pruning when computing interactions, a scene with 10 people far away from each other might be easier to solve than with 6 people closely interacting.

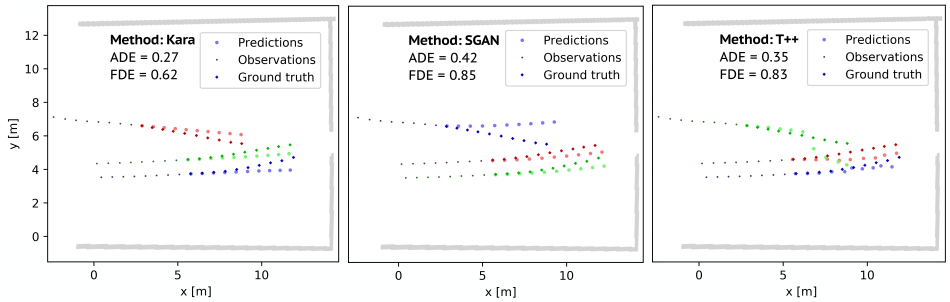


Figure 9: Predictions in the ETH scenario

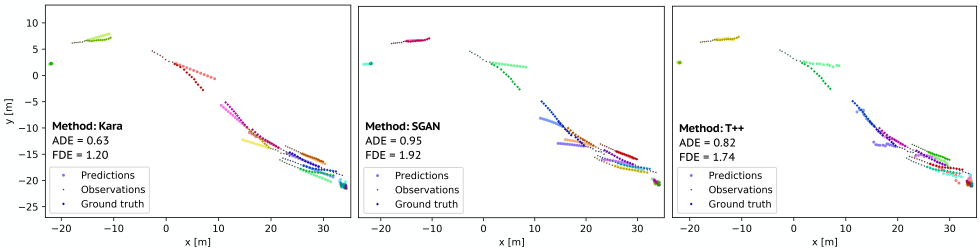


Figure 10: Predictions in the ATC scenario

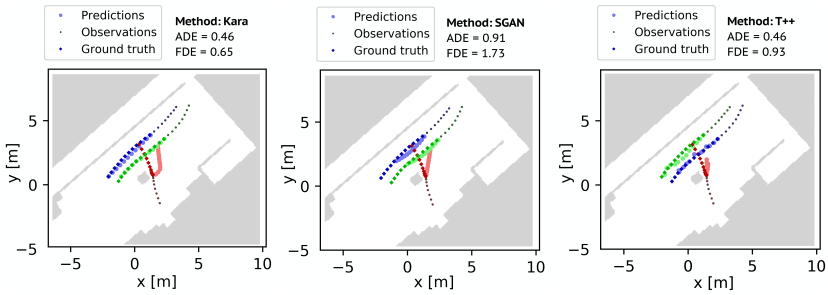


Figure 11: Predictions in the THÖR1 scenario

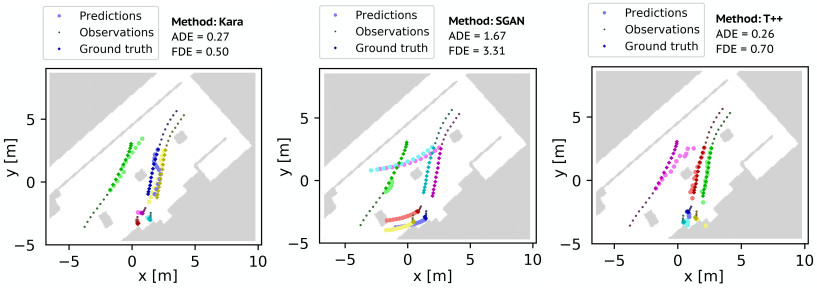


Figure 12: Predictions in the THÖR3 scenario

4 THÖR-MAGNI: A Large-scale Indoor Motion Capture Recording of Human Movement and Robot Interaction

Summary: We present a new large dataset of indoor human and robot navigation and interaction, called THÖR-MAGNI, that is designed to facilitate research on social human navigation: e.g., modeling and predicting human motion, analyzing goal-oriented interactions between humans and robots, and investigating visual attention in a social interaction context. Unlike existing datasets, THÖR-MAGNI includes a broader set of contextual features and offers multiple scenario variations to facilitate factor isolation. The dataset includes many social human-human and human-robot interaction scenarios, rich context annotations, and multi-modal data, such as walking trajectories, gaze tracking data, and lidar and camera streams recorded from a mobile robot.

4.1 Introduction

Modern approaches for modeling human motion require plentiful data recorded in diverse environments and settings to train on, as well as for the evaluation [11]. Among the growing numbers of human trajectory datasets, most focus on capturing interactions between the moving agents in indoor [24], outdoor [25], and automated driving [30] settings. These datasets are designed to study how people interact and avoid collisions in social settings by describing their motion through position and velocity information. Further datasets attempt to capture full-body motion in various activities and human-object interactions in household settings [31, 32, 33].

Human motion is influenced by many exogenous factors, which cumulatively amount to the *context* in which people move and interact. Among those are numerous environmental factors: motion and activities of other people and robots, locations of obstacles, semantic attributes such as points of common interest, direction signs, and special zones. Motion datasets should not only capture these factors to enable computational analysis of how people navigate but also vary them systematically to support factor isolation in various conditions. Datasets with access to rich context can help to better explain, model, and predict human motion.

Furthermore, beyond the environment context, there are various aspects of the specific person — *target agent cues* [11] — which are helpful in better understanding their intention, ongoing activity, attention, and distraction, preferences, and abilities. These cues include head orientations, full body positions, gaze directions, social grouping, and past activity patterns. Multi-modal approaches for human motion modeling and prediction can provide more accurate results by combining these cues [5], and their development is subject to the availability of high-quality multi-modal data.

Existing datasets in human motion analysis often lack the comprehensive inclusion of the exogenous factors and the target agent cues necessary for holistic studies of human motion dynamics. This research gap hinders the development of robust models that capture the relationship between contextual cues and human behavior in different scenarios. To address this gap, we present a novel dataset incorporating a broader set of contextual features and multiple variations to support factor isolation. By integrating diverse modalities such as walking trajectories, eye tracking data, and environmental sensory inputs captured by a mobile robot (see Figure 13), our dataset fosters the exploration and analysis of human motion in various scenarios with increased fidelity and granularity. In this paper, we propose a novel dataset of accurate human and robot navigation and interaction in diverse indoor contexts, building on the previous THÖR dataset [26].

The THÖR dataset, recorded in our previous H202 EU project ILIAD, pioneered weakly-scripted scenario-based data collection with motion capture in a controlled environment,

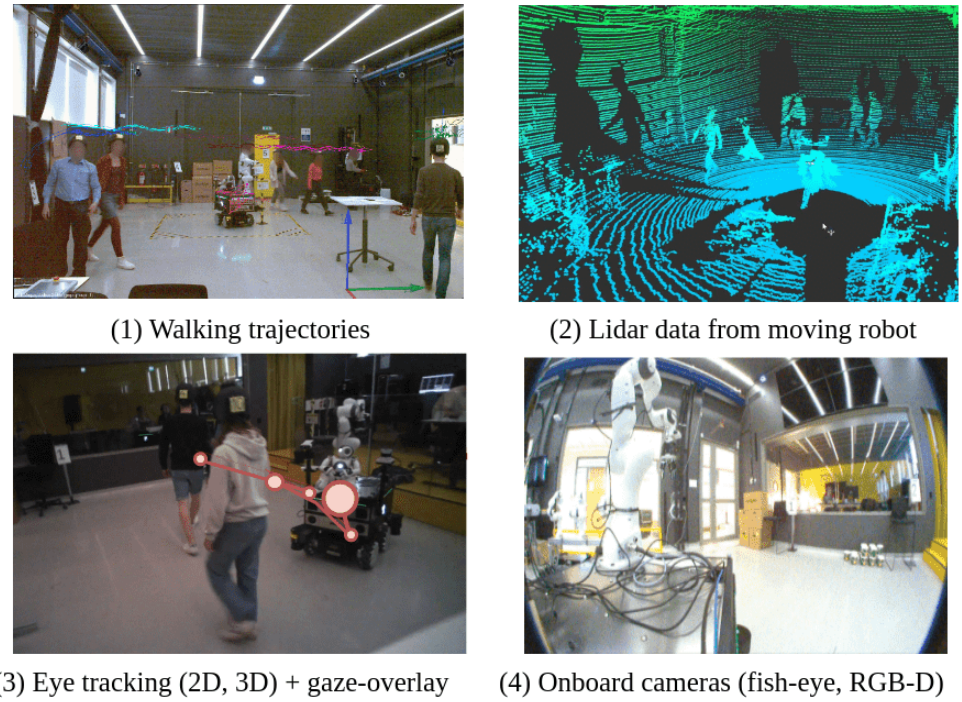


Figure 13: THÖR-MAGNI data modalities. (1) motion capture trajectories of participants in a workplace setting shared with other humans and robots; (2) lidar sweep recorded with the DARKO robot; (3) snapshot from an eye tracker's gaze overlay video; (4) fish-eye camera image from the DARKO robot, showing object stashes and two goal points from our scenarios.

recording continuous activities involving meaningful social navigation towards randomized targets in the environment. Our new THÖR-MAGNI dataset extends this effort with rich context annotations, time-synchronized multi-modal data, human-robot interaction scenarios, and diverse navigation modes of a mobile robot. The THÖR dataset established a foundation for collecting open-source data on human social navigation toward randomized targets in a controlled setting using motion capture technology with minimal scripting. In particular, the THÖR-MAGNI dataset represents a significant advancement, enhancing data quality and features to provide rich insights into human motion and interactions within a larger room. The publicly available THÖR datasets, especially THÖR-MAGNI, facilitate more comprehensive human-robot interaction and human social navigation research.

The THÖR-MAGNI data collection is designed around systematic variation of environmental factors to allow building cue-conditioned models of human motion and verifying hypotheses on factor impact. To that end, we propose several scenarios in which the participants, in addition to primary navigation, need to move objects, interact with each other and the robot, and respond to remote instructions. The dataset includes differential and omnidirectional robot navigation, semantic zones, environmental direction signs, and many other aspects. We provide position and head orientation for each moving agent, as well as 3D lidar scans and gaze tracking. Finally, we provide tools to visualize the dataset's multiple modalities and preprocess the trajectory data. In total, THÖR-MAGNI captures 3.5 hours of motion of 40 participants over five days of recording, which is available for download[4]. Furthermore, we note the continuity between the THÖR and THÖR-MAGNI recordings due to their shared environment (in diverse configurations), motion capture

[4]https://doi.org/10.5281/zenodo.10407223

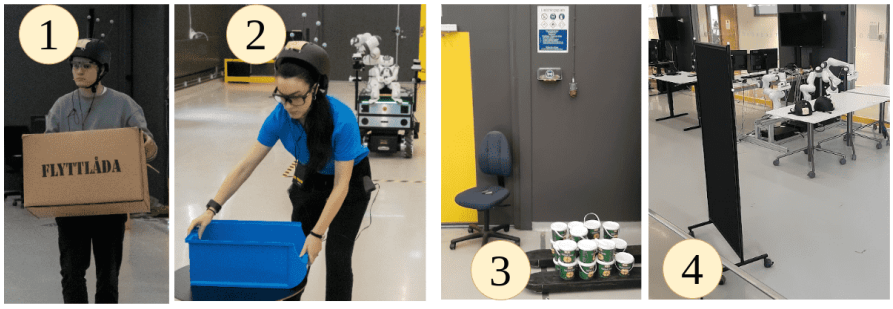


Figure 14: Participants in the role of Carrier were transporting various objects in different sizes and shapes. (1) *Carrier–Box* carrying a medium-sized card box with two hands. (2) *Carrier–Storage Bin HRI* placing the bin at a goal point (3) Stash of small objects transported by the *Carrier–Bucket* (4) Large Object (poster stand) moved by two *Carrier–Large Object*.

system, and complimentary scenario composition.

4.2 Description of the Dataset

The THÖR-MAGNI consists of 52 four-minute recordings (runs) of participants performing various activities related to navigating alone and in groups, finding and transporting small and large objects, and interacting with robots. THÖR-MAGNI contains over 3.5 hours of motion data for 40 participants, including position, velocity, and head orientation. Eye tracking data is available for 16 of them, totaling 8.3 hours for eight activities (see Table 4). In 24 runs, THÖR-MAGNI also includes the robot sensor data of 3D point clouds from an Ouster lidar. Additionally, videos recorded by an Azure Kinect camera and a Basler fish-eye camera onboard a mobile robot are available on request.

4.2.1 Environment Design

We conducted the data acquisition in a laboratory at Örebro University, the same as in the THÖR dataset [26]. The room has seven goal positions to drive purposeful human navigation through the available space, generating frequent interactions in the center. We include several environmental layouts (i.e., obstacle maps) in the THÖR-MAGNI dataset, which vary the placement of static obstacles (robotic manipulators and tables) in the room to prevent walking between goals in a straight path. Apart from static obstacles, two robots are in the room: a static robotic arm near the podium and an omnidirectional mobile robot with a robotic arm on top (see Section 4.2.2).

4.2.2 Navigation and Interaction Design

The interaction and navigation design in THÖR-MAGNI extends the weakly-scripted motion recording procedure introduced in the THÖR dataset [26]. This procedure facilitates realistic motion in controlled settings, in which accurate ground truth motion capture and eye tracking data are collected using specialized equipment. Our key idea is to assign meaningful activities and tasks to the recording's participants, allowing them to concentrate on their continuous activity during which they freely move inside the room shared with other people and robots. To generate a diverse range of interactions, we developed several scenes that vary in the composition of tasks, robot operation, and other contextual cues.

Table 4: Amount of eye tracking- and trajectory data recorded for various activities with all three devices: Tobii 2, Tobii 3, and Pupil Invisible glasses

Activity	Eye tracking (min.)	Trajectory data (min.)
Visitors–Alone	108	392
Visitors–Group 2	124	344
Visitors–Group 3	52	168
Visitors–Alone HRI	64	112
Carrier–Bucket	32	96
Carrier–Box	60	96
Carrier–Large Object	92	192
Carrier–Storage Bin HRI	16	16
Total	548	1416

Tasks, Activities and Roles Requiring Search and Navigation Aligned with the DARKO project objectives, we aimed to simulate authentic scenes that reflect the different activities individuals perform in a workplace environment. To that end, we designed several tasks that require search, navigation, and interaction with objects, other participants, and a mobile robot. Participants engaged in those tasks according to their assigned *role*.

Our dataset has two types of roles: **Visitors** and **Carriers**. **Visitors** navigate either individually (*Visitors–Alone*) or in groups of two (*Visitors–Group 2*) or three (*Visitors–Group 3*) between target points in the environment. The **Visitors** role includes a human-robot interaction component denoted by *Visitors–Alone HRI*, where participants interact with a robot in a joint navigation task (see Section 4.2.2). In addition, **Carriers** are involved in transporting various objects, including *Carrier–Bucket*, *Carrier–Box*, *Carrier–Storage Bin HRI* and *Carrier–Large Object* (see Figure 14). **Carriers** transport objects between pre-defined target points, and objects themselves representing different levels of difficulty for navigation, categorized as small (lowest difficulty), medium (medium difficulty), and large (highest difficulty).

Modes of Robot Navigation and HRI Our dataset includes the DARKO robot, which acts as a static obstacle in some scenes and moves in others. This range of behaviors enables the study of participants' movements and gaze behaviors concerning the stationary and mobile status of the robot. In certain scenes, the robot was teleoperated and moved omnidirectionally, enabling it to reach any 2D position from a stationary position. In some, it moved directionally with a predetermined orientation (front). In others, the DARKO robot navigated semi-autonomously with manually set goal points. When acting semi-autonomously, the robot interacted with participants through a communication intermediary called the "Anthropomorphic Robot Mock Driver" (ARMoD), a novel concept designed and implemented in WP5 T5.2 to facilitate natural communication with the DARKO platform.

