2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Extended Force and Velocity Prediction in Human-Robot Collaborative Transportation through Future Environment Representation Estimation

J. E. Domínguez-Vidal

***Abstract*— In this work, we address the challenge of predicting human-applied force and velocity during collaborative object transportation over extended distances (5–8 m). We enhance state-of-the-art predictors by refining their input data processing, which significantly improves prediction accuracy. Furthermore, we extend the temporal prediction horizon from 1 s to 2 s without compromising performance, by introducing an extra environmental prediction module that conditions force and velocity estimations based on anticipated sensory input. This integration captures the contextual dependency of human behaviour during joint transport. Experimental evaluations, both on dataset and in real-world settings, validate the effectiveness of our approach. Specifically, our best model manages to achieve success rates in testset of up to 90.4% in predicting the human's exerted force and up to 93.0% in the velocity of the human-robot pair during the next 2 s, and up to 87.1% and 91.3% respectively in real experiments.**

***Index Terms*— Physical Human-Robot Interaction, Object Transportation, Force Prediction, Human-in-the-Loop**

I. Introduction

In the past decade, we have witnessed a remarkable surge in the collaborative capabilities of robotics when performing tasks alongside humans in increasingly efficient and seamless ways. This includes a wide range of tasks such as collaborative assembly [1], [2], handover [3], [4], or collaborative object transportation [5], [6], among others. The primary reason behind this rise has been the significant advances in fields such as automatic control and artificial intelligence, and more prominently, in Deep Learning (DL). In turn, robotics has served as a testbed for these disciplines, further promoting their development. Specifically, this interplay between disciplines has focused on the development of both predictors of the human's future actions—enabling it to anticipate and respond accordingly [7]—and decision-making algorithms, allowing the robot to determine how to act based on the information provided by the human [8], [9].

In this work, we focus on the task of human–robot collaborative transportation (see Fig.1) and leverage several advances in DL to improve the predictive capabilities of the tools available in the current literature. Specifically, our goal is to enhance the accuracy of existing predictors of human-applied force and velocity during the joint transportation of an object across relatively long distances (5-8 m) within the environment in which the task is performed [10].

To achieve this, we first select the best-performing predictors currently available in the literature and improve their performance by addressing a common issue across all of them regarding the processing of input information used to generate predictions. Secondly, we extend the temporal prediction horizon from a maximum of 1 s to 2 s without incurring any significant loss in accuracy. To do so, we introduce an additional prediction of the environment the robot will perceive next, which we use to condition the other two predictions, given the crucial influence of the surroundings on how the task unfolds. Last but not least, we publicly release the full dataset used to train all of our predictors, this being the third of our contributions since previous authors have not made their datasets available, thereby hindering the advancement of such tools.

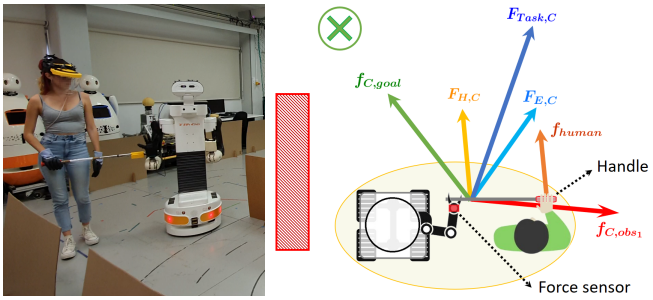

Fig. 1. **Example of one of the collaborative transportation set-ups.** *Left* - The human and the robot must transport an aluminium bar until a predefined goal through a complex scenario with multiple routes. OptiTrack to detect and track the human, the robot and the goal. The robot has a force sensor to detect human's exerted force and a LiDAR to detect the environment. *Right* - Main forces considered in the same situation. Goal attractive force in green, obstacle repulsive force in red, environment force in yellow, human real force in orange, human's transformed force in light blue and total force in dark blue.

In the remainder of the article, Section II presents the state of the art related to this work. Section III presents the architecture of the force and velocity predictors considered in this article as well as the considerations about their training. Section IV shows the results obtained regarding the performance of each predictor both in dataset and real experiments. Finally, Section V presents the conclusions.