4.2.3 Scenario Design

We address the context of agent movement by including both humans and robots, as previously discussed, in five specifically designed scenes we call "scenarios". Scenario 1 captures the dynamics of motion because of semantic attributes of the environment and

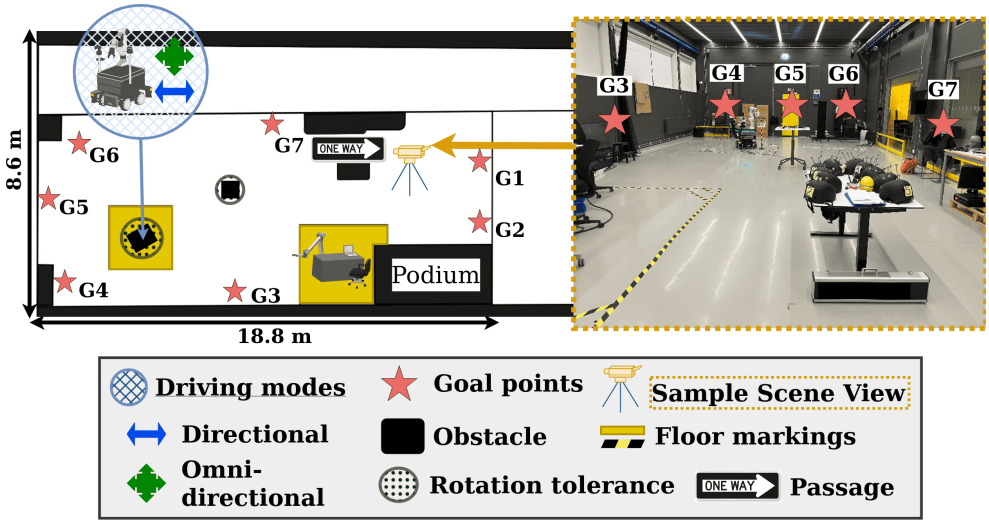


Figure 15: Varying environmental layouts for the room configuration of Scenarios 1–3. **Right:** Sample scene view for the site used for data acquisition of the THÖR-MAGNI dataset showing the room configuration for Scenarios 1–3 with the environment layout for Scenario 1B. **Left:** Overview of the room configuration and the scenario-specific layout changes. **Bottom:** Legend explaining layout elements, including driving styles for the robot in Scenario 3, semantic elements specific for Scenario 1 (Floor markings, Passage), and position of goals and obstacles. Upon placement, some objects were subject to a slight rotation between runs, which is accounted for in the layouts with the rotation tolerance.

sets up a baseline for goal-directed social human navigation. Scenario 2 adds role-specific motion for some participants navigating the environment. Subsequently, Scenario 3 explores the impact of different robot motion styles on these role-specific patterns. Figure 15 depicts a detailed overview of the room configuration and varying environmental layouts for Scenarios 1–3. Scenario 1's conditions A and B capture regular social behavior in a static environment with and without additional floor markings and a one-way passage. Scenario 2 maintains the same layout as Scenario 1A but introduces individuals performing tasks, emulating industrial activities. Scenario 3 explores human-robot interactions by varying the driving modes of the mobile robot teleoperated by experimenters on a podium.

Transitioning to a smaller room configuration, we present two scenarios to explore human motion and intended interactions between humans and robots: Scenarios 4 and 5. In Scenario 4, participants engaged in intermittent interaction with a mobile robot. This robot communicated in two interaction styles through another entity to mediate joint navigation with participants toward goal points. In Scenario 5, the robots and a human co-worker collaborated actively in transporting small storage bins. For a comprehensive overview of roles and scenarios, see Figure 16.

4.2.4 Post Processing

Multi-modal data synchronization is key to our data collection. We used ROS and custom Python scripts to align the data streams while maintaining temporal integrity. To achieve synchronicity between the motion capture and eye tracking data, we strategically placed custom events associated with precise timestamps in the two data streams using the respective software of the eye tracking devices such as Tobii Pro Lab and Pupil Player as well as the Qualisys Track Manager (QTM) for the motion capture system. This procedure resulted in CSV files where all modalities' timestamps are synchronized on the motion

Information	**Scenario 1:** Capturing Motion Dynamics in the Environment	**Scenario 2:** Role-Specific Motion Patterns in Industrial Environments	**Scenario 3:** Impact of Mobile Robot Motion on Human Behavior	**Scenario 4:** Spatial HRI and Navigation in a Shared Environment	**Scenario 5:** Spatial HRI, Proactive Robotic Assistance
Roles	*Visitors-Alone* *Visitors-Group 2* *Visitors-Group 3*	*Visitors-Alone* *Visitors-Group 2* *Visitors-Group 3* *Carrier-Box* *Carrier-Bucket* *Carrier-Large Object*	*Visitors-Alone* *Visitors-Group 2* *Visitors-Group 3* *Carrier-Box* *Carrier-Bucket* *Carrier-Large Object*	*Visitors-Alone* *Visitors-Alone HRI* *Visitors-Group 2*	*Visitors-Alone* *Visitors-Group 2* *Carrier-Storage Bin HRI*
Robot-Motion	Stationary (Obstacle)	Stationary (Obstacle)	Condition based (Teleoperated)	Directional (Semi-Autonomous)	Directional (Semi-Autonomous)
Environment-Layout					
Conditions	Condition A Layout without- Condition B with ***semantics***	No conditions	Condition A Differential- Condition B Omnidirectional-Driving	Condition A Verbal-Only HRI Condition B Mutlimodal HRI	No conditions
Duration and Recording Day	64 min. on Day 1-4	32 min. on Day 1-4	64 min. on Day 1-4	32 min. on Day 5	16 min. on Day 5

Figure 16: Scenario definitions in the THÖR-MAGNI dataset, including roles, DARKO robot motion status (e.g., autonomous or teleoperated), environment layout (i.e., obstacle maps), specific scenario conditions, and duration and recording days. Each recording day has a unique set of participants. Day 1 has nine participants; days 2-4 have seven participants each. Three mobile eye-tracking devices were used daily for three participants. On day 5, two devices were used for two sets of participants. The duration of recorded trajectory and eye-tracking data is provided in Table 4.

capture system's timestamp. Within these files, eye tracking data is available for frames where the motion capture system tracks all rigid body markers, as it is a prerequisite to determine the 3D gaze vector using a correct head orientation. The frame numbers for each respective eye tracker's scene recording are indexed in the column named "SceneFNr" in the corresponding CSV file.

To facilitate a thorough analysis of the eye tracking data in our study, we offer access to the raw data from the Tobii glasses, along with essential synchronization details. The scene recordings are provided in a blurred format to ensure data protection and removed audio data. Access to the raw data from the Pupil Invisible glasses can be granted upon individual request, providing careful and ethical distribution of sensitive data.

An extensive post-processing stage followed the data acquisitions, including synchronization and alignment. It aimed to refine and validate the collected data and ensure the protection of sensitive data. This stage involved several vital procedures, such as eliminating artifacts and noise caused by marker occlusion, lighting variations, and camera disruptions. We also rectified misidentified trajectories through spatial and temporal consistency evaluations, applying manual adjustments when needed.

4.3 Analysis of the Dataset

We deploy a set of rigorous metrics to assess the dataset's effectiveness in capturing human motion dynamics and interactions in various scenarios. Key metrics include *tracking duration*, *minimum distance between people*, and the *number of 8-second tracklets*. Tracking duration measures the average continuous tracking time for human agents, with higher values indicating more extended tracking periods, which are beneficial for long-term human motion prediction. The minimum distance between people measures the closest proximity observed between individuals, providing insight into social interactions and personal space dynamics. The number of 8-second tracklets provides a standardized measure for comparing tracklet continuity across different datasets.

The THÖR-MAGNI dataset, compared to existing human motion datasets such as ETH/UCY and THÖR, displays an advance in capturing extended and varied interactions. In particular, the participants in THÖR-MAGNI have more complex and non-linear motion, longer tracking duration, closer interactions and more variety in their velocities. Further evaluation details are presented in [9].

4.4 Conclusion

THÖR-MAGNI dataset compliments and extends the THÖR efforts, initiated as part of the previous ILIAD project. Both datasets have already proven to be instrumental in formulating and validating hypotheses about human motion and human-robot in the DARKO project, as evidently follows from the deliverables D5.1 and D5.2. THÖR-MAGNI provides valuable ground truth data on the important motion cues of the environment and individual humans, which enabled us to design and evaluate several novel methods for human motion modeling and prediction, presented in the next sections.

5 Long-term human motion prediction using Maps of Dynamics

Summary: In this section, we propose to exploit *maps of dynamics* (MoDs), a class of general representations of place-dependent spatial motion patterns, learned from prior observations) for long-term human motion prediction (LHMP). We present a new MoD-informed human motion prediction approach, named CLiFF-LHMP, which is data efficient, explainable, and insensitive to errors from an upstream tracking system. Our approach uses CLiFF-map, a specific MoD trained with human motion data recorded in the same environment. We bias a constant velocity prediction with samples from the CLiFF-map to generate multi-modal trajectory predictions. In two public datasets we show that this algorithm outperforms the state of the art for predictions over very extended periods of time, achieving 45% more accurate prediction performance at 50s compared to the baseline.

5.1 Introduction

In this section we consider methods for predicting human motion in an extended time frame, i.e. reaching beyond the typical $3-5$ s prediction horizon used by robots for collision avoidance. Such predictions are useful to associate observed tracklets in sparse camera networks, or inform the robot of the long-term environment dynamics on the path to its goal [34, 35], for instance when following a group of people [36]. Very long-term predictions are useful for global motion planning to produce socially-aware unobtrusive trajectories, and for coordinating connected multi-robot systems with sparse perception fields.

Human motion is complex and may be influenced by several hard-to-model factors, including social rules and norms, personal preferences, and subtle cues in the environment that are not represented in geometric maps. Accordingly, accurate motion prediction is very challenging [11]. Prediction on the very long-term scale (i.e., over 20 s into the future) is particularly hard as complex, large-scale environments influence human motion in a way that cannot be summarized and contained in the current state of the moving person or the observed interactions but rather have to be modelled explicitly [37].

To predict very long-term human motion, we exploit *maps of dynamics* (MoDs) that encode human dynamics as a feature of the environment. Specifically, we use Circular Linear Flow Field map (CLiFF-map) [38], which captures multimodal statistical information about human flow patterns in a continuous probabilistic representation over velocities. The motion patterns represented in a CLiFF-map implicitly avoid collisions with static obstacles and follow the topological structure of the environment, e.g., capturing the dynamic flow through a hall into a corridor (see Fig. 17). Our CLiFF-LHMP approach predicts stochastic trajectories by sampling from a CLiFF-map to guide a velocity filtering model [1]. Examples of prediction results are shown in Fig. 17.

In qualitative and quantitative experiments we demonstrate our CLiFF-LHMP approach is 45% more accurate than the baseline at 50 s, with average displacement error (ADE) below 5 m up to 50 s. In contrast to prior art in long-term environment-aware motion prediction [37], our method does not make any assumptions on the optimality of human motion and instead generalizes the features of human-space interactions from the learned MoD. Furthermore, our method does not require a list of goals in the environment as input, in contrast to prior planning-based prediction methods. Finally, our method can flexibly estimate the variable time end-points of human motion, predicting both short- and long-term trajectories, in contrast to the prior art which always predicts up to a fixed prediction horizon.

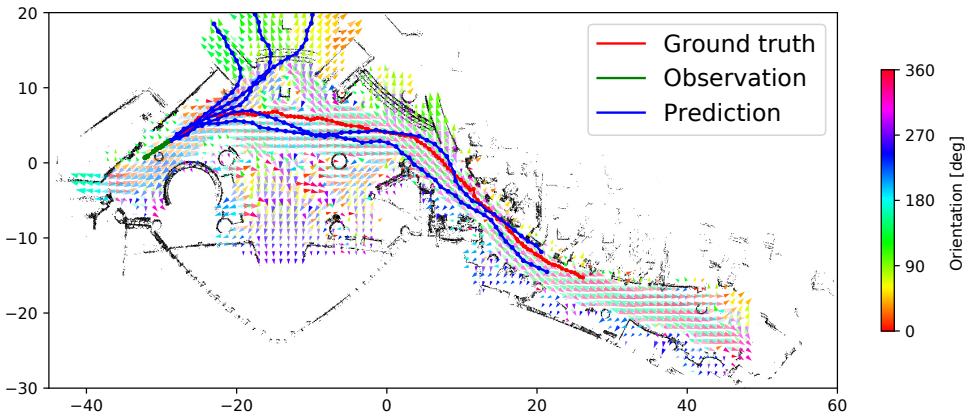


Figure 17: Long-term (50 s) motion prediction result obtained with CLiFF-LHMP for one person in the ATC dataset. **Red** line: ground truth trajectory. **Green** line: observed trajectory. **Blue** lines: predicted trajectories. The CLiFF-map is shown with colored arrows.

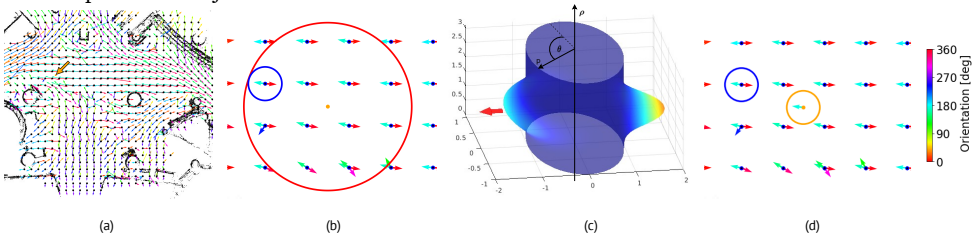


Figure 18: Steps of sampling a direction θ_s from the CLiFF-map. **(a)** CLiFF-map built from the ATC data. The location to sample from is marked with an orange arrow. **(b)** Selection of SWGMMs in the CLiFF-map: The red circle contains all SWGMMs within r_s distance to the sampling location. From these SWGMMs, the SWGMM with the highest motion ratio is selected (marked with a blue circle). **(c)** The SWGMM distribution in the selected location wrapped on a unit cylinder. The speed is represented by the position along the ρ axis and the direction is θ. The probability is represented by the distance from the surface of the cylinder. A velocity vector (marked with a red arrow) is sampled from this SWGMM. **(d)** The direction value θ_s of the sampled velocity is shown in the sampled direction and marked with an orange circle.

5.2 Method

In this section, we first describe the CLiFF-map representation for site-specific motion patterns and then present the CLiFF-LHMP approach for single-agent long-term motion prediction exploiting the information accumulated in a CLiFF-map.

Circular-Linear Flow Field Map (CLiFF-map): To predict human trajectories we exploit the information about local flow patterns represented in a CLiFF-map as a multimodal, continuous distribution over velocities. CLiFF-map [38] is a probabilistic framework for mapping velocity observations (independently of their underlying physical processes), i.e., essentially a generalization of a vector field into a Gaussian mixture field.

Each location in the map is associated with a Gaussian mixture model (GMM). A CLiFF-map represents motion patterns based on local observations and estimates the likelihood of motion at a given query location.

CLiFF-maps represent speed and direction jointly as velocity $\mathbf{V} = [\theta, \rho]^T$ using direction θ and speed ρ, where $\rho \in \mathbb{R}^+$, $\theta \in [0, 2\pi)$.

As the direction θ is a circular variable and the speed is linear, a mixture of *semi-wrapped* normal distributions (SWNDs) is used in CLiFF-map. At a given location, the

Algorithm 1: CLiFF-LHMP

Input: $\mathscr{H}$, x_{t_0}, y_{t_0}, Ξ
Output: $\mathscr{T}$
1 $\mathscr{T} = \{\}$
2 $\rho_{\text{obs}}, \theta_{\text{obs}} \leftarrow$ getObservedVelocity($\mathscr{H}$)
3 $s_{t_0} = (x_{t_0}, y_{t_0}, \rho_{\text{obs}}, \theta_{\text{obs}})$
4 **for** $t = t_0 + 1, \ldots, t_0 + T_p$ **do**
5 $\quad x_t, y_t \leftarrow$ getNewPosition(s_{t_1})
6 $\quad \theta_s \leftarrow$ sampleDirectionFromCLiFFmap(x_t, y_t, Ξ)
7 $\quad (\rho_t, \theta_t) \leftarrow$ predictVelocity(θ_s, ρ_{t_1}, θ_{t_1})
8 $\quad s_t \leftarrow (x_t, y_t, \rho_t, \theta_t)$
9 $\quad \mathscr{T} \leftarrow \mathscr{T} \cup s_t$
10 **return** $\mathscr{T}$

semi-wrapped probability density function (PDF) over velocities can be visualized as a function on a cylinder. Direction values θ are wrapped on the unit circle and the speed ρ runs along the length of the cylinder. An SWND $\mathscr{N}^{SW}_{\boldsymbol{\Sigma},\boldsymbol{\mu}}$ is formally defined as $\mathscr{N}^{SW}_{\boldsymbol{\Sigma},\boldsymbol{\mu}}(\mathbf{V}) = \sum_{k\in\mathbb{Z}} \mathscr{N}_{\boldsymbol{\Sigma},\boldsymbol{\mu}}([\theta,\rho]^T + 2\pi[k,0]^T)$, where $\boldsymbol{\Sigma}, \boldsymbol{\mu}$ denote the covariance matrix and mean value of the directional velocity $(\theta,\rho)^T$, and k is a winding number. Although $k \in \mathbb{Z}$, the PDF can be approximated adequately by taking $k \in \{-1, 0, 1\}$ for practical purposes. To preserve the multi-modal characteristic of the flow, a semi-wrapped Gaussian mixture model (SWGMM) is used, which is a PDF represented as a weighted sum of J SWNDs: $p(\mathbf{V}|\xi) = \sum_{j=1}^{J} \pi_j \mathscr{N}^{SW}_{\boldsymbol{\Sigma}_j,\boldsymbol{\mu}_j}(\mathbf{V})$, where $\boldsymbol{\xi} = \{\xi_j = (\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j, \pi_j) | j \in \mathbb{Z}^+\}$ denotes a finite set of components of the SWGMM, and π_j denotes the mixing factor and satisfies $0 \le \pi_j \le 1$.