II. Related Work

In the field of collaborative object manipulation between humans and robots, early approaches commonly employed control strategies based on admittance [11], [12] and impedance [13], [14] variation. These methods were primarily designed to ensure the robot could rapidly respond to human-applied forces in a smooth manner. Some research efforts also incorporated predictive components—either forecasting human motion [15] or anticipating object trajectories [16]—to enhance the controllers' performance. However,

The author is with the Polytechnic University of Catalonia (UPC). Jordi Girona, 31, 08034, Barcelona, Spain. jose.enrique.dominguez@upc.edu.

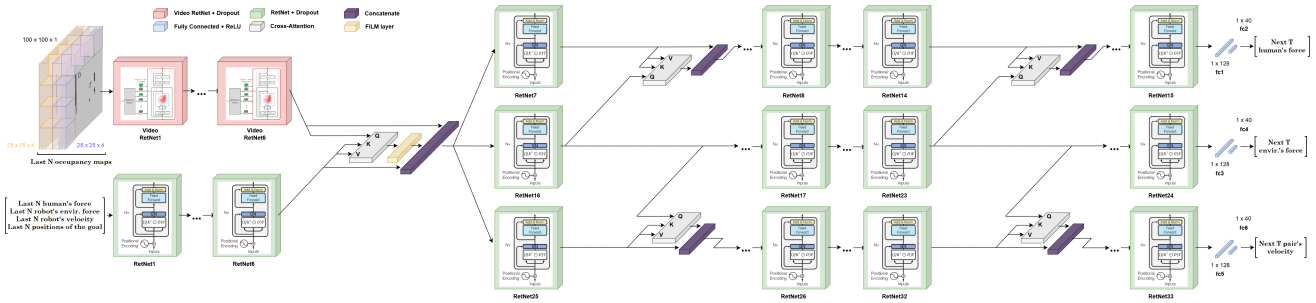

Fig. 2. **Model architecture for RetNet-CA-SFM force/velocity predictor.** Two input streams in parallel, one to process occupancy maps obtained from LiDAR using Video RetNets and another to process secuencial inputs with vanila RetNets. Both streams combined using Cross-Attention and FiLM layers and processed using three output streams, one of them to generate a prediction of the environmental force with which conditioning the other two outputs.

it wasn't until the advancement and widespread use of Deep Learning (DL) techniques that the accuracy of such predictions significantly improved. For instance, [17] adopts Reinforcement Learning (RL) to model the uncertainty in human behaviour, which is then integrated into a Model Predictive Control (MPC) framework. Additionally, works such as [18], [19] leverage Learning from Demonstration (LfD) either to acquire the task representation or to infer the velocity profile that the human partner aims to follow.

Most previous works may or may not use the human-applied force as an input, but they never attempt to predict it. Instead, they typically estimate other task-relevant variables, such as the desired trajectory of the transported object or its velocity profile. This is not the case in the work by Fusco et al. [5], where a Transformer-based model [20] is proposed to predict both human motion and the contact forces over a time window of up to 800 ms in a short-distance transportation task (involving only arm movements and no actual displacement). Similarly, [21]–[23] also perform a prediction of the future force that the human will exert, as well as the velocity profile that the human–robot pair will follow over the next 1 s, but in long-distance collaborative transportation tasks. Moreover, they show that both the force prediction [21] and the velocity prediction [22] can be used to generate an estimate of the trajectory to be followed by both agents, providing a richer representation capable of detecting rapid changes in human intent earlier.

However, these latter works do not leverage the relationships between the various input variables considered, thereby limiting the temporal horizon of their predictions. In this work, we employ Retentive Networks (RetNet) [24], [25] as our core architecture, and we incorporate several techniques such as the use of Cross-Attention modules [26], [27] and the inclusion of FiLM layers [28] among others to address the aforementioned limitation.

III. 3-Output Stream Force/Velocity Predictors for Collaborative Object Transportation

In order to compare our results with those reported in the literature, we adopt a setup similar to the one used in [23]. Accordingly, we use an indoor environment of size 5x8 m, featuring multiple walls and columns in the middle, so that the human-robot pair has several alternative paths to transport the object from the initial point to the goal, both of which are pre-defined. They may change their path on the fly as many times as they wish.