Human Motion Prediction Using CLiFF-map: We frame the task of predicting a person's future trajectory as inferring a sequence of future states. The algorithm is presented in Alg. 1. With the input of an observation history of O_p past states of a person and a CLiFF-map Ξ, the algorithm predicts T_p future states. The length of the observation history is $O_s \in \mathbb{R}^+$ s, equivalent to $O_p > 0$ observation time steps. With the current time-step denoted as the integer $t_0 \ge 0$, the sequence of observed states is $\mathscr{H} = \langle s_{t_0-1}, \ldots, s_{t_0-O_p} \rangle$, where s_t is the state of a person at time-step t. A state is represented by 2D Cartesian coordinates (x, y), speed ρ and direction θ: $s = (x, y, \rho, \theta)$.

From the observed sequence $\mathscr{H}$, we derive the observed speed ρ_{obs} and direction θ_{obs} at time-step t_0 (line 2 of Alg. 1). Then the current state becomes $s_{t_0} = (x_{t_0}, y_{t_0}, \rho_{\text{obs}}, \theta_{\text{obs}})$ (line 3 of Alg. 1). The values of ρ_{obs} and θ_{obs} are calculated as a weighted sum of the finite differences in the observed states, as in the Atlas benchmark (Sec. 3.3) [1]. With the same parameters as in [1], the sequence of observed velocities is weighted with a zero-mean Gaussian kernel with $\sigma = 1.5$ to put more weight on more recent observations, such that $\rho_{\text{obs}} = \sum_{t=1}^{O_p} v_{t_0-t}\, g(t)$ and $\theta_{\text{obs}} = \sum_{t=1}^{O_p} \theta_{t_0-t}\, g(t)$, where $g(t) = (\sigma\sqrt{2\pi} e^{\frac{1}{2}(\frac{t}{\sigma})^2})^{-1}$.

Given the current state s_{t_0}, we estimate a sequence of future states. Similar to past states, future states are predicted within a time horizon $T_s \in \mathbb{R}^+$ s. T_s is equivalent to $T_p > 0$ prediction time steps, assuming a constant time interval Δt between two predictions. Thus, the prediction horizon is $T_s = T_p \Delta t$. The predicted sequence is then denoted as $\mathscr{T} = \langle s_{t_0+1}, s_{t_0+2}, \ldots, s_{t_0+T_p} \rangle$.

To estimate $\mathscr{T}$, for each prediction time step, we sample a direction from the CLiFF-map at the current position (x_t, y_t) to bias the prediction with the learned motion patterns represented by the CLiFF-map. The main steps for each iteration are shown in lines 5–9 of Alg. 1.

Algorithm 2: sampleDirectionFromCLiFFmap(x, y, Ξ)

Input: x, y, Ξ
Output: θ_s
1 $\Xi_{\text{near}} \leftarrow$ getNearSWGMMs(x, y, Ξ)
2 $\xi \leftarrow$ selectSWGMM(Ξ_{near})
3 $\theta_s \leftarrow$ sampleDirectionFromSWGMM(ξ)
4 **return** θ_s

For each iteration, we first compute the predicted position (x_t, y_t) at time step t from the state at the previous time step (line 5 of Alg. 1):

$$\begin{aligned} x_t &= x_{t-1} + \rho_{t-1} \cos\theta_{t-1} \Delta t, \\ y_t &= y_{t-1} + \rho_{t-1} \sin\theta_{t-1} \Delta t, \end{aligned} \tag{1}$$

Afterwards, we estimate the new speed and direction using constant velocity prediction biased by the CLiFF-map. The bias impacts only the estimated direction of motion, speed is assumed to be unchanging.

To estimate direction at time t, we sample a direction from the CLiFF-map at location (x_t, y_t) in the function `sampleDirectionFromCLiFFmap()` (line 6 of Alg. 1). Alg. 2 outlines its implementation. The inputs of Alg. 2 are: the sample location (x, y) and the CLiFF-map Ξ of the environment. The sampling process is illustrated in Fig. 18. To sample a direction at location (x, y), from Ξ, we first get the SWGMMs Ξ_{near} whose distances to (x, y) are less than the sampling radius r_s (line 1 of Alg. 2). In a CLiFF-map, each SWGMM is associated with a motion ratio. To sample from the location with the highest intensity of human motions, in line 2, from Ξ_{near}, we select the SWGMM ξ with highest motion ratio. In line 3 of Alg. 2, from ξ, an SWND is sampled from the selected SWGMM, based on the mixing factor π. A velocity is drawn randomly from the sampled SWND. Finally, the direction of the sampled velocity is returned and used for motion prediction.

With the direction sampled from the CLiFF-map, we predict the velocity (ρ_t, θ_t) in line 7 of Alg. 1 assuming that a person tends to continue walking with the same speed as in the last time step, $\rho_t = \rho_{t-1}$, and bias the direction of motion with the sampled direction θ_s as:

$$\theta_t = \theta_{t-1} + (\theta_s - \theta_{t-1}) \cdot K(\theta_s - \theta_{t-1}), \tag{2}$$

where $K(\cdot)$ is a kernel function that defines the degree of impact of the CLiFF-map. We use a Gaussian kernel with a parameter β that represents the kernel width:

$$K(x) = e^{-\beta \|x\|^2}. \tag{3}$$

In the end of each iteration, we add s_t to the predicted trajectory $\mathcal{T}$ (line 9 of Alg. 1) and update t for the next iteration. After iterating for T_p times, the output is a sequence $\mathcal{T}$ of future states that represents the predicted trajectory.

5.3 Experiments

Accurate map-aware long-term motion predictions are typically addressed with Markov Decision Process (MDP) based methods [39, 40, 41, 42, 37]. Among them, as the baseline for CLiFF-LHMP, we chose the recent IS-MDP approach [37], developed by the consortium members in the previous ILIAD project. We also compare our method with the constant velocity predictor [13, 1]. We evaluate the predictive performance of our method using the **ATC** [24] and **THÖR** [26] datasets.

Dataset	Horizon	ADE / FDE (m)		
		CLiFF-LHMP	IS-MDP	CVM
ATC	50 s	**4.6** / **9.6**	8.4 / 21.3	12.4 / 27.1
THÖR1	12 s	**1.5** / **2.6**	1.6 / 3.5	1.8 / 3.8
THÖR3	12 s	**1.3** / **2.6**	1.5 / 3.6	2.8 / 6.1

Table 5: Long-term prediction horizon results on different datasets. With $O_s = 3.2$ s, error reported are ADE/FDE in meters.

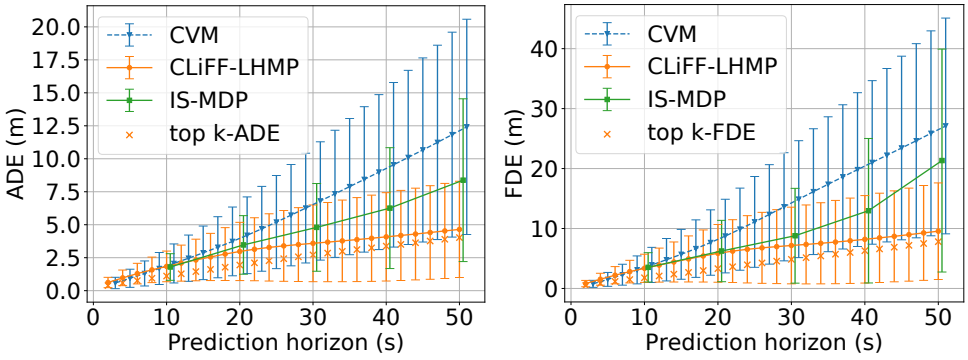


Figure 19: ADE/FDE (mean ± one std. dev.) in the ATC dataset with prediction horizon 1–50 s.

Quantitative Results: We show that our approach is substantially better than IS-MDP when the prediction horizon is above 20 s since it implicitly exploits location-specific motion patterns, thus overcoming a known limitation of MDP-based methods [37]. At 50 s in the ATC dataset, our method achieves a 45% ADE and 55% FDE improvement in performance compared to IS-MDP. At 12 s in THÖR1 and THÖR3, our method achieves an improvement of 6.3% and 13.3% ADE (25.7%, 27.8% FDE) over IS-MDP, respectively.

Further evaluation details are available in [3, 2], in Table 5 and Figures 19–20.

Qualitative Results: Fig. 21, 22 show qualitative results with example predictions. Our approach correctly captures the motion patterns in each scenario, utilizing the environment information during the prediction. Fig. 22 shows that the predicted trajectories avoid the obstacles, even though an obstacle map is not used for predictions. Furthermore, using maps of dynamics built from the observations of human motion makes it possible to predict motion through regions which appear as obstacles in an occupancy map, for example across stairs and through narrow passages (see Fig. 21). Similarly, using the MoD input keeps predictions in more intensively used areas of the environment, avoiding semantically-insignificant and empty regions, e.g., corners of the room (see Fig. 22).

Real world demonstrations of the CLiFF-LHMP approach on the DARKO robot are also shown in Fig. 38.

5.4 Extensions of the MoD-LHMP methods

5.4.1 Time-conditioned MoD-LHMP

Human flows in real environments exhibit distinct patterns in different times of the day. In the general CLiFF-LHMP method, one CLiFF-map is used for all predicted trajectories,

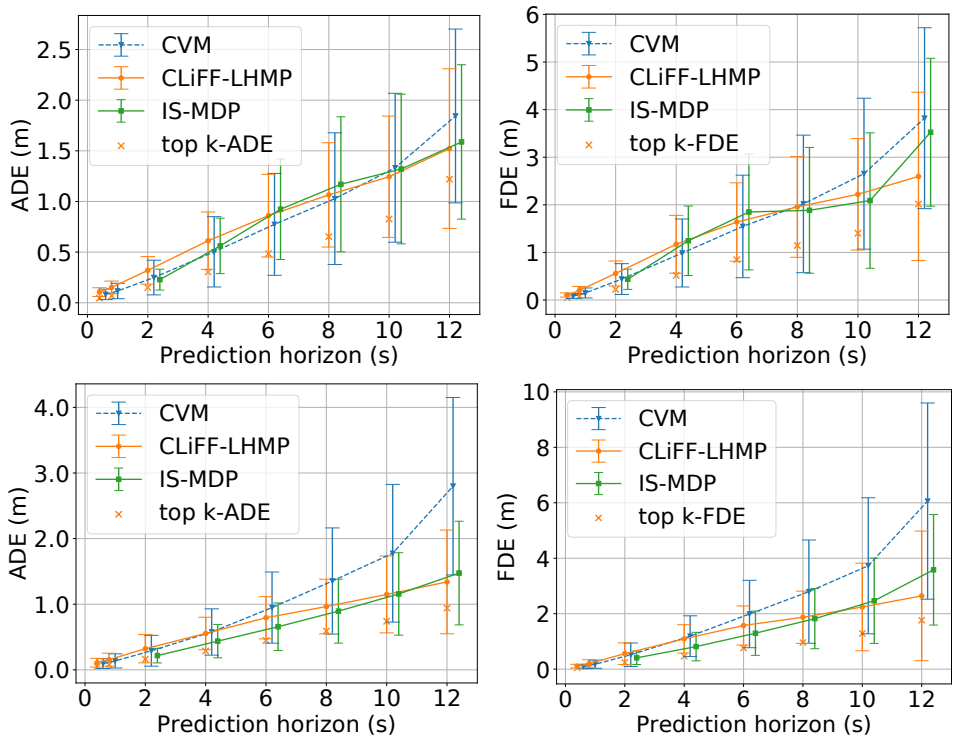


Figure 20: ADE/FDE (mean ± one std. dev.) in the THÖR1 **(top)** and THÖR3 **(bottom)** dataset with prediction horizon 0.4–12 s.

no matter what time the trajectory happens. However, motion patterns of human can vary over time. In this work, we present time-conditioned CLiFF-maps to represent motion patterns in different time in a day.

One day is divided into n time intervals. For each fixed time interval, one CLiFF-map is trained using the trajectories appear over that time interval. For one day, time-conditioned CLiFF-map consists of n CLiFF-maps are generated corresponding to n time intervals. Fig. 23 shows the CLiFF-map of 10:00, 14:00 and 18:00 in the first day of ATC dataset. From CLiFF-maps in Fig. 23, the change of human motion pattern over a day can be visualized clearly. To visualize how human motion patterns vary by hour in a day, as an example, Fig. 24 shows the CLiFF-map in one location of east corridor in ATC environment. With time-conditioned CLiFF-map, human motion can be represented more accurately compared with general CLiFF-map, which is trained on data over the whole day.

To predict human movement, the current time of a person, t_0, decides which time interval the future trajectory appears at. The corresponding time-conditioned CLiFF-map is then used for prediction. After choosing the corresponding CLiFF-map in time-conditioned CLiFF-maps, the rest of the process is same as with CLiFF-LHMP. The proposed method improves performance in presence of timed events, such as the morning and evening rushes, lunch hours, etc.

5.4.2 Class-conditioned MoD-LHMP

DARKO robot is expected to operate in real complex environments, shared with diverse agents which may be involved in dedicated activities, have heterogeneous motion patterns or belong to different classes (e.g. walking humans, delivery vehicles, other types of robots,

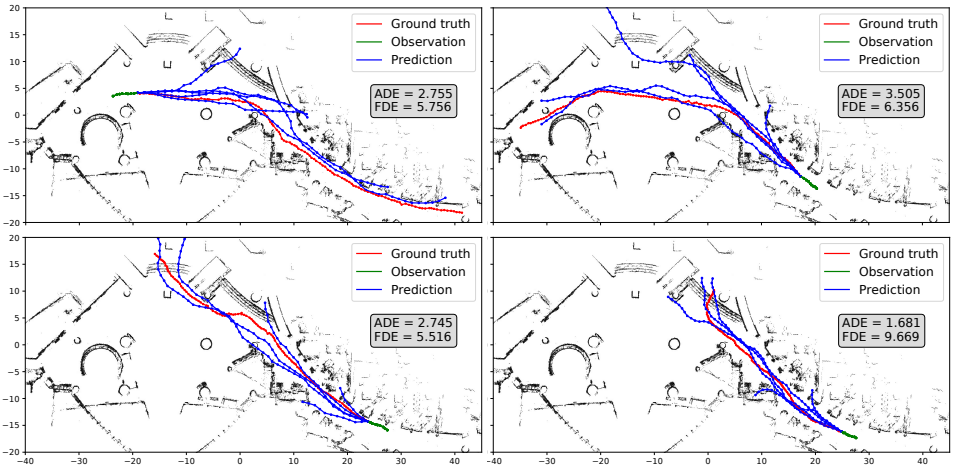


Figure 21: Predictions in ATC with $T_s = 50$ s. **Red** line shows the ground truth trajectory. **Green** line shows the observed trajectory and **blue** lines show the predicted trajectories. Note that we correctly predict trajectories crossing obstacles such as stairs (top of the map) and exits (left of the map).

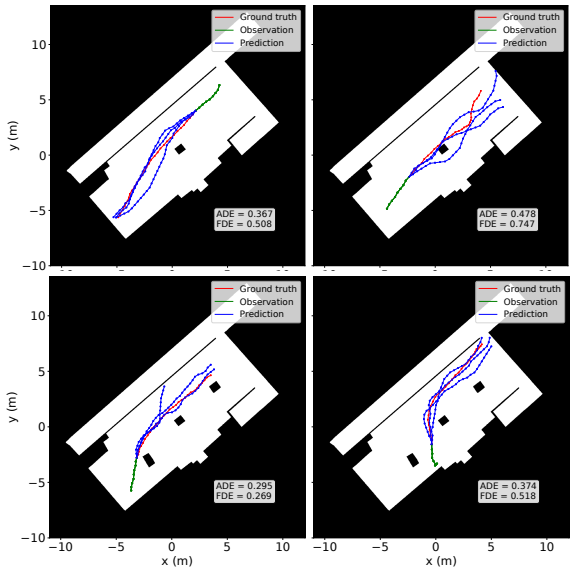


Figure 22: Predictions in THÖR1 **(top)** and THÖR3 **(bottom)** with $T_s = 12$ s. **Red** line shows the ground truth trajectory. **Green** line shows the observed trajectory and **blue** lines show the predicted future trajectories

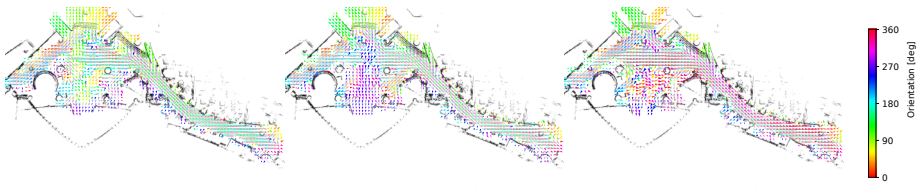


Figure 23: Time-conditioned CLiFF-map for 10:00 (**left**), 14:00 (**middle**) and 18:00 (**right**). Colored arrow shows the mean value of the component with maximum weight in CLiFF-map.

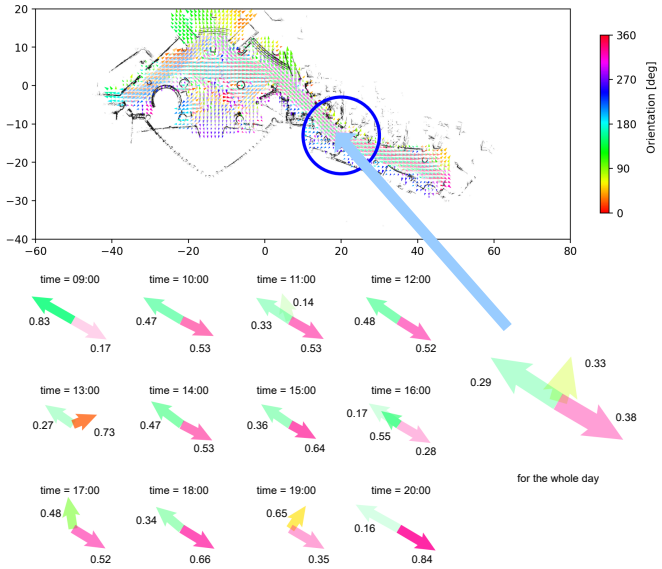


Figure 24: CLiFF-maps at one example location of east corridor. For each hour between 9:00 to 21:00, time-conditioned CLiFF-map of the example location is shown, together with general CLiFF-map of the whole day at the same location. The colored arrow shows the mean value of each component in the Semi-Wrapped Gaussian Mixture model (SWGMM), which represents speed and orientation jointly with a multimodal distribution. The weight of each component is shown and the colored arrow gets more transparent as the weight of the component is larger.

etc). Class- or activity-conditioned motion prediction is thus an appealing way to reduce forecast uncertainty and get more accurate predictions for heterogeneous agents. However, this is hardly explored in the prior art, especially for mobile robots and in limited data applications.

In CLiFF-LHMP, a single CLiFF-map is used for all predicted trajectories, irrespective of the agent class. However, their motion patterns often differ, as detailed in Fig. 25. In DARKO, we analyse different class-conditioned trajectory prediction methods on two datasets. We propose a set of conditional pattern-based and efficient deep learning-based baselines, and evaluate their performance on robotics and outdoors datasets (THÖR-MAGNI with diverse activities in difference roles and Stanford Drone Dataset with diverse agent classes).

Our experiments show that all methods improve accuracy in most of the settings when considering class labels, as shown in qualitative trajectory prediction comparisons in Fig. 26. More importantly, we observe that there are significant differences when learning from imbalanced datasets, or in new environments where sufficient data is not available.

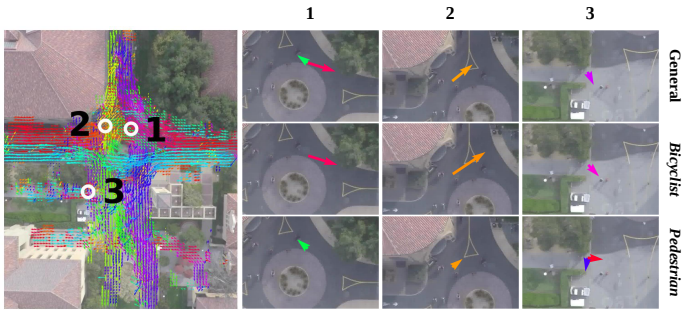


Figure 25: CLiFF-maps at example locations in SDD [**robicquet16**]. Both general and class-conditioned CLiFF-maps of *Bicyclist* and *Pedestrian* of three locations are shown on the **right**. General CLiFF maps may depict combinations of multiple classes (point 1) or median speed and orientation (points 2 and 3).

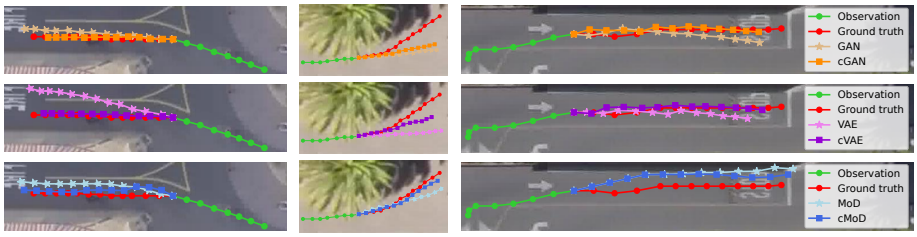


Figure 26: Prediction examples of *Bicyclist* (**left**), *Pedestrian* (**middle**) and *Car* (**right**) in SDD with 4.8 s prediction horizon.

In particular, we find that deep learning methods perform better on balanced datasets, but in applications with limited data, e.g., cold start of a robot in a new environment, or imbalanced classes, pattern-based methods may be preferable.

5.4.3 LaCE-LHMP

Detecting and identifying abnormal trajectories is a major challenge in motion modelling and prediction. Existing methods typically identify abnormal motions by comparing them to expected behaviors [43] or measuring deviations from normal motions [44]. However, these approaches require labelled data for supervised learning.

The modelling approach based on the CLiFF map may struggle to differentiate dominant human flow from irregular motion, and therefore the prediction accuracy may be affected by anomalous data.

To address the limitations of prior work, we propose the Laminar Component Enhanced LHMP approach (LaCE-LHMP). Our approach is inspired by data-driven airflow modelling, which estimates laminar and turbulent flow components and uses predominantly the laminar components to make flow predictions. Based on the hypothesis that human trajectory patterns also manifest laminar flow (that represents predictable motion) and turbulent flow components (that reflect more unpredictable and arbitrary motion), LaCE-LHMP extracts the laminar patterns in human dynamics and uses them for human motion prediction.

As shown in Table 6 and Fig. 28, LaCE-LHMP performs on par with CLiFF-LHMP in the short-term perspective, and outperforms it by 6.0% ADE and 6.4% FDE.

To evaluate the relation between prediction performance and the degree of laminar dominance in the environment, we present a heatmap of FDE values of our approach

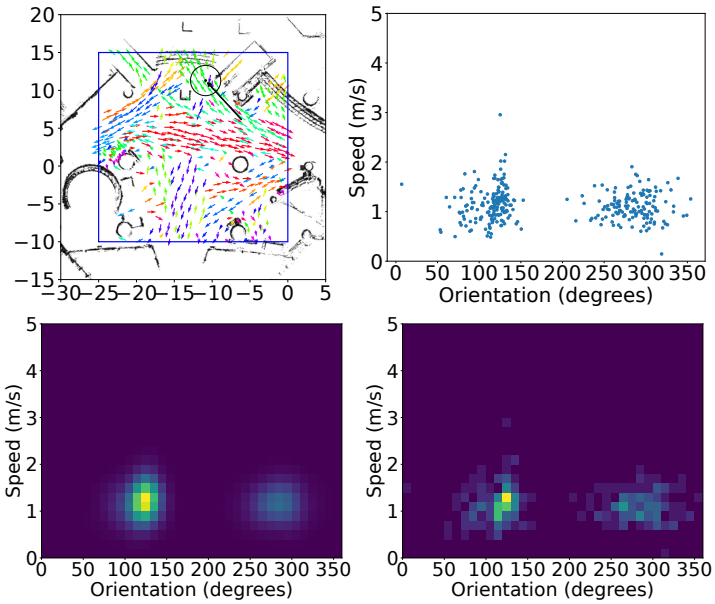


Figure 27: Example of laminar component extraction in LaCE-LHMP. **Upper-left**: LaCE model of a location in a shopping mall. Colored arrows show flow directions with highest likelihoods; **Upper-right**: raw data (velocity observations) in the $\omega - v$ domain (i.e. speed and orientation) at a specific location; **Lower-left:** histogram of the raw data Γ^R; **Lower-right**: extracted laminar component Γ^L. The laminar component is used for motion prediction in LaCE-LHMP.

Method	**ADE / FDE**	**Top-k ADE / FDE**
CVM	4.26 / 9.01	-
Trajectron++	6.09 / 12.86	**2.96** / **5.86**
CLiFF-LHMP	3.52±0.009 / 7.40±0.021	3.00 / 6.09
LaCE-LHMP (Ours)	**3.31±0.006** / **6.93±0.013**	3.00 / 6.13

Table 6: Long-term prediction (20 s) results on the ATC dataset. With $O_s = 3$ s, errors are reported as ADE/FDE in meters.

for prediction horizon 20 s in Fig. 29. In laminar-dominated regions, predictions made using the LaCE model are more accurate than in regions with more turbulent patterns, indicating that the former are more predictable.

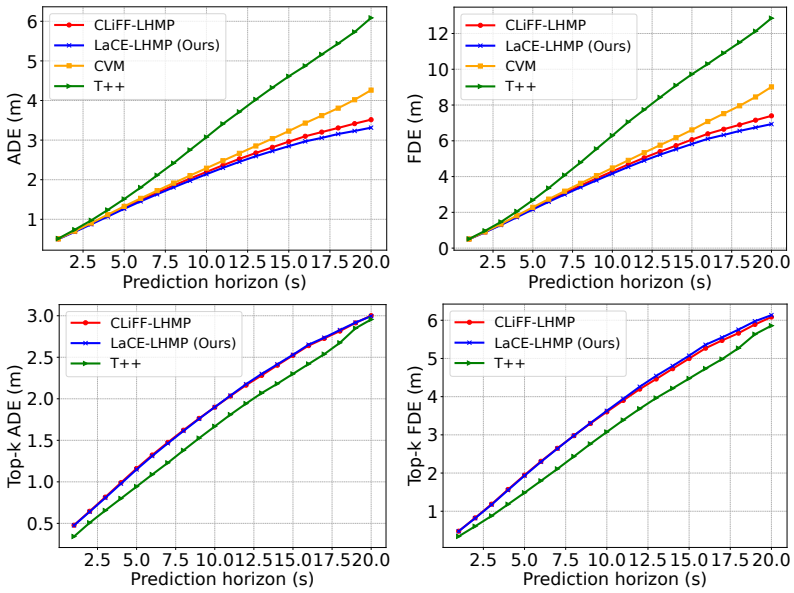


Figure 28: ADE/FDE (**top**) and top-k ADE/FDE (**bottom**) in the ATC dataset with a prediction horizon 1–20 s. Predictions with the LaCE model are more accurate during the whole considered period, as indicated by lower ADE/FDE values, which signify improved performance.

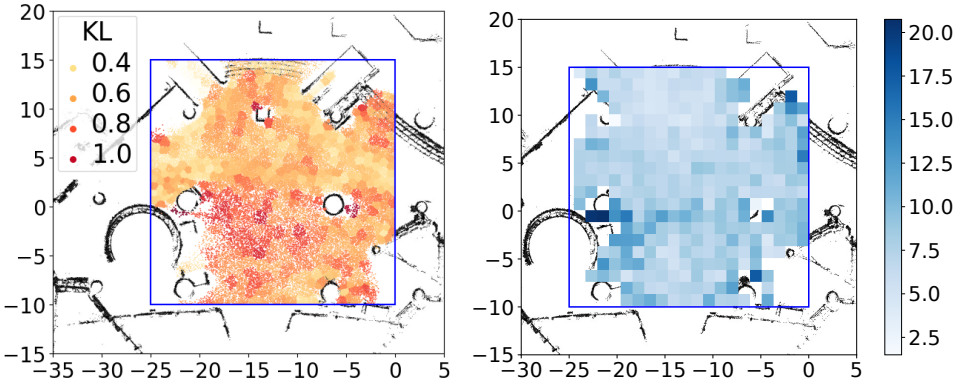


Figure 29: Left: KL divergence between Γ^R and Γ^L. **Right:** A heatmap illustrating the FDE values of LaCE-LHMP in the ATC dataset, with a prediction horizon of 20 s. Predictions exhibit higher accuracy in the central region. Predictions exhibit higher accuracy in the central region, which is predominantly laminar, as indicated by lower KL divergence.

6 Unified 3D Human Pose Dynamics and Trajectory Prediction

Summary: Human 3D body pose is a strong descriptor of walking dynamics and intention. It gives the robot a finer representation of space occupied by the person in comparison to a geometric point on the top-down 2D map with a standard collision avoidance radius. In this section, we introduce a unified approach to forecast the dynamics of human keypoints and the motion trajectory from a short input sequence of poses, developed as a joint effort with WP2 T2.5. We utilize the 3D human pose estimation input from WP2 followed by a graph attention network to encode the skeleton structure. We propose a novel motion transformation technique to predict full-body 3D joint positions directly in a global coordinate frame. The backbone of our prediction architecture is a compact non-autoregressive Transformer, capable of accurate and real-time pose and trajectory prediction. In an extensive evaluation, including public and novel datasets with specific focus on activities that are relevant for mobile robot navigation, we show our approach to be faster and more accurate in pose and trajectory prediction compared to the prior art.

6.1 Introduction

Refining trajectory predictions with full-body poses gives complete information about human behaviour with many promising applications in human-robot interaction [45, 46], automated driving [47], surveillance [48] and healthcare [49].

Historically speaking, the research in trajectory prediction and full-body pose prediction, with a few notable exceptions [50, 45, 51], progressed independently and targeted distinct application scenarios. There are, however, benefits of solving the task in a unified manner: considering the gait, head orientation and other full-body pose features can refine the dynamics modeling in trajectory prediction [52, 11], but also the full-body poses can be predicted more accurately when rooted in global coordinates with respect to the walking surface [45]. Furthermore, full-body pose prediction in trajectory coordinates can improve robot response in approach and handover applications, and better plan the collision avoidance maneuvers in close proximity to the robot [53].

Prior art addressed the problem of pose and trajectory prediction as two separate tasks [54, 55] or solved it in a decoupled manner with separate modules [56, 45]. Furthermore, only a handful of works specifically focused on navigation activities, which are of critically interest for predictive planning of the DARKO robot, considered over the more diverse actions without distinct locomotion. This challenge is also reflected in the prior art datasets of full-body motion, which are often dominated by static activities such as bending, reaching, handing over, standing up, etc.

In this section we present a novel Unified human Pose and Trajectory predictor ("UPTor"). We propose to encode the skeleton features using a graph attention network and use a non-autoregressive Transformer model [55, 45] with pose input sequence in global 3D coordinates. To support this form of input, we propose a coordinate transformation technique applied to the training sequences and during inference. We evaluate our method on the H3.6M [57] and CMU-Mocap [58] datasets, and contribute a novel DARKO dataset of 17 subjects performing diverse navigation-related activities. In qualitative and quantitative experiments we show that our method is more accurate in full-body pose and trajectory prediction of walking people, while being faster and more compact than prior art.

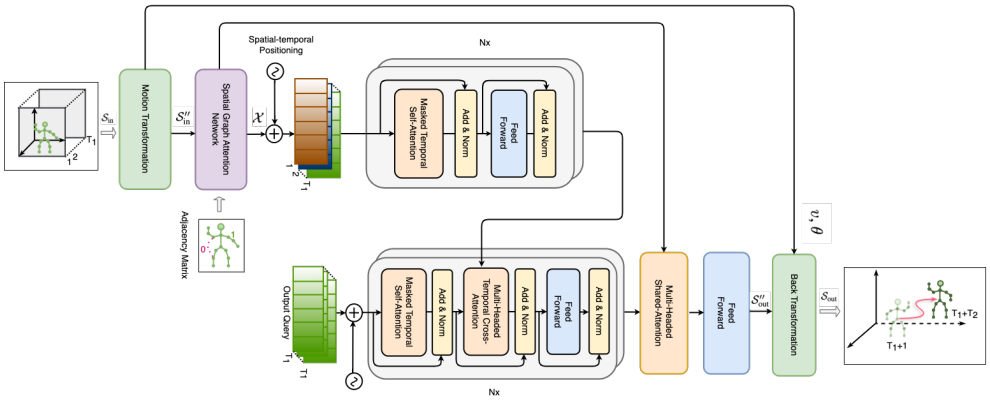


Figure 30: UPTor: Unified 3D Human Pose Dynamics and Trajectory Prediction Transformer

6.2 Methodology

6.2.1 Problem Formulation

Let $\mathscr{P}(t) \in \mathbb{R}^{3N}$ denote the 3D human pose at time t comprising N joints: $\mathscr{P}(t) = \{j_1(t), j_2(t), \ldots, j_N(t)\}$ where each $j_i(t) \in \mathbb{R}^3$ represents the (x, y, z) coordinates of the i^{th} joint in the global coordinate frame at time t. We define an input sequence as a set of poses from time 1 to time T_1: $\mathscr{S}_{\text{in}} = \{\mathscr{P}(1), \mathscr{P}(2), \ldots, \mathscr{P}(T_1) \in \mathbb{R}^{T_1 \times 3N}$. The objective of the model is to predict sequence of poses from $T_1 + 1$ to $T_1 + T_2$ in global coordinate frame: $\mathscr{S}_{\text{out}} = \{\mathscr{P}(T_1 + 1), \ldots, \mathscr{P}(T_1 + T_2)\} \in \mathbb{R}^{T_2 \times 3N}$. The complete motion sequence $\mathscr{S}$ is given by $\mathscr{S} = \mathscr{S}_{\text{in}} \cup \mathscr{S}_{\text{out}}$.