A. Problem Formulation

Two predictions are to be obtained. First, an estimate of the next T force measurements in the (x, y) plane applied by the human on the transported object, $Y^{force}_{N+1:N+T} \in \mathbb{R}^{2,T}$. Second, an estimate of the next T measurements of the (linear and angular) velocity of the human-robot pair, $Y^{vel}_{N+1:N+T} \in \mathbb{R}^{2,T}$. This dual-output system has proven particularly useful for this task, as the first variable can be used to detect rapid (albeit typically less precise) changes in the human's intention (in this case understood as implementation intention according to [29]), while the second variable provides a prediction of the trajectory that the pair will follow. The z-axis torque is not predicted making our work comparable with the state-of-the-art.

To make both predictions, five input information flows are used. The first two represent the environment in which the task is being performed in two different ways. The first involves converting the robot's LiDAR readings into occupancy maps of size 100x100 pixels. Each pixel indicates whether the corresponding 10x10 cm cell in the environment is occupied by an obstacle or not. The second approach consists of post-processing these maps and applying a logic inspired by the well-known Social Force Model (SFM) [30]. To this end, clustering techniques are applied to identify the O obstacles visible from the robot's perspective. Each of these obstacles is assigned a repulsive force, $\boldsymbol{f_{C,obs_i}} \in \mathbb{R}^2$, emulating the pair's tendency to avoid that obstacle. Simultaneously, the robot uses a global planner to generate the optimal path to the goal, enabling the use of waypoints obtained from this planner to which an attractive force, $\boldsymbol{f_{C,goal}} \in \mathbb{R}^2$, is assigned representing the tendency of both agents to follow the optimal path. These forces can be combined in a weighted manner into a single virtual force, $\boldsymbol{f_{E,C}} \in \mathbb{R}^2$, to represent the practical implications of the current environmental situation[1].

The remaining three sources of input information are as follows: First, the force applied by the human and measured

[1]For an example of how to calculate $\boldsymbol{F_{E,C}}$ and the related experiments, see: `https://youtu.be/CO3V3bLQ6Jw`

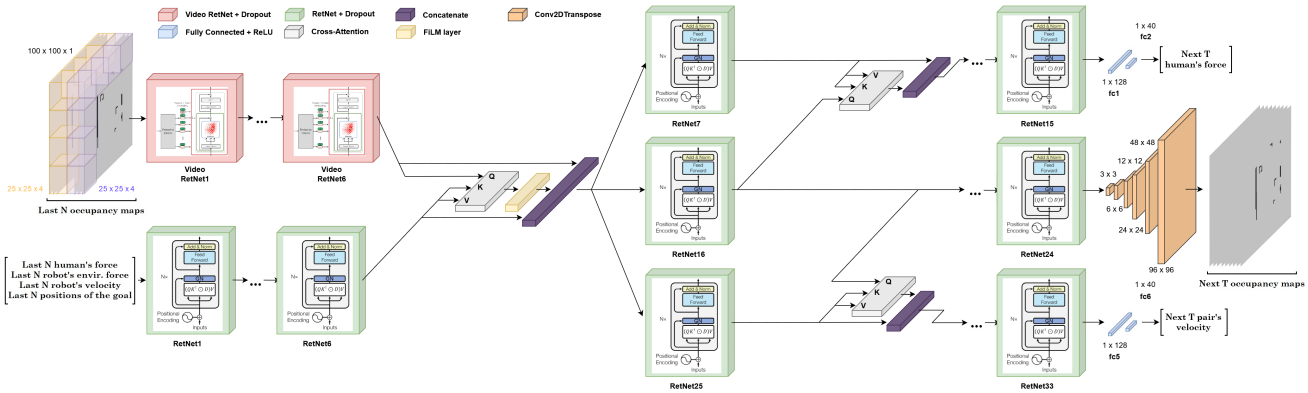

Fig. 3. **Model architecture for RetNet-CA-OM force/velocity predictor.** Same input structure as with RetNet-CA-SFM predictor and also using three output streams. In this case, one of them to generate a prediction of the next occupancy maps with which conditioning the other two outputs.

by the robot using a force sensor mounted on its wrist, $\boldsymbol{F_{H,C}} \in \mathbb{R}^2$. Second, the (linear and angular) velocity commands used by the robot to move its base. These commands are computed by jointly processing the force applied by the human and the virtual force representing the environment, $\boldsymbol{F_{E,C}}$. Finally, the fifth information stream corresponds to the distance (magnitude and angle) between the human-robot pair and the goal. To ensure that all these variables can be used effectively, they are normalised to the range $[-1, 1]$. Specifically, a maximum magnitude of 12 N is assumed for each force, maximum robot velocities of 0.65 m/s (linear) and 1 rad/s (angular), and a maximum distance to the goal of 7 m.