6.2.2 Model Architecture

Motion Transformation: We address the challenge of training motion sequences from global coordinate frame with varying initial positions and motion orientations. To that end, we propose a systematic method to normalize the motion sequences and achieve global and orientation invariance, as summarized in Fig. 31.

Global invariance: To ensure that our predictions commence from consistent coordinates, we translate the entire motion sequence using the translation vector v, derived as the negative counterpart of the root joint position at the last pose of the input sequence $j_{\text{root}}(T_1)$ i.e., $v = -j_{\text{root}}(T_1)$. We add this translation vector to every pose in the sequence to translate the entire sequence towards origin. The resulting translated poses are formally defined as:

$$\mathscr{P}'(t) = \mathscr{P}(t) + v \quad \forall t \in [1, T_1 + T_2] \tag{4}$$

Through this translation, the last human pose of our input sequence is anchored to the origin with its root joint at $(0, 0, 0)$ thus providing a uniform initiation for subsequent motion predictions.

Orientation invariance: To align various motion directions along with the positive x-axis, the rotation angle, θ is computed in radians between the input motion direction and the positive x-axis. This angle is calculated as the arctangent of the ratio of the differences in the y and x coordinates of the root joint's position at the last input pose T_1 and the pose at $(T_1 - \delta)$, with δ being a predetermined interval to measure motion direction at the end of the input horizon.

$$\theta = \arctan 2(\Delta y, \Delta x) \begin{cases} \Delta x = j_{\text{root}}(T_1)_x - j_{\text{root}}(T_1 - \delta)_x \\ \Delta y = j_{\text{root}}(T_1)_y - j_{\text{root}}(T_1 - \delta)_y \end{cases} \tag{5}$$

Finally, to rotate the entire sequence, we take the dot product of the rotation matrix around the z-axis, R_z with all translated poses, $\mathscr{P}'(t)$, resulting in a motion transformed sequence $\mathscr{S}''$. The rotated poses of $\mathscr{S}''$ are formally defined as:

$$\mathscr{P}''(t) = \begin{bmatrix} \cos(-\theta) & -\sin(-\theta) & 0 \\ \sin(-\theta) & \cos(-\theta) & 0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \mathscr{P}'(t) \quad \forall t \in [1, T_1 + T_2] \tag{6}$$

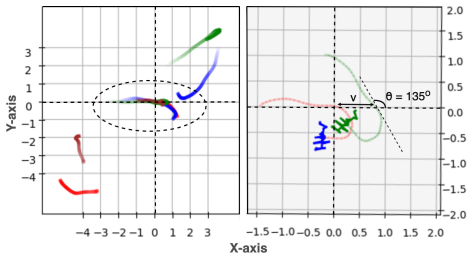


Figure 31: Left: Top-down view of 4 color-coded motion sequences, with corresponding sequences after motion transformation encircled. Faded colors mark the motion start. **Right:** Transformation of a single sequence. Original motion is represented by a green trajectory and skeleton, accompanied by calculations of the angle between the motion direction and the positive x-axis, as well as the translation vector at T_1.

Spatial Graph Embedding: We utilize a Graph Attention Network (GAT) [59] to generate graph embedding for each pose $\mathscr{P}''(t)$ in the input sequence. To that end, we represent the human pose as a graph, where each joint corresponds to a node and bones to the edges. Input to the spatial graph attention module is a reshaped sequence, $\mathscr{S}''_{\text{in}} \in \mathbb{R}^{T_1 \times N \times 3}$. Each joint node $j_i(t) \in \mathbb{R}^3$ has 3 spatial features in our graph representation. Edges, E are determined based on the kinematic chain of the body skeleton, these can vary depending on the dataset.

The GAT layer produces a new set of node features $\mathscr{X} \in \mathbb{R}^{T_1 \times N \times J_{\text{dim}}}$ as its output, yielding the joint embedding. Subsequently, joint embedding from the same pose are flattened to create the pose embedding, which has a dimension of $\mathscr{X} \in \mathbb{R}^{T_1 \times (N \times J_{\text{dim}})}$. Thus, the dimensionality D of the transformer model is given by $N \times J_{\text{dim}}$. The GAT here is employed to facilitate intra-frame attention mechanisms among joints, effectively capturing the spatial relationships. The output from the GAT, which represents spatial embedding, is subsequently fed into a transformer module. The transformer is designed to learn temporal relations across frames, ensuring a comprehensive understanding of both spatial and temporal dynamics of human motion.

Spatial-Temporal Positioning: We incorporate dual layers of positional encoding to capture human dynamics in detail, building on the formulation from [60]. First we generate a sinusoidal spatial positional encoding to establish differentiation amongst various joints within each pose. For every joint, this method produces a positional encoding of dimension J_{dim}, accounting for all N joints. Subsequently, to differentiate between poses over time, we generate a temporal encoding. Temporal encodings have dimension $J_{\text{dim}} \times N$, accounting for all T_1 poses, elucidating the sequential dynamics from one frame to the next.

Transformer Encoder Decoder: The basic structure of the Transformer layers are adopted from [60] with a non-autoregressive decoder inspired by POTR [55]. The transformer takes the spatio-temporally positioned input poses and processes it through a number of Nx encoder and decoder layers. Each layer of encoder and decoder shown in Figure 30 incorporates a Temporal Self-Attention component adopted from [61] that emphasize the relative distances between tokens in a sequence. In this technique, for each pose, attention scores are weighted more heavily towards its immediate neighboring poses. This is particularly beneficial for human pose sequences, where not only the order of

positions are crucial, but the relative transition between frames is also extremely important. In addition, to ensure that the human pose at current step is dependent only on prior poses and not on any future poses, we employ casual masking to Temporal Self-Attention components.

The output from encoder block projects the encoded sequence into a latent space $Z = [z_1, z_2, \ldots, z_T]$. The decoder utilizes this latent space to produce the output pose sequence. Decoder queries are initialised with $\mathscr{X}(T_1)$ which is the encoder's last input pose repeated over target length times.

After the decoding phase, a Multi-Headed Shared-Attention Mechanism is employed, wherein the query from the output of the decoder attends to the output of the Graph Attention Network, $\mathscr{X}$. Subsequently, these output embedding are propagated through linear layers and then transformed back to their original motion orientation and global coordinate space using $\vec{v}$ and θ giving rise to the forecasted 3D poses, $\mathscr{S}_{\text{out}}$.

Given that each pose data has a dimensionality of $3N$, with a predicted pose sequence $\hat{y}_{T_1+1}, \hat{y}_{T_1+2}, \ldots, \hat{y}_{T_1+T_2}$ and a ground truth pose sequence $y_{T_1+1}, y_{T_1+2}, \ldots, y_{T_1+T_2}$, our model was trained using a combined pose and trajectory loss function L:

$$L = \frac{1}{3N(T_2 - T_1 - 1)} \sum_{t=T_1+1}^{T_1+T_2} \|\hat{y}_t - y_t\|^2 + \frac{1}{3N(T_2 - T_1)} \sum_{t=T_1+1}^{T_1+T_2-1} \|(\hat{y}_{t+1} - \hat{y}_t) - (y_{t+1} - y_t)\|^2$$

6.3 Experiments

6.3.1 Setup

Datasets: We evaluate our approach on the common public datasets, Human3.6M [57] and CMU-Mocap [58], and on our own novel DARKO dataset. We categorize the motion sequences from those datasets into two broad action categories: navigational motion, such as walking, running or greeting, and static activities without distinct locomotion, such as discussing, smoking, or eating. The details of the Human3.6M and CMU-Mocap dataset processing are available in [10]. The DARKO walking humans dataset includes diverse locomotion modes (walking, pacing, running, turning), captured over a span of 3 hours from 17 participants, separated into 508 motion sequences, summing up to 35k frames across all actors.

We use Average and Final Displacement Errors (ADE and FDE), measured on the root joint (e.g. ADE_{Tr}) and the pose joints (e.g. ADE_{Po}) in meters. ADE is the mean square error between predicted and ground truth joint coordinate, calculated over the entire output sequence, whereas FDE measures displacement in the last time step. When calculating ADE_{Po} and FDE_{Po} to isolate pose prediction accuracy, we subtract the root joint to remove global translation. Furthermore, we report Runtime (R) in milliseconds to illustrate duration of a single forward pass.

6.3.2 Results

Human3.6M [57]: Table 7 presents a quantitative comparison of our method, UPTor, against four established baselines [45, 54, 51]. The baseline values are taken from [45], and we make sure to strictly follow their evaluation setup.

Experimental results in Table 7 demonstrate the competitive performance of our approach, UPTor, against the state of the art in pose prediction (Dlow and DMMGAN). DLow is not designed to predict trajectories and cannot handle global translation, whereas DMMGAN is not a feasible model for real-time prediction and has restricted use for robotic applications due to the large model size and higher runtime. The major part of the H3.6M evaluation set is comprised of static actions such as sitting, smoking, eating, or discussing.

Method	ADE/FDE$_{Po}$ (m)↓	ADE/FDE$_{Tr}$ (m)↓	R (msec) ↓
DLow [54]	0.48 / 0.62	0.19 / 0.45	**20**
DMMGAN [51]	**0.44** / **0.52**	**0.12** / **0.23**	100
HipOnly [51]	-	0.15 / 0.30	18
STPOTR [45]	0.50 / 0.75	0.13 / 0.27	25
UPTor	0.55 / 0.76	**0.12** / 0.25	-

Table 7: Evaluation on Human3.6M across 15 distinct actions, including both navigational and static activities.

Method	Params ↓	R (msec) ↓
STPOTR [45]	43,276,992	73
UPTor	**23,165,184**	**61**

Table 8: Runtime and size comparison on H3.6M

Method	DARKO Dataset		CMU-Mocap	
	ADE/FDE$_{Po}$(m)↓	ADE/FDE$_{Tr}$(m)↓	ADE/FDE$_{Po}$(m)↓	ADE/FDE$_{Tr}$(m)↓
STPOTR [45]	0.47 / 0.65	0.17 / 0.34	0.58 / 0.86	0.09 / 0.18
UPTor	**0.39** / **0.55**	**0.13** / **0.26**	**0.45** / **0.71**	**0.07** / **0.14**

Table 9: Error metrics comparison on DARKO data & locomotion actions subset of CMU-Mocap

UPTor model variant	Original		Translate		Rotate		Trans + Rot	
	ADE$_{Po}$ (m) ↓	ADE$_{Tr}$ (m) ↓	ADE$_{Po}$ (m) ↓	ADE$_{Tr}$ (m) ↓	ADE$_{Po}$ (m) ↓	ADE$_{Tr}$ (m) ↓	ADE$_{Po}$ (m) ↓	ADE$_{Tr}$ (m) ↓
w/o Transformation	0.46	**0.13**	1.12	5.51	1.08	2.47	1.24	6.40
with Transformation	**0.39**	**0.13**	**0.39**	**0.13**	**0.39**	**0.13**	**0.39**	**0.13**

Table 10: Evaluation on DARKO data under diverse spatial transformations and motion orientations.

Given that these actions are heavily influenced by individual behavioral patterns, the superior performance of these two models is due to their generative nature, which allows them to forecast a wide range of potential future poses/trajectories, and the error metrics are computed on the closest generated sample. Our UPTor surpasses the HipOnly method which is the trajectory prediction module from DMMGAN.

UPTor also excels in trajectory prediction across all 15 actions of the H3.6M dataset surpassing STPOTR in terms of ADE$_{Tr}$ and FDE$_{Tr}$. STPOTR uses a distinct module and loss function for pose prediction, targeting on minor variations in pose dynamics, thereby improving its pose prediction accuracy. Instead, we utilize a unified architecture and loss function for parallel training and prediction of both pose and trajectory, thus trajectories are predicted with high precision as they account for a significant portion of training error. Moreover, our model uses pose dynamics as an added context for effective trajectory learning.

The modular approach of STPOTR comes at a computational cost, as highlighted by the runtime comparison presented in Table 8. For an assessment of computational efficiency and model complexity, we conducted run-time tests with our model and STPOTR on a laptop equipped with an 11th Gen Intel® Core™ i7-11850H CPU and an NVIDIA RTX A3000 GPU. In Table 8, “Params" refers to the number of transformer parameters, shedding light on model size and complexity, while run-time represents the duration for a single forward pass with batch size of 1. Empirical results demonstrate our model’s capability to infer human pose dynamics and trajectory patterns with a 50% reduction in model size and enhanced run-time performance. These results align with our objective for achieving accurate trajectory prediction with better computational efficiency compared to baseline models for mobile robotic applications. Similarly, quantitative results in Table 9 indicate superior performance of our model in predicting both pose and trajectory for navigational actions from the CMU-Mocap dataset.

Finally, Fig. 32 illustrates exemplary predicted poses along with the trajectory displacement.

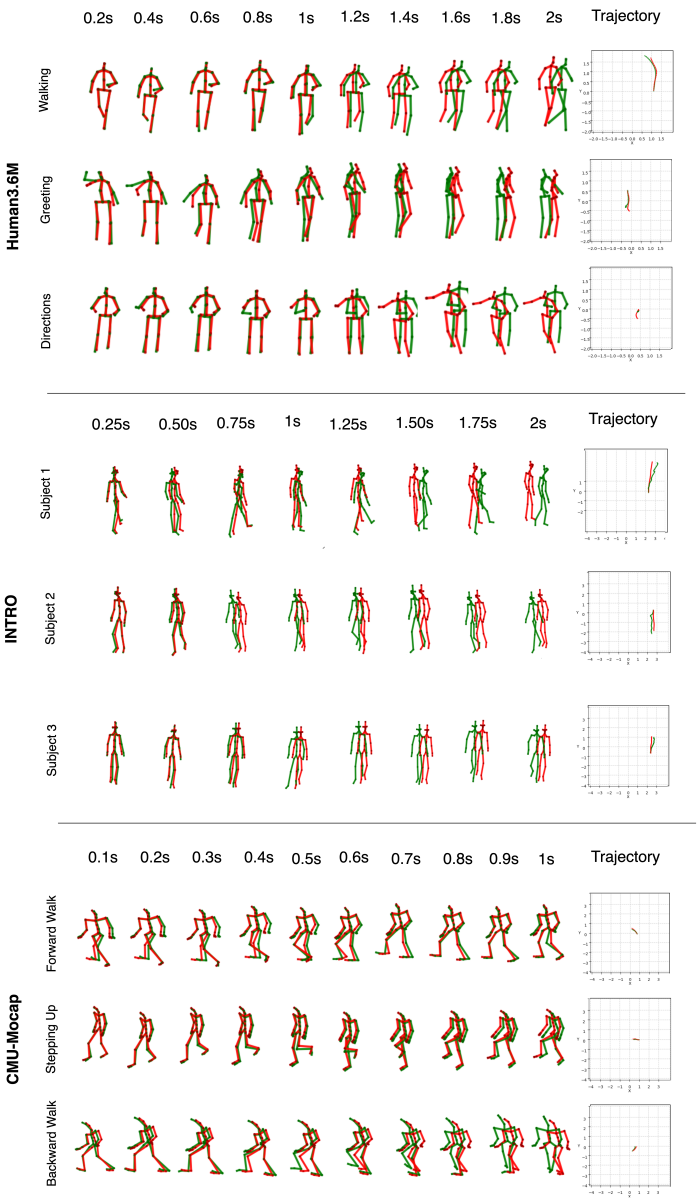


Figure 32: Example predictions in all datasets. Predicted poses are shown with red skeletons, while the green ones depict the ground truth, followed by the top-down view of the entire trajectory in global space. For H3.6M and DARKO, motion is predicted for 2 sec and for 1 sec for CMU-Mocap.