Taking all the above into account, we use the latest N occupancy maps generated from the LiDAR readings, $X^{map}_{1:N}$, as well as the concatenation of the most recent N samples from the other sources of information, $X^{f}_{1:N}$ with $x^{f}_{i} \in \mathbb{R}^8$. The goal is to predict the next T samples of the force that the human will apply, $Y^{force}_{N+1:N+T}$, as well as the velocity of the human-robot pair, $Y^{vel}_{N+1:N+T}$. In this case, $N = 20$ and $T = 20$, meaning that the past 2 s are used to predict the following 2 s, as the system runs at 10 Hz.

B. Force/Velocity Prediction Models

As in [23], we also employ Retentive Networks (RetNet) [24] to process our sequential signals, due to their demonstrated ability to achieve performance comparable to that of Transformers, while offering a more efficient use of computational resources. Similarly, we adopt the generalisation introduced in [23] to enable the use of such networks for processing video sequences (occupancy map sequences in our case). Accordingly, the input signals are processed using two parallel pipelines. However, unlike [23], we do not concatenate their outputs directly, as this would result in the loss of dependencies between the occupancy maps and the other input variables. Instead, we use a Cross-Attention block to model these dependencies, using the features extracted from the processed maps as Queries, and the features corresponding to the remaining inputs as Keys and Values. The output of this block is passed through a FiLM layer [28], and its result is concatenated with the original features extracted from both the occupancy maps and the other inputs. This model will be called the *RetNet-3D-CA* version in the following sections.

Fig. 2 illustrates this architecture, along with the second of the modifications introduced. This modification builds upon the observation that the environment is the most decisive factor when predicting both the force exerted by the human and the velocity of the pair [22]. To address this, a third output stream is incorporated with the aim of predicting the evolution of the virtual force representing the environment, $\boldsymbol{f_{E,C}}$, over the next T time steps. However, the goal is not to predict this signal, but rather to use the intermediate representations obtained during its computation to condition the prediction of the two target variables. To achieve this, Cross-Attention (CA) blocks are inserted between streams, using the layer-wise outputs of the RetNets from the environment output stream as Queries, and the corresponding layer outputs of the other two streams as Keys and Values. The output of each CA block is concatenated with the corresponding layer output, thus preserving the residual connection. This version will be called *RetNet-3D-CA-SFM*.

Fig. 3 illustrates a variation of this strategy. $\boldsymbol{f_{E,C}}$ is a compressed, two-dimensional representation of the environment in which the task is being performed, whereas an occupancy map represents the state of the setup in an uncompressed form. For this reason, if we want to use a better representation of the (future) environment we can replace the prediction of the next T samples of $\boldsymbol{f_{E,C}}$ with an estimation of the next T occupancy maps, using a standard autoencoder architecture. As in the previous case, Cross-Attention blocks with residual connections are used to condition the generation of the two target variables. We will call this version *RetNet-3D-CA-OM*.

Regarding the internal parameters used across all models, the first input stream processes $X^{map}_{1:N}$ by employing tubelets containing $L = 4$ consecutive patches of size 25x25 pixels, which are passed through 8 layers of Video RetNets. These layers use $h = 8$ self-attention heads, with a projection dimensionality of 128 and a dropout probability of $p = 0.3$. The second input stream processes $X^{f}_{1:N}$ using 6 RetNet

TABLE I

EVOLUTION OF MEAN ERROR AND PERCENTAGE OF CORRECT PREDICTIONS IN TESTSET. VARIABLE Y REPRESENTS FORCE (F) OR VELOCITY (Vel).

Measure		Time [ms]							
		Force ($Y = F$)				Velocity ($Y = Vel$)			
		500	1000	1500	2000	500	1000	1500	2000
Error Y_x [N or m/s]	CNN+LSTM [21]	0.269	0.293	0.388	0.552	–	–	–	–
	ViViT+T [22]	0.220	0.244	0.309	0.439	0.0072	0.0090	0.0144	0.0252
	RetNet-3D [23]	0.216	0.236	0.298	0.422	0.0068	0.0085	0.0135	0.0238
	RetNet-3D-CA	0.209	0.225	0.273	0.368	0.0063	0.0078	0.0118	0.0197
	RetNet-3D-CA-SFM	0.202	0.214	0.253	0.322	0.0059	0.0070	0.0102	0.0160
	RetNet-3D-CA-OM	**0.200**	**0.209**	**0.247**	**0.309**	**0.0058**	**0.0066**	**0.0096**	**0.0148**
Error Y_y [N or rad/s]	CNN+LSTM [21]	0.136	0.150	0.203	0.294	–	–	–	–
	ViViT+T [22]	0.109	0.122	0.158	0.231	0.0050	0.0065	0.0112	0.0204
	RetNet-3D [23]	0.106	0.117	0.152	0.221	0.0047	0.0060	0.0103	0.0191
	RetNet-3D-CA	0.103	0.111	0.138	0.192	0.0042	0.0054	0.0088	0.0154
	RetNet-3D-CA-SFM	0.099	0.105	0.127	0.166	0.0039	0.0047	0.0074	0.0122
	RetNet-3D-CA-OM	**0.098**	**0.103**	**0.124**	**0.158**	**0.0038**	**0.0044**	**0.0069**	**0.0113**
Error $\lvert\boldsymbol{Y}\rvert < 0.1 \cdot Y_{max}$ & Error $\angle\boldsymbol{Y} < 18°$ [%]	CNN+LSTM [21]	92.2	91.1	86.8	79.4	–	–	–	–
	ViViT+T [22]	94.4	93.3	90.4	84.5	97.2	96.2	93.1	87.0
	RetNet-3D [23]	94.5	93.6	90.9	85.3	97.4	96.5	93.7	87.9
	RetNet-3D-CA	94.9	94.2	92.0	87.7	97.7	96.9	94.7	90.3
	RetNet-3D-CA-SFM	95.2	94.7	92.9	89.8	97.9	97.4	95.6	92.4
	RetNet-3D-CA-OM	**95.3**	**94.9**	**93.2**	**90.4**	**98.0**	**97.6**	**95.9**	**93.0**

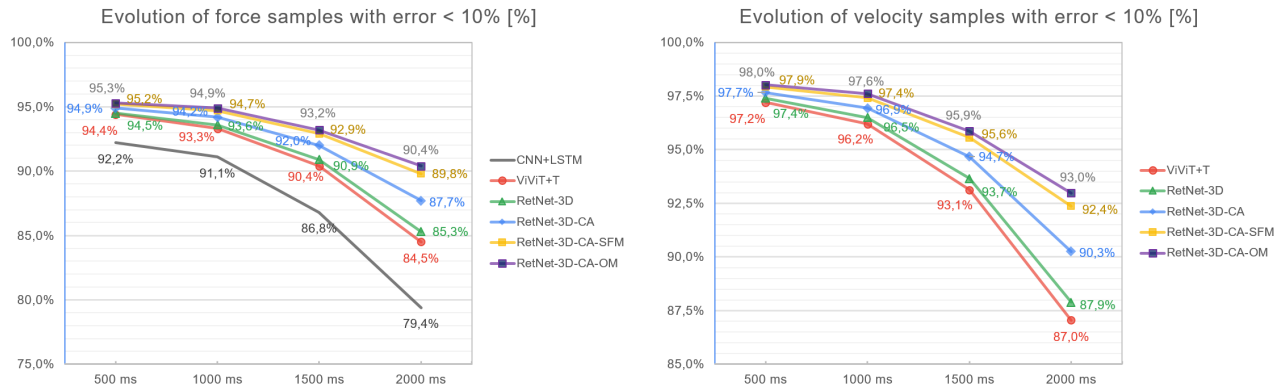

Fig. 4. **Evolution of percentage of predicted samples below considered threshold from 500 to 2000 ms using each model.** *Left* - Performance variation for force prediction. *Right* - Performance variation for velocity prediction. *CNN+LSTM* missing since it does not generate a velocity prediction.

layers with $h = 8$ heads and dimensionalities of 64 for the projections, 512 for the inner fully connected layers, and 128 for the sub-layer outputs. In this case, dropout is applied with a probability of $p = 0.25$. For the output streams, 9 RetNet layers are used per branch, with the same parameters as in the previous case.