7 Human Gaze and Head Rotation during Navigation, Exploration and Object Manipulation

Summary: The human gaze is an important cue to signal intention, attention, distraction, and the regions of interest in the immediate surroundings. Gaze tracking can transform how robots perceive, understand, and react to people, enabling new modes of robot control, interaction, and collaboration. In this section, we use gaze tracking data from THÖR-MAGNI to investigate the coordination between gaze direction and head rotation of humans engaged in various indoor activities involving navigation, interaction with objects, and collaboration with a mobile robot. In particular, we study the spread and central bias of fixations in diverse activities and examine the correlation between gaze direction and head rotation. We introduce various human motion metrics to enhance the understanding of gaze behavior in dynamic interactions. Finally, we apply semantic object labeling to decompose the gaze distribution into activity-relevant regions. This work is accepted to the RO-MAN 2024 conference [7].

7.1 Introduction

Gaze has been described as a window into the human mind. It provides information related to human attention and intention. Integrating gaze tracking into Human-Robot Interaction (HRI) approaches can help robots better understand human behavior and, in turn, help robots navigate shared spaces more effectively and participate in collaborative tasks with greater awareness and adaptability.

Studies of human gaze during navigation and dynamic human-robot interactions are still scarce, not least due to the complexity of tracking the gaze of a moving person. Thus, head orientation is often used as a proxy for gaze direction and using head orientation has been shown to improve the interaction between humans and robots [62]. Furthermore, head orientation is successfully used in automated driving settings to infer the attention and intention of pedestrians and cyclists [63]. However, relying solely on head orientation as an indicator of gaze has limitations due to the complex nature of human attention, which often involves subtle eye movements not captured by head orientation alone [64], (see Figure 33).

In the course of the DARKO project, we analyze the gaze patterns of people moving and interacting in a dynamic environment shared with robots. We utilize the THÖR-MAGNI dataset, unique for its synchronized data on head orientation, eye movement patterns, and walking trajectories across a diverse group of individuals [9]. In particular, we show the potential and limitations of using head orientation as a proxy for gaze and the complex relationship between head movements and gaze direction.

Our study employs various analytical approaches to examine and describe human gaze patterns. Firstly, we focus on the distribution of visual fixations on the 2D tracker plane to evaluate the uncertainty caused by eye rotation relative to head orientation. We extend the analysis of fixations by examining participants' activities and the specific micro-actions they performed during tasks and interactions. We use heatmaps to visualize fixations and identify patterns of visual engagement and attention allocation.

To offer a geometric representation of where participants fixated most frequently on these heatmaps in the 2D tracker plane, we apply ellipse-fitting techniques to summarize and analyze areas of highest fixation density, referred to as "central tendencies". Additionally, the levels of engagement are quantified by calculating the average duration and rate of fixations. This allows for a deeper understanding of how participants interacted with their environment and the robots within it. Through this analysis, we aim to provide

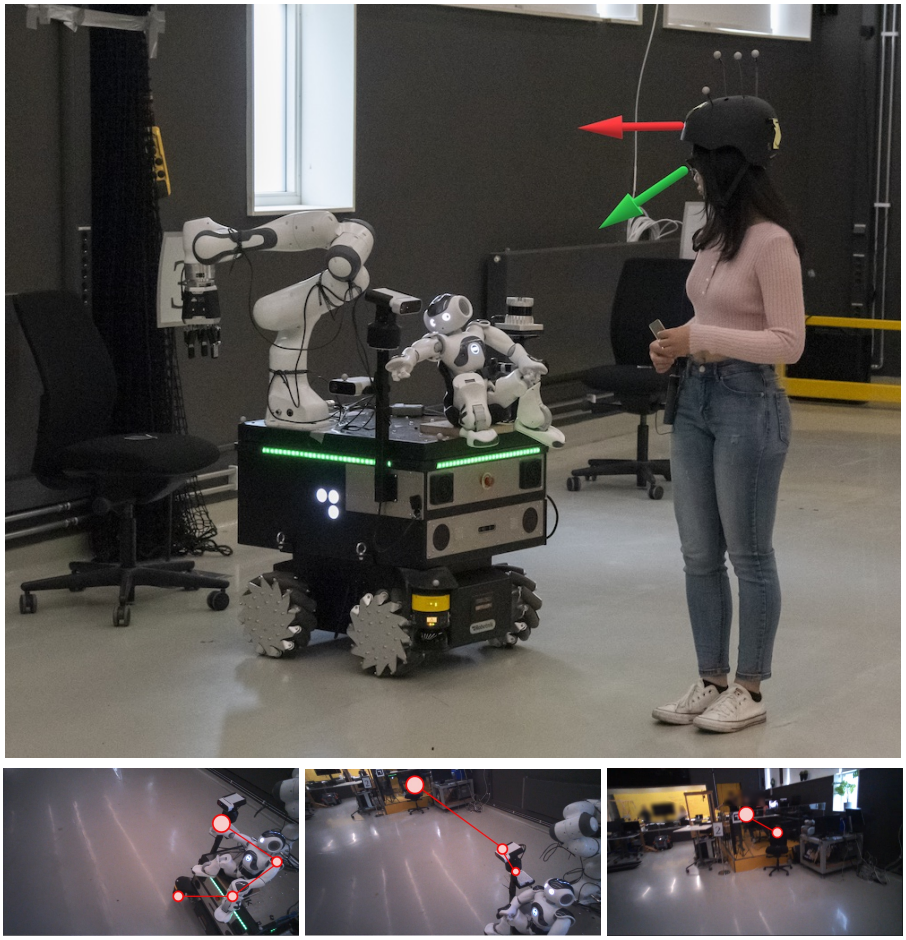


Figure 33: A participant of the THÖR-MAGNI dataset attends to instructions of the mobile robot [65]. **Top:** Illustration of the visual difference between the head orientation (**red**) and gaze direction (**green**). **Bottom:** a sequence of gazes on the mobile robot, followed by a shift of attention to the goal point that the robot cued. This shift is followed by a head rotation to center the visual field on the goal point. Fixations are shown with **white circles**, and their sequences are connected by **red lines**.

more effective support for gaze-informed predictions in dynamic settings and highlight the nuanced ways human attention is directed and sustained during human-robot interaction.

Furthermore, we investigate the coordination between eye and head movements during attention shifts. We compare our findings in the indoor settings with prior studies in outdoor environments. We correlate head orientation and gaze vectors with motion metrics to link visual attention with physical movement. With this analysis, we seek to support the deployment of appearance-based gaze estimation methods, which struggle with head and eye coordination variability [66], especially in dynamic environments.

Lastly, we leverage the YOLO object detection model to qualify the objects human gaze at more precisely. By identifying and categorizing objects or areas that attract significant visual focus, we gain insights into the semantics of targets of participants' gaze, enriching our understanding of attention allocation in dynamic settings, especially during locomotion. Applying modern computer vision techniques to eye-tracking data is a promising approach to contextually interpreting human attention within the context of HRI.

Table 11: Amount of eye-tracking data available from the Tobii glasses for various activities in the THÖR-MAGNI dataset.

Activity	Recorded minutes	Scenario
Visitors-Alone	108	All
Visitors-Group 2	124	All
Visitors-Group 3	52	All
Carrier-Bucket	32	2–3
Carrier-Box	60	2–3
Carrier-Large Object	92	2–3
Total	468	

7.2 Analysis of Gaze Patterns in Navigation and Interaction Tasks

We develop and describe a methodology to analyze human gaze in dynamic environments. We study and compare gaze behavior across various activities and tasks, which include search and navigation towards goals in the room, manipulation of objects, social interactions and receiving instructions from the robot. We have two goals in mind: first, to provide tools and methods to support human activity understanding and prediction from mobile gaze trackers, and second, to quantify human gaze in relation to head orientation in scenarios where systems may need to rely on the head orientation as a proxy for head direction.

In this section we present the analysis of gaze distribution and its bias in Sec. 7.2.1, 7.2.2 and 7.2.3. In Sec. 7.2.4, we introduce head orientation into the analysis and discuss the head- and eye-rotation comfort ranges. In Sec. 7.2.5, we discuss auxiliary motion metrics, and, in Sec. 7.2.6, examine the distribution of gazes towards static and dynamic semantic objects in the environment.

7.2.1 Overall Gaze Distribution

First, we study gaze points where participants focused their attention, known as fixations [67]. An example of fixation sequences obtained with eye-tracking glasses is shown in Figure 33 (bottom). We visualize the accumulation of fixations using heatmaps, also called "attention maps". These heatmaps, generated from the THÖR-MAGNI dataset's eye-tracking data, illuminate how participants distribute visual attention within environments shared with robots and other humans.

Figure 34 displays the heatmap created for the entire gaze fixations. We identify a preference for gaze points along the vertical center of the visual field, accompanied by a stronger variation in vertical fixation positions compared to horizontal ones. The participants' focus in our study is slightly shifted to the right of the vertical center line. Additionally, there was a general trend of participants directing their gaze more toward the upper portion of the images.

Using heatmaps, we can also quantify these shifts of the gaze distribution. To that end, we use an ellipse fitting technique [68], which leverages the covariance matrix, eigenvalues, and chi-squared distributions to accurately delineate areas of concentrated gaze. This method, also visualized in Figure 34, encapsulates areas representing 25%, 50%, 80%, and 90% of all collected fixations, providing a quantitative measure of where participants' gazes converge most frequently. Our findings reveal that ellipses covering 28% and 39% of the image area encapsulate 80% and 90% of all fixations, respectively, underscoring participants' central focus in the visual field.

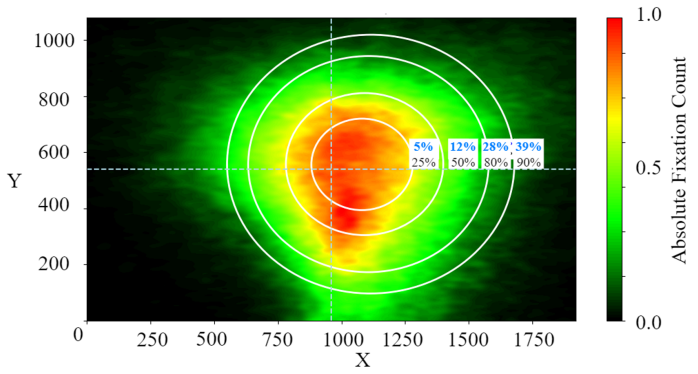


Figure 34: Fixation locations in the THÖR-MAGNI dataset. **Ellipses** represent areas containing 25%, 50%, 80%, and 90% (**Black Labels**) of all recorded gazes. **Blue Labels** indicate the percentage of the 1920x1080 image included in each ellipse.

7.2.2 Gaze Distribution Across Activities and Micro-Actions

We examine how gaze distribution varies across participant activities and micro-actions to understand the intricacies of the human gaze in motion and around robots within the THÖR-MAGNI dataset's dynamic settings. We analyze activities such as navigating to goal points and manipulating objects. For consistency reasons, our analysis focuses on the initial three scenarios involving the roles of "Visitors" and "Carriers". Therefore, we do not include Scenarios 4 and 5, with a stronger focus on HRI, as there is already a study concerning the human gaze in these scenarios [65], which is part of WP5 T5.2 and deliverable D5.2. Figure 35 illustrates the spatial distribution of fixations, revealing nuanced visual attention patterns across roles and micro-actions, supported by Tables 12 and 13 contain numerical values describing the center of mass and spread of the central fixation tendencies. "Carrier-Box" and "Carrier-Bucket" participants exhibit concentrated gazes toward the center during object transportation, indicating focused attention necessary for this task. In contrast, the "Carrier-Large Object" group shows a more dispersed gaze pattern, particularly favoring the upper hemisphere. This dispersion likely reflects the need for broader environmental awareness in localizing the object and maneuvering oversized items, partially occluding their visual field. "Visitors-Group 2" and "Visitors-Alone" display varied gaze distributions, highlighting the impact of group size and task complexity. Notably, "Visitors-Group 3" participants prefer the lower hemisphere, possibly indicating a different visual engagement strategy due to group dynamics or task demands.

7.2.3 Quantifying Gaze Distribution: Central Bias and Spread

This study examines the distribution of visual fixations during various participant activities, including micro-actions such as "Walk between goals" or "Draw Card" and tasks involving object manipulation. The findings are presented in Tables 12 and 13 as tuples indicating the 2D coordinates of the fixations' center of mass relative to the image's central point (illustrated by the crossing of the dotted lines in Figure 35). Following the methods described in subsection 7.2.1, we fit ellipses to encompass 80% of the fixation distributions. We list the percentages that detail the proportion of the 2D eye-tracking plane these occupy alongside the coordinates in the tables. This analysis sheds light on visual attention patterns across different tasks, marking the shifts in focus with precise distances from the image's center. For the "Carrier-Box" and "Carrier-Bucket" groups, the analysis revealed that the central displacements of their fixation hotspots were generally close to the image's

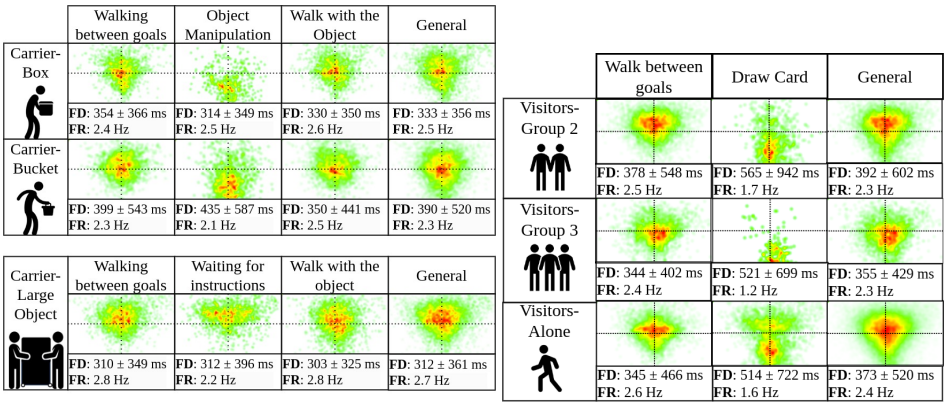


Figure 35: Heatmaps of fixation locations in the dataset with average **Fixation duration (FD)** and overall **Fixation rate (FR)** per role and the comprised micro-actions. Visual axes are shown with dotted lines. Central coordinates and the spread of the central biases are listed in Tables 12 and 13.

Table 12: Central shift of gaze distribution in activities and micro-actions from Figure 35 (top). In each cell, the tuple indicates the 2D coordinates of the distribution center of mass with respect to the center of the frame (rounded to the nearest 5 pixels). The percentage indicates the area of the ellipse (that encompasses 80% of the fixations) with respect to the area of the frame.

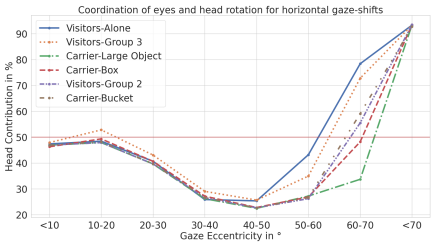
Activity	**Walk between goals**	**Object Manipulation**	**Walk with the Object**	**General**
Carrier-Box	(-20, 0), 18.6%	(-40, -240), 10.6%	(-100, 50), 15.9%	(0, 20), 27.3%
Carrier-Bucket	(0, 150), 28.5%	(60, -260), 14.8%	(-80, 30), 18.2%	(-10, -40), 19.1%
	Walk between goals	**Wait for instructions**	**Walk with the Object**	**General**
Carrier-Large Object	(0, 75), 19%	(-35, 150), 22.7%	(50, 50), 21.2%	(0, 125), 25%

Table 13: Central shift of gaze distribution in activities and micro-actions from Figure 35 (bottom). In each cell, the tuple indicates the 2D coordinates of the distribution center of mass with respect to the center of the frame (rounded to the nearest 5 pixels). The percentage indicates the area of the ellipse (that encompasses 80% of the fixations) with respect to the area of the frame.

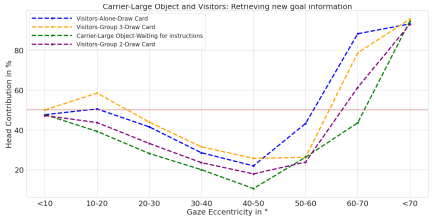
Activity	**Walk between goals**	**Draw Card**	**General**
Visitors-Group 2	(100, 50), 22.8%	(-30, -270), 14.7%	(10, 250), 24.9%
Visitors-Group 3	(80, -80), 22.7%	(80, -420), 8.5%	(80, -100), 17%
Visitors-Alone	(80, -100), 15.4%	(30, -250), 34.1%	(50, 50), 24.2%

center, indicating a concentrated area of visual attention during most activities, except during “Object Manipulation,” where the focal areas significantly diverged from the center. This pattern suggests that tasks requiring detailed object interaction prompt broader visual engagement, as evidenced by the larger displacement values. Conversely, the “Carrier-Large Object” group’s fixations were predominantly in the image’s upper hemisphere, indicating a consistent focus area across their activities.

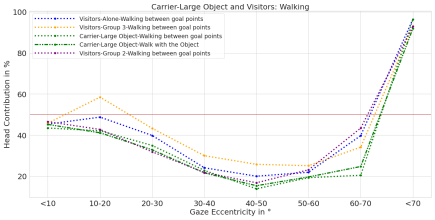
During the analysis of the visitors’ activities, we observe discernible patterns in the distribution of visual attention across different micro-actions. Specifically, during the “Draw Card” tasks, there was a noticeable shift of focus toward the lower hemisphere of



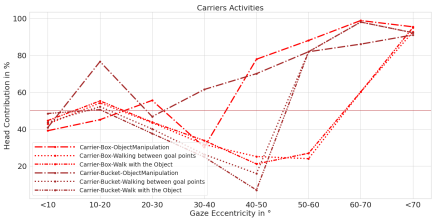
(a) Gaze eccentricity: During motion peaks with eye and head rotations in the same direction, binned in 10-degree intervals. **Overall Head Contribution** to these shifts is shown on the other axis.



(b) Goal actions: Visitors and Carriers-Large Object **drawing a card** or **waiting for instructions** for new goal points.



(c) Walking actions: Visitors and Carriers-Large Object.



(d) Object manipulation and walking: Carrier-Bucket and Carrier-Box, walking actions include with and without object.

Figure 36: Showing how much the head direction contributes to the total gaze direction, depending on the eccentricity of the gaze angle with respect to the body orientation. This figure applies to movements where the head and eyes move horizontally in the same direction.

the image, indicating a heightened level of visual engagement unique to this activity. This gaze concentration is distinct from the more varied attention patterns associated with other tasks, indicating that specific actions can significantly influence where and how visual attention is assigned. In tasks other than "Draw Card," the "Visitors-Alone" and "Visitors-Group 2" categories exhibited behaviors consistent with active visual exploration and navigation within the space. Conversely, "Visitors-Group 3" primarily focused their gaze on the center of the image, indicating a desire to facilitate communication and coordination within the group. These observations highlight the dynamic and task-specific nature of visual attention among the "Visitors", emphasizing how the context and demands of different activities subtly shape the collective and individual focus within groups.

7.2.4 Quantifying Eye-Head Coordination in Gaze Shifts

To better understand eye-head coordination, we analyze instances where participants' eyes and heads rotated horizontally in unison. We focus on horizontal rotations because observers preferentially spread their gaze horizontally to explore their surroundings, and horizontal gaze movements are more common than vertical ones when walking over flat terrain [69].

Our systematic categorization of eye and head movement coordination across various activities revealed distinct patterns in horizontal gaze shifts. For minor attention shifts, contributions from both eyes and head were nearly equal, especially during locomotion (see Figure 36c). For larger attention shifts, the relative contribution of the head decreases, reaching a minimum of around 45 degrees. This is slightly higher than the 35 degrees observed by Stahl et al. [70] for eye movements in seated free viewing tasks. However,

Table 14: Walking Speed, Acceleration, and Straightness Index (SI) for the roles in Scenarios 1–3 of the THÖR-MAGNI dataset

Role	Speed [m/s]	Acceleration [m/s²]	SI
Visitors-Alone	0.88 ± 0.55	0.28 ± 0.37	0.6
Visitors-Group 2	0.81 ± 0.5	0.24 ± 0.31	0.69
Visitors-Group 3	0.80 ± 0.5	0.24 ± 0.36	0.77
Carrier-Box	1.07 ± 0.47	0.25 ± 0.38	0.95
Carrier-Bucket	1.16 ± 0.4	0.27 ± 0.41	0.97
Carrier-Large Object	0.65 ± 0.51	0.22 ± 0.27	0.75

head movements were more dominant across all gaze shifts and micro-actions in groups of three visitors (see Figures 36a and 36b). Beyond the 40-50 degree range, head movement contributions increased variably across different activities, with Visitors-Alone and those in Visitors-Group 3 adjusting their gaze more rapidly than those carrying objects. Eye movements dominated between 20 and 50 degrees, while more significant shifts above 70 degrees significantly increased head contributions. This pattern aligns with outdoor environment findings [69] but with a more pronounced reliance on eye movements in indoor settings. An exception to these trends is the micro-actions of Carrier-Box and Bucket during object manipulation, which strongly preferred head contributions to gaze shifts (see Figure 36d).

7.2.5 Correlating Motion with Gaze Alignment

To explore how human motion dynamics impact eye-head coordination, we propose several metrics of human motion, namely the straightness of participants' trajectories, walking speed, and acceleration. The straightness index (SI) ranges from 0 (non-linear) to 1 (perfectly straight), with higher values indicating more linear trajectories and lower values indicating more explorative trajectories. Mean walking speeds and accelerations are presented with standard deviations to highlight inter-individual variability. To calculate these motion metrics, we follow the preprocessing methods outlined by de Almeida et al. [5]. Results are outlined in Table 14.

Our analysis extends to correlating (Spearman correlation) these motion metrics with the alignment between head orientation and gaze vector, offering novel insights into human navigation strategies. We discovered a subtle yet statistically significant negative correlation between head and eye rotation alignment with walking speed ($\rho = -0.04$, $p < 0.01$), suggesting that increased linear velocity tends to enhance the synchronization of eye and head movements. A weaker yet also significant negative correlation with linear acceleration ($\rho = -0.01$, $p < 0.01$) indicates a less pronounced impact on the alignment of eyes and head compared to velocity.

The motion metrics in Table 14 vary across participant activities and reveal distinct movement patterns. "Carriers", characterized by high-velocity, linear movements and minimal head movement contribution to gaze shifts and centralized gazes, contrast with "Visitors-Alone," who exhibit more dynamic movements with less linear trajectories and more explorative gaze distributions. "Visitors-Group 2" and "Visitors-Group 3" show similar speeds and accelerations to "Visitors-Alone" but follow straighter paths and demonstrate different gaze and head movement dynamics. "Visitors-Group 3" mainly displays a more considerable head contribution to gaze shifts.

Moving in pairs, the "Carrier-Large Object" participants displayed the slowest walking speeds and accelerations, a wide range of velocities, and the highest fixation rates, under-

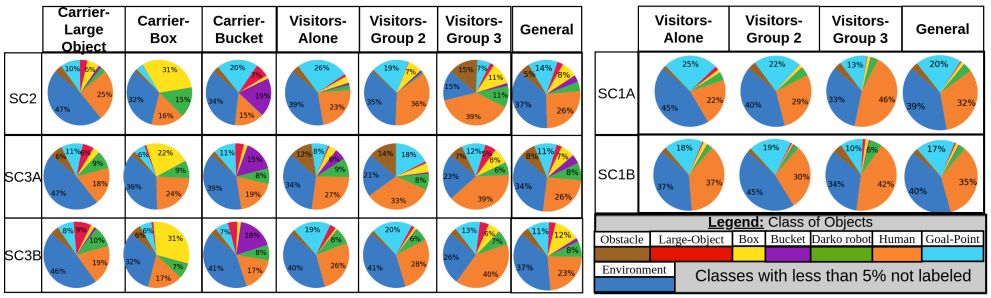


Figure 37: Distribution of fixations on objects by participants in Scenarios 2, 3A, and 3B (left) and 1A and 1B (right).

lining a unique interaction pattern. Their contribution of head movement to gaze shifts notably stayed below 50% for most activities despite their interactions with objects like the other carriers (see Figure 36), emphasizing the role's distinct coordination patterns. These findings emphasize the intricate relationship between physical motion and gaze behavior, contributing valuable perspectives to developing intuitive and responsive robot interactions in shared spaces.

7.2.6 Quantifying Attention with Object Detection

The static and dynamic objects heavily influence the human gaze in the environment. We employed an object detection method for the video frames from the eye-tracking glasses to achieve a finer decomposition of attention into classes of semantic objects. We used YOLOv8 [71], pre-trained on the COCO dataset, and refined with a custom dataset with labeled objects from THÖR-MAGNI. The classes, listed in Figure 37, include role-dependent objects (e.g. boxes and buckets), other walking people and the DARKO robot (see also Figure 33).

Our custom dataset, consisting of 355 images annotated with seven classes, facilitated a focused analysis of participants' gaze during motion, particularly near the DARKO robot. Through this methodology, we observed notable shifts in attention distribution across different scenarios and activities, as visualized in Figure 37. The pie charts illustrate a change in attention allocation from the environment and other participants in Scenario 1 to the more diversified attention towards the DARKO robot in subsequent scenarios, underscoring its significant presence in the shared space.

The statistical analysis, employing t-tests and calculating Cohen's d-effect sizes, supports these observations with significant findings. Specifically, the transition from Scenarios 1 and 2 to Scenarios 3A and 3B reveals a marked increase in attention towards DARKO, with effect sizes of $[-1.6, -0.8]$, all with $p < 0.1$, respectively, indicating a strong influence of the robot's presence on participant attention. This influence is further supported by the lack of statistically significant differences in attention between the different driving styles of DARKO in Scenarios 3A and 3B, suggesting that it is the robot's presence as a static or dynamic entity rather than its motion pattern that primarily captures human attention.

7.3 Discussion

Our analysis emphasizes the utility of using head orientation as a baseline assumption for gaze direction, which is particularly advantageous for onboard sensor-based gaze tracking. While this correlation is crucial for designing natural and seamless HRI systems, more than head orientation is required. It serves as an excellent standalone measure but must

be complemented with gaze information to account for more dynamic settings, enabling refined interaction models that accommodate the complexity of human attention in diverse activities.

We observed that the contribution of eye gaze to attentional shifts decreases when objects are being carried or manipulated, indicating a preference for head movements over eye movements in these scenarios. Eye gaze is centralized during these tasks and focuses on objects and goal points relevant to the task fulfillment. For actions such as card drawing and navigating between goal points, where eye gaze plays a more significant role, understanding central gaze tendencies becomes essential to support gaze estimation via head orientation. Specifically, during navigation and visual exploration, gazes are primarily directed toward the upper hemisphere of the mobile eye-tracker (horizon) and on the lower hemisphere during card drawing or object manipulation.

Moreover, our observations of participants in groups of three show that they divert their attention from robots to social interactions, displaying a slightly higher head contribution overall than other roles. This highlights the need for mobile robots to integrate sophisticated detection and anticipation algorithms in crowded areas, particularly around groups. Such capabilities are vital for navigating social environments, ensuring human safety, and optimizing robot operational efficiency. Understanding group dynamics provides valuable insights for designing robots that can navigate human social settings, adjust their behavior to minimize disruptions and promote coexistence.

Additionally, our research on human-robot interaction has revealed a significant finding regarding the perception of robots. Specifically, the gaze distribution was similar for robots driving directionally and omnidirectionally, suggesting that the perception of these two mobility styles may not differ substantially. This similarity in gaze distribution indicates a potential versatility in human acceptance of different robotic mobility styles, which opens up avenues for innovative robot designs without compromising the user experience. The affirmation of technological advancements in robot locomotion encourages confidence in their acceptance within human-centered environments.

In conclusion, examining human gaze behavior in HRI contexts enriches our understanding of the interplay between human attention, perception, and robot design. These insights can advance the development of robotic systems that align with human behaviors and expectations, improving safety, efficiency, and integration into shared spaces. The broad applicability of these advancements, from collaborative manufacturing to autonomous vehicles, highlights the importance of gaze analysis in future HRI research and development.

7.4 Conclusion

We strongly believe that, with the evident progress in mobile eye-tracking, human gaze will increase its importance in Human-Robot interaction, bi-directional communication, activity recognition and direct robot control. Along with the head orientation and full-body pose, as discussed in Section 6, gaze is an important visual cue for the robots to track and exploit in their operation. In this section we outlined a novel methodology to measure and generalize human gaze data in navigation and dynamic interaction scenarios.

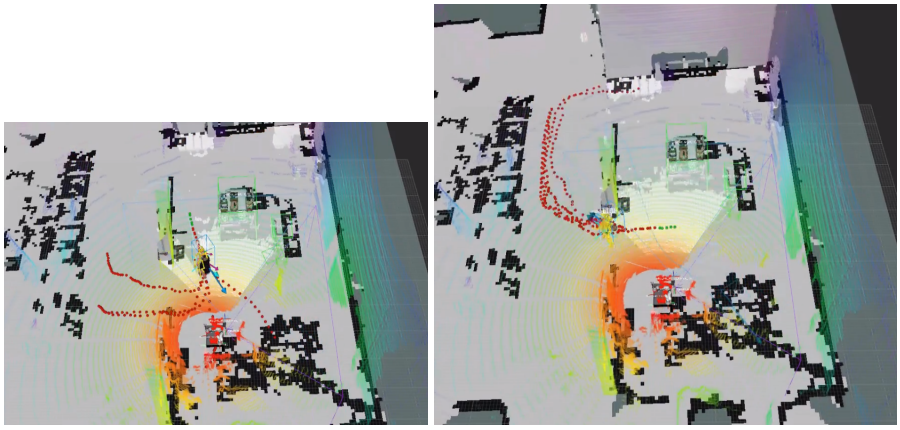


Figure 38: CLiFF-LHMP prediction of people at ARENA2036 running live on the DARKO robot [3].

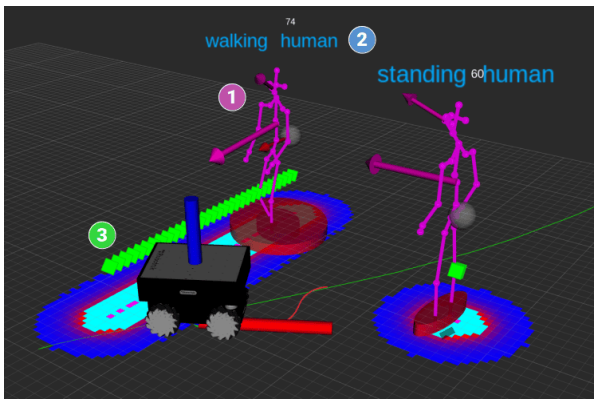


Figure 39: Context-aware collision avoidance of the DARKO robot, considering (1) full-body 3D human skeleton poses projected as red ellipses, (2) detected activities, and (3) 2D motion predictions. The mobile robot proactively clears the path of a walking human, while preparing to bypass a standing person. This method [53] is developed in WP6.

8 Conclusion

In this deliverable, the DARKO consortium has introduced novel methods, datasets, and experimental results to predict the future motions and intents of the surrounding people. These results, in combination with the work carried out in T5.2, T5.3 and T5.4 of WP5, represent major steps forward towards the research goals of DARKO and the final, fully integrated robot demonstration at the end of the project.

Several novel methods described in this deliverable have been successfully deployed and demonstrated during the MS3 demonstration milestone and the associated stakeholder meeting at the KI.FABRIK in Munich in June 2024. Specifically, we have shown the data efficient long-term trajectory prediction using maps of dynamics, trained from only 5 minutes of human motion data, see Fig. 38. We have also integrated the fast short-term trajectory prediction into the context-aware Model Predictive Control (MPC) method for collision avoidance, see Fig. 39. Further practical demonstrations occurred at the Automatica fair in Munich in June 2023 and the project status day event at ARENA 2036 in September 2023, and have raised great interest among the audiences.

As future steps, we plan to integrate and demonstrate the full-body prediction system,

described in Sec. 6, on the robot, and achieve tighter integration with the rest of the WP5 components. Towards the final demonstration in 2025, we wish to plan sets of experiments that demonstrate the benefit of using the proposed architecture in cluttered intralogistic environments.

9 References

WP5 publications

[1] Andrey Rudenko, Luigi Palmieri, Wanting Huang, Achim J Lilienthal, and Kai O Arras. "The atlas benchmark: An automated evaluation framework for human motion prediction". In: *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE. 2022, pp. 636–643.

[2] Yufei Zhu, Andrey Rudenko, Tomasz P Kucner, Achim J Lilienthal, and Martin Magnusson. "A Data-Efficient Approach for Long-Term Human Motion Prediction Using Maps of Dynamics". In: *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA), Workshop on Long-term Human Motion Prediction*. 2023.

[3] Yufei Zhu, Andrey Rudenko, Tomasz P Kucner, Luigi Palmieri, Kai O Arras, Achim J Lilienthal, and Martin Magnusson. "CLiFF-LHMP: Using Spatial Dynamics Patterns for Long-Term Human Motion Prediction". In: *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2023, pp. 3795–3802.

[4] Yufei Zhu, Han Fan, Andrey Rudenko, Martin Magnusson, Erik Schaffernicht, and Achim J Lilienthal. "LaCE-LHMP: Airflow Modelling-Inspired Long-Term Human Motion Prediction By Enhancing Laminar Characteristics in Human Flow". In: *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*. 2024.