C. Dataset Acquisition and Training

The dataset used in this study is publicly available on GitHub[2]. A total of 120 volunteers participated in up to 448 collaborative transportation experiments conducted across 6 scenarios of varying complexity, each including at least two routes to reach the goal (see Fig. 1) except for the most complex scenario, which offers at least eight possible routes . As a result, the dataset contains 17400 sub-sequences, obtained by dividing each experiment into blocks of $N + T$ samples, with each sub-sequence containing the five specified input signals. Additionally, an overlap of $(N+T)/2$ samples is applied between sub-sequences for data augmentation purposes. The dataset is split into the standard partitions: training (90%: 15660 sub-sequences), validation (5%: 870 sub-sequences), and testing (5%: 870 sub-sequences).

All models were implemented using TensorFlow. The Adam optimiser was used, with an initial learning rate of $lr = 5 \times 10^{-4}$ and a learning rate decay factor of 0.97, down to a minimum of $lr_{min} = 3 \times 10^{-5}$. Early stopping was applied to prevent overfitting, and a warm-up phase of 1 epoch was introduced for the output stream responsible

[2]Dataset URL: `https://github.com/JEDominguezVidal/human_force_prediction_extended_dataset`

TABLE II

PERFORMANCE OBTAINED FOR EACH MODEL WITH DIFFERENT GRAPHIC CARDS

Model	Frames Per Second (min. / avg. / max.)				AVG.
	GTX 1060 Mobile (80 W)	GTX 1660 Ti Desktop	RTX 3060 Mobile (80 W)	RTX 3080 Ti Desktop	
CNN+LSTM [21]	13.1 - 13.9 - 14.2	18.2 - 18.9 - 20.2	21.1 - 22.5 - 24.0	54.5 - 58.6 - 62.1	28.5
ViViT+T [22]	6.79 - 7.29 - 7.89	9.10 - 9.89 - 10.7	10.6 - 11.8 - 12.7	30.1 - 32.2 - 34.3	15.3
RetNet-3D [23]	8.56 - 9.20 - 9.95	11.6 - 12.3 - 13.2	14.4 - 15.5 - 16.5	40.0 - 42.7 - 45.3	19.9
RetNet-3D-CA	8.47 - 9.11 - 9.81	11.5 - 12.2 - 13.0	14.3 - 15.3 - 16.3	39.9 - 42.2 - 44.9	19.7
RetNet-3D-CA-SFM	7.06 - 7.51 - 8.12	9.32 - 10.1 - 10.9	11.2 - 12.4 - 13.3	31.6 - 33.6 - 35.7	15.7
RetNet-3D-CA-OM	6.05 - 6.47 - 6.96	8.08 - 8.80 - 9.58	9.61 - 10.8 - 11.6	27.3 - 29.2 - 31.3	13.8

for predicting the environment-related variable (i.e., virtual force in *RetNet-3D-CA-SFM* and occupancy maps in *RetNet-3D-CA-OM*). All models were trained on an NVIDIA RTX 3080 Ti graphics card, with training times ranging from 110 to 480 minutes depending on the model. This wide variability is explained by the fact that the *RetNet-3D-CA-OM* model required up to 3.9 times more training time, as it generates images in one of its output streams. No model exceeded epoch 97 due to early stopping.

IV. RESULTS

To evaluate the effectiveness of our models, three tests have been conducted. Firstly, the testset split was used to assess the prediction error of these models in comparison with those reported in the literature. Secondly, their performance was evaluated in real-world experiments. To this end, 15 new experiments not included in the original dataset and, therefore, with different human preferences are used. These experiments were conducted without the use of any predictor in order to avoid influencing the robot's movements. Subsequently, and in an offline manner, each predictor was executed encapsulated within a ROS node using the recorded data to assess its actual performance. Thirdly, since we also recorded the trajectory followed by the pair using the OptiTrack motion tracking system, we are able to evaluate the usefulness of each predictor as a trajectory estimator. All the experiments presented in this work have been performed with the approval of the ethics committee of the associated university.

A. Force/Velocity Predictor Performance in Testset split

To assess the accuracy of our predictors, we employ the metrics proposed in [21]. Accordingly, we compute the mean absolute error for each Cartesian axis in the force prediction, and the mean absolute error for both linear and angular velocity with respect to their ground truth values. Additionally, we calculate the percentage of samples with an error—both in magnitude and direction—below 10% (force: $1.2N$, velocity: $0.065m/s$ and $0.1\ rad/s$), as this threshold is considered in the literature to be the maximum acceptable.

Table I presents, in its first two row groups of rows, the evolution of absolute errors in each axis for both predictions, and in its final group, the evolution of the percentage of samples deemed acceptable. This latter metric is graphically represented in Fig. 4 for easier visualisation.

As expected, *ViViT+T* outperforms *CNN+LSTM*, demonstrating that ViViTs can indeed outperform CNNs when trained on sufficiently large datasets [31], [32]. Similarly, the performance gain between *ViViT+T* and *RetNet-3D* is residual. However, when relationships among input variables are taken into account through the modifications introduced in the *RetNet-3D-CA* model, an improvement comparable to that achieved when moving from *CNN+LSTM* to *ViViT+T* is observed. Finally, a similar improvement is achieved with the *RetNet-3D-CA-SFM* and *RetNet-3D-CA-OM* models, the latter offering the best performance across all evaluated metrics, albeit by a very narrow margin compared to *RetNet-3D-CA-SFM*. Specifically, *RetNet-3D-CA-OM* model achieves prediction accuracy at 2000 ms comparable to that of *CNN+LSTM* at a time horizon of just 1000 ms, or that of *RetNet-3D* at 1500 ms, demonstrating the effectiveness of the implemented enhancements.

While the *RetNet-3D-CA-OM* model is the most accurate, this comes at the cost of a significant computational burden. Table II shows the frame rate in Frames Per Second (FPS) at which each model can deliver a complete prediction (calculated as the inverse of the inference time after receiving a new set of input data) using various graphics cards. It is worth noting that, since the graphics cards belong to different generations, the same graphics drivers have not been used for all of them. Instead, the most up-to-date version available for each card was employed. As can be observed, the *RetNet-3D-CA-OM* model is the most demanding, with even an RTX 3060 Mobile struggling to consistently deliver the desired 10 FPS (the robot's LiDAR typically operates at 10 Hz, which is the standard frequency for most systems). This makes the *RetNet-3D-CA-SFM* model more advisable for setups with hardware limitations.

B. Predictor Performance in Real Experiments

It is essential to evaluate the performance of our models with real-world experiments not included in the dataset, as the ultimate goal of such tools is to be employed in real use cases. To this end, we used 15 new experiments not present in the dataset (ethics approval ID: 2023.05, given by Polytechnic University of Catalonia) and evaluated the

TABLE III

MEAN ERROR AND PERCENTAGE OF CORRECT PREDICTIONS IN REAL EXPERIMENTS. VARIABLE Y REPRESENTS FORCE (F) OR VELOCITY (Vel).

Measure		Time [ms]							
		Force ($Y = F$)				Velocity ($Y = Vel$)			
		500	1000	1500	2000	500	1000	1500	2000
Error Y_x [N or m/s]	RetNet-3D [23]	0.270	0.297	0.366	0.500	0.0096	0.0115	0.0168	0.0274
	RetNet-3D-CA-SFM	0.254	0.273	0.321	0.399	0.0086	0.0098	0.0133	0.0193
	RetNet-3D-CA-OM	**0.252**	**0.266**	**0.311**	**0.386**	**0.0084**	**0.0095**	**0.0128**	**0.0181**
Error Y_y [N or rad/s]	RetNet-3D [23]	0.136	0.152	0.192	0.266	0.0069	0.0086	0.0133	0.0222
	RetNet-3D-CA-SFM	0.128	0.139	0.166	0.210	0.0061	0.0071	0.0102	0.0153
	RetNet-3D-CA-OM	**0.126**	**0.136**	**0.161**	**0.203**	**0.0060**	**0.0069**	**0.0098**	**0.0143**
Error $\lvert\boldsymbol{Y}\rvert < 0.1 \cdot Y_{max}$ & Error $\angle\boldsymbol{Y} < 18°$ [%]	RetNet-3D [23]	92.2	91.0	87.9	81.9	95.9	94.9	91.9	86.0
	RetNet-3D-CA-SFM	92.9	92.1	90.0	86.5	96.4	95.8	93.9	90.6
	RetNet-3D-CA-OM	**93.0**	**92.4**	**90.4**	**87.1**	**96.5**	**96.0**	**94.2**	**91.3**

best model available in the literature, *RetNet-3D*, as well as our best model, *RetNet-3D-CA-OM*. We also included the *RetNet-3D-CA-SFM* model in this comparison, as it offers the best performance-resource consumption ratio.