[5] Tiago Rodrigues De Almeida, Andrey Rudenko, Tim Schreiter, Yufei Zhu, Eduardo Gutierrez Maestro, Lucas Morillo-Mendez, Tomasz P Kucner, Oscar Martinez Mozos, Martin Magnusson, Luigi Palmieri, et al. "THOR-Magni: Comparative Analysis of Deep Learning Models for Role-Conditioned Human Motion Prediction". In: *Workshop Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 2200–2209.

[6] Tiago Rodrigues de Almeida, Yufei Zhu, Andrey Rudenko, Tomasz P Kucner, Johannes A Stork, Martin Magnusson, and Achim J Lilienthal. "Trajectory Prediction for Heterogeneous Agents: A Performance Analysis on Small and Imbalanced Datasets". In: *IEEE Robotics and Automation Letters* (2024).

[7] Tim Schreiter, Andrey Rudenko, Martin Magnusson, and Achim J Lilienthal. "Human Gaze and Head Rotation during Navigation, Exploration and Object Manipulation in Shared Environments with Robots". In: *Proc. of the IEEE Int. Symp. on Robot and Human Interactive Comm. (RO-MAN)*. 2024.

[8] Tim Schreiter, Tiago Rodrigues de Almeida, Yufei Zhu, Eduardo Gutiérrez Maestro, Lucas Morillo-Mendez, Andrey Rudenko, Tomasz P Kucner, Oscar Martinez Mozos, Martin Magnusson, Luigi Palmieri, et al. "The Magni Human Motion Dataset: Accurate, Complex, Multi-Modal, Natural, Semantically-Rich and Contextualized". In: *IEEE International Conference on Robot and Human Interactive Communication Workshop Proceedings: Towards Socially Intelligent Robots In Real World Applications (SIRRW 2022)*. IEEE. 2023.

[9] Tim Schreiter, Tiago Rodrigues de Almeida, Yufei Zhu, Eduardo Gutierrez Maestro, Lucas Morillo-Mendez, Andrey Rudenko, Luigi Palmieri, Tomasz Piotr Kucner, Martin Magnusson, and Achim J. Lilienthal. "THÖR-MAGNI: A Large-scale Indoor Motion Capture Recording of Human Movement and Robot Interaction". In: *Submitted to IJRR, arXiv preprint arXiv:2403.09285* (2024).

[10] Nisarga Nilavadi Chandregowda, Andrey Rudenko, and Timm Linder. "Unified 3D Human Pose Dynamics and Trajectory Prediction for Robotic Applications". In: *Submitted to the Conference on Robotic Learning (CoRL)*. 2024.

Other references

[11] A. Rudenko, L. Palmieri, M. Herman, K. M. Kitani, D. M. Gavrila, and K. O. Arras. "Human motion trajectory prediction: A survey". In: *Int. J. of Robotics Research* 39.8 (2020), pp. 895–935.

[12] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese. "Social LSTM: Human trajectory prediction in crowded spaces". In: *Proc. of the IEEE Conf. on Comp. Vis. and Pat. Rec. (CVPR)*. 2016, pp. 961–971.

[13] Christoph Schöller, Vincent Aravantinos, Florian Lay, and Alois Knoll. "What the constant velocity model can teach us about pedestrian motion prediction". In: *IEEE Robotics and Automation Letters* 5.2 (2020), pp. 1696–1703.

[14] Stefan Becker, Ronny Hug, Wolfgang Hubner, and Michael Arens. "Red: A simple but effective baseline predictor for the trajnet benchmark". In: *Proc. of the Europ. Conf. on Comp. Vision (ECCV) Workshops*. 2018.

[15] Parth Kothari, Sven Kreiss, and Alexandre Alahi. "Human trajectory forecasting in crowds: A deep learning perspective". In: *IEEE Trans. on Intell. Transp. Syst. (TITS)* (2021).

[16] Marius Lindauer, Katharina Eggensperger, Matthias Feurer, Stefan Falkner, André Biedenkapp, and Frank Hutter. *SMAC v3: Algorithm Configuration in Python*. https://github.com/automl/SMAC3. 2017.

[17] D. Helbing and P. Molnar. "Social force model for pedestrian dynamics". In: *Physical review E* 51.5 (1995), p. 4282.

[18] I. Karamouzas, P. Heil, P. van Beek, and M. H. Overmars. "A predictive collision avoidance model for pedestrian simulation". In: *Int. Workshop on Motion in Games*. Springer. 2009, pp. 41–52.

[19] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi. "Social GAN: Socially Acceptable Trajectories With Generative Adversarial Networks". In: *Proc. of the IEEE Conf. on Comp. Vis. and Pat. Rec. (CVPR)*. June 2018.

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

[21] S. Pellegrini, A. Ess, K. Schindler, and L. van Gool. "You'll never walk alone: Modeling social behavior for multi-target tracking". In: *Proc. of the IEEE Int. Conf. on Computer Vision (ICCV)*. 2009, pp. 261–268.

[22] A. Lerner, Y. Chrysanthou, and D. Lischinski. "Crowds by example". In: *Computer Graphics Forum*. Vol. 26. 3. Wiley Online Library. 2007, pp. 655–664.

[23] B. Majecka. "Statistical models of pedestrian behaviour in the forum". In: *Master's thesis, School of Informatics, University of Edinburgh* (2009).

[24] D. Brščić, T. Kanda, T. Ikeda, and T. Miyashita. "Person tracking in large public spaces using 3-D range sensors". In: *IEEE Trans. on Human-Machine Systems* 43.6 (2013), pp. 522–534.

[25] A. Robicquet, A. Sadeghian, A. Alahi, and S. Savarese. "Learning social etiquette: Human trajectory understanding in crowded scenes". In: *Proc. of the Europ. Conf. on Comp. Vision (ECCV)*. Springer. 2016, pp. 549–565.

[26] A. Rudenko, T. P. Kucner, C. S. Swaminathan, R. T Chadalavada, K. O. Arras, and A. J. Lilienthal. "THÖR: Human-Robot Navigation Data Collection and Accurate Motion Trajectories Dataset". In: *IEEE Robotics and Automation Letters* 5.2 (2020), pp. 676–682.

[27] F. Zanlungo, T. Ikeda, and T. Kanda. “Social force model with explicit collision prediction”. In: *EPL (Europhysics Letters)* 93.6 (2011), p. 68005.

[28] Wanting Huang. “Benchmarking Local Interaction Models for Human Motion Prediction in Social Spaces”. In: *M.Sc. Thesis*. 2020.

[29] Stefano Pellegrini, Andreas Ess, Konrad Schindler, and Luc van Gool. “You’ll Never Walk Alone: Modeling Social Behavior for Multi-target Tracking”. In: *Int. Conf. on Computer Vision*. 2009.

[30] Julian Bock, Robert Krajewski, Tobias Moers, Steffen Runde, Lennart Vater, and Lutz Eckstein. “The ind dataset: A drone dataset of naturalistic road user trajectories at german intersections”. In: *2020 IEEE Intelligent Vehicles Symposium (IV)*. IEEE. 2020, pp. 1929–1934.

[31] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot. “NTU RGB+D 120: A large-scale benchmark for 3d human activity understanding”. In: *IEEE transactions on pattern analysis and machine intelligence* 42.10 (2019), pp. 2684–2701.

[32] Philipp Kratzer, Simon Bihlmaier, Niteesh Balachandra Midlagajni, Rohit Prakash, Marc Toussaint, and Jim Mainprice. “MoGaze: A Dataset of Full-Body Motions that Includes Workspace Geometry and Eye-Gaze”. In: *IEEE Robotics and Automation Letters (RAL)* (2020).

[33] Mahsa Ehsanpour, Fatemeh Sadat Saleh, Silvio Savarese, Ian D. Reid, and Hamid Rezatofighi. “JRDB-Act: A Large-scale Dataset for Spatio-temporal Action, Social Group and Activity Detection”. In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), pp. 20951–20960.

[34] Luigi Palmieri, Tomasz P Kucner, Martin Magnusson, Achim J Lilienthal, and Kai O Arras. “Kinodynamic motion planning on Gaussian mixture fields”. In: *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*. IEEE. 2017, pp. 6176–6181.

[35] Chittaranjan Srinivas Swaminathan, Tomasz Piotr Kucner, Martin Magnusson, Luigi Palmieri, and Achim J Lilienthal. “Down The CLiFF: Flow-aware Trajectory Planning under Motion Pattern Uncertainty”. In: *Proc. of the IEEE Int. Conf. on Intell. Robots and Syst. (IROS)*. IEEE. 2018, pp. 7403–7409.

[36] Rudolph Triebel, Kai Arras, Rachid Alami, Lucas Beyer, Stefan Breuers, Raja Chatila, Mohamed Chetouani, Daniel Cremers, Vanessa Evers, Michelangelo Fiore, et al. “Spencer: A socially aware service robot for passenger guidance and help in busy airports”. In: *Field and service robotics*. Springer. 2016, pp. 607–622.

[37] A. Rudenko, L. Palmieri, A. J. Lilienthal, and K. O. Arras. “Human Motion Prediction under Social Grouping Constraints”. In: *Proc. of the IEEE Int. Conf. on Intell. Robots and Syst. (IROS)*. 2018.

[38] Tomasz Piotr Kucner, Martin Magnusson, Erik Schaffernicht, Victor H. Bennetts, and Achim J. Lilienthal. “Enabling Flow Awareness for Mobile Robots in Partially Observable Environments”. In: *IEEE Robotics and Automation Letters* 2.2 (Apr. 2017), pp. 1093–1100.

[39] B. D. Ziebart, N. Ratliff, G. Gallagher, C. Mertz, K. Peterson, J. A. Bagnell, M. Hebert, A. K. Dey, and S. Srinivasa. “Planning-based prediction for pedestrians”. In: *Proc. of the IEEE Int. Conf. on Intell. Robots and Syst. (IROS)*. 2009, pp. 3931–3936.

[40] V. Karasev, A. Ayvaci, B. Heisele, and S. Soatto. “Intent-aware long-term prediction of pedestrian motion”. In: *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*. 2016, pp. 2543–2549.

[41] E. Rehder, F. Wirth, M. Lauer, and C. Stiller. “Pedestrian prediction by planning using deep neural networks”. In: *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*. 2018, pp. 1–5.

[42] A. Rudenko, L. Palmieri, and K. O. Arras. “Joint Prediction of Human Motion Using a Planning-Based Social Force Approach”. In: *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*. 2018, pp. 1–7.

[43] W. Liu, D. Lian W. Luo, and S. Gao. “Future Frame Prediction for Anomaly Detection – A New Baseline”. In: *Proc. of the IEEE Conf. on Comp. Vis. and Pat. Rec. (CVPR)*. 2018.

[44] Tharindu Fernando, Simon Denman, Sridha Sridharan, and Clinton Fookes. “Soft+Hardwired attention: An LSTM framework for human trajectory prediction and abnormal event detection”. In: *Neural networks* 108 (2018), pp. 466–478.

[45] Mohammad Mahdavian, Payam Nikdel, Mahdi TaherAhmadi, and Mo Chen. “STPOTR: Simultaneous Human Trajectory and Pose Prediction Using a Non-Autoregressive Transformer for Robot Follow-Ahead”. In: *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*. IEEE. 2023, pp. 9959–9965.

[46] Judith Bütepage, Hedvig Kjellström, and Danica Kragic. “Anticipating many futures: Online human motion prediction and generation for human-robot interaction”. In: *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*. IEEE. 2018, pp. 4563–4570.

[47] Karttikeya Mangalam, Ehsan Adeli, Kuan-Hui Lee, Adrien Gaidon, and Juan Carlos Niebles. “Disentangling human dynamics for pedestrian locomotion forecasting with noisy supervision”. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 2784–2793.

[48] Suyuan Li and Xin Song. “Future Frame Prediction Network for Human Fall Detection in Surveillance Videos”. In: *IEEE Sensors Journal* (2023).

[49] Łukasz Kidziński, Bryan Yang, Jennifer L Hicks, Apoorva Rajagopal, Scott L Delp, and Michael H Schwartz. “Deep neural networks enable quantitative movement analysis using single-camera videos”. In: *Nature communications* 11.1 (2020), p. 4054.

[50] Zhe Cao, Hang Gao, Karttikeya Mangalam, Qi-Zhi Cai, Minh Vo, and Jitendra Malik. “Long-term human motion prediction with scene context”. In: *Proc. of the Europ. Conf. on Comp. Vision (ECCV)*. Springer. 2020, pp. 387–404.

[51] Payam Nikdel, Mohammad Mahdavian, and Mo Chen. “DMMGAN: Diverse Multi Motion Prediction of 3D Human Joints using Attention-Based Generative Adversarial Network”. In: *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2023, pp. 9938–9944.

[52] V. V. Unhelkar, C. Pérez-D’Arpino, L. Stirling, and J. A. Shah. “Human-robot co-navigation using anticipatory indicators of human walking motion”. In: *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*. 2015, pp. 6183–6190.

[53] Elisa Stefanini, Luigi Palmieri, Andrey Rudenko, Till Hielscher, Timm Linder, and Lucia Pallottino. “Efficient Context-Aware Model Predictive Control for Human-Aware Navigation”. In: *IEEE Robotics and Automation Letters*. 2024.

[54] Ye Yuan and Kris Kitani. “Dlow: Diversifying latent flows for diverse human motion prediction”. In: *Proc. of the Europ. Conf. on Comp. Vision (ECCV)*. Springer. 2020, pp. 346–364.

[55] Angel Martínez-González, Michael Villamizar, and Jean-Marc Odobez. “Pose transformers (POTR): Human motion prediction with non-autoregressive transformers”. In: *Proc. of the IEEE Conf. on Comp. Vis. and Pat. Rec. (CVPR)*. 2021, pp. 2276–2284.

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

[57] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. "Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments". In: *IEEE Trans. on Patt. Anal. and Mach. Intell. (PAMI)* 36.7 (2014), pp. 1325–1339.

[58] *CMU Graphics Lab Motion Capture Database*. Accessed: 01-09-2023.

[59] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. "Graph attention networks". In: *arXiv preprint arXiv:1710.10903* (2017).

[60] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need". In: *Proc. of the Advances in Neural Information Processing Systems (NIPS)* 30 (2017).

[61] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. *Self-Attention with Relative Position Representations*. 2018. arXiv: 1803.02155 [cs.CL].

[62] Oriane Dermy, François Charpillet, and Serena Ivaldi. "Multi-modal intention prediction with probabilistic movement primitives". In: *Human Friendly Robotics: 10th International Workshop*. Springer. 2019, pp. 181–196.

[63] F.B. Flohr. "Vulnerable road user detection and orientation estimation for context-aware automated driving". PhD thesis. Faculty of Science (FNWI): Universiteit van Amsterdam, 2018, p. 194.

[64] Oskar Palinko, Francesco Rea, Giulio Sandini, and Alessandra Sciutti. "Robot reading human gaze: Why eye tracking is better than head tracking for human-robot collaboration". In: *Proc. of the IEEE Int. Conf. on Intell. Robots and Syst. (IROS)*. 2016.

[65] Tim Schreiter, Lucas Morillo-Mendez, Ravi T Chadalavada, Andrey Rudenko, Erik Billing, Martin Magnusson, Kai O Arras, and Achim J Lilienthal. "Advantages of Multimodal versus Verbal-Only Robot-to-Human Communication with an Anthropomorphic Robotic Mock Driver". In: *32nd IEEE RO-MAN, 2023, Busan, Korea*. IEEE. 2023, pp. 293–300.

[66] Shreya Ghosh, Abhinav Dhall, Munawar Hayat, Jarrod Knibbe, and Qiang Ji. "Automatic gaze analysis: A survey of deep learning based approaches". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46.1 (2023), pp. 61–84.

[67] Tanja Blascheck, Kuno Kurzhals, Michael Raschke, Michael Burch, Daniel Weiskopf, and Thomas Ertl. "Visualization of eye tracking data: A taxonomy and survey". In: *Computer Graphics Forum*. Vol. 36. 8. Wiley Online Library. 2017, pp. 260–284.

[68] Flora Ioannidou, Frouke Hermens, Timothy Hodgson, et al. "The centrial bias in day-to-day viewing". In: *Journal of Eye Movement Research* 9.6 (2016), pp. 1–13.

[69] John M Franchak, Brianna McGee, and Gabrielle Blanch. "Adapting the coordination of eyes and head to differences in task and environment during fully-mobile visual exploration". In: *PLoS one* 16.8 (2021).

[70] John S Stahl. "Amplitude of human head movements associated with horizontal saccades". In: *Experimental brain research* 126 (1999), pp. 41–54.

[71] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. *Ultralytics YOLOv8*. https://github.com/ultralytics/ultralytics. 2023.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101017274