Table III shows the results. As can be seen, both *RetNet-3D-CA-SFM* and *RetNet-3D-CA-OM* outperform *RetNet-3D* across all metrics, with the latter being the best of the two. Focusing on the percentage of samples considered correct, the *RetNet-3D-CA-SFM* model shows a drop in performance of up to 3.3% in force prediction and up to 1.8% in velocity prediction. In contrast, the *RetNet-3D-CA-OM* model exhibits drops of up to 3.3% and 1.7%, respectively. These drops are due to the fact that the dataset includes experiments conducted in up to 6 different setups to achieve greater model generalisation, while here only the sixth and most complex setup is used.

C. Force/Velocity Prediction for Movement Estimation

The prediction of the human-robot pair's velocity can be integrated to provide an estimation of the trajectory both agents will follow. Table IV shows the average error made by the various evaluated models when comparing this estimation with the actual trajectory followed by the pair. It also includes, for illustration, the error that would be made by an estimation obtained by projecting a second-order polynomial fitted to the last 2 s. As can be seen, the *RetNet-3D-CA* model surpasses the previous state of the art. In turn, this model is outperformed by both *RetNet-3D-CA-SFM* and *RetNet-3D-CA-OM*, with the latter achieving the lowest error.

Although the reduction in error with the *RetNet-3D-CA-OM* model compared to *RetNet-3D* is approximately 10%, observing the practical implications of this reduction shows just how significant it can be. Fig.5 shows two rare cases associated with the particularities of each user. Fig.5-A presents a situation where the human makes an extra effort to avoid the obstacle on the right, as it is the one closest to them. The *RetNet-3D* model does not account for the evolution of the environment, thus it tends to assume that the human will continue exerting this extra effort. Meanwhile, the *RetNet-3D-CA-OM* model predicts that once the obstacle is passed, the human will cease to apply this extra force, thus providing a more accurate prediction. Fig.5-B depicts a situation in which the human temporarily applies less force than usual. The *RetNet-3D* model assumes that the human will continue to exert reduced force, while the *RetNet-3D-CA-OM* model takes into account that when the pair approaches the next obstacle, the human will necessarily need to exert more force in order to decide which direction to turn.

TABLE IV

COMPARISON OF MEAN ERROR ESTIMATING HUMAN-ROBOT PAIR FUTURE TRAJECTORY WITH DIFFERENT MODELS.

Model	L2 [m]		
	500 ms	1000 ms	2000 ms
2nd order polynomial	0.123	0.278	0.543
CNN+LSTM [21]	0.096	0.203	0.425
ViViT+T [22]	0.063	0.143	0.312
RetNet-3D [23]	0.061	0.138	0.297
RetNet-3D-CA	0.058	0.129	0.280
RetNet-3D-CA-SFM	**0.057**	0.126	0.271
RetNet-3D-CA-OM	**0.057**	**0.125**	**0.269**

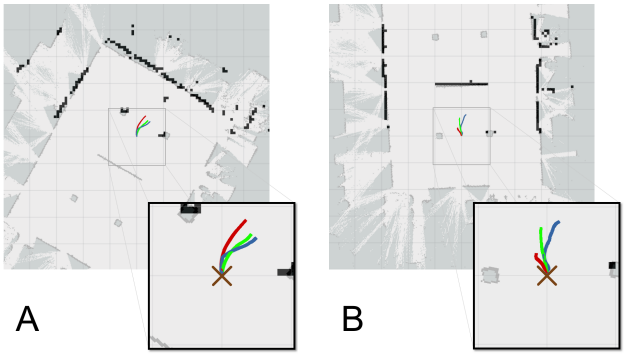

Fig. 5. **Human-robot pair's real and estimated trajectory for the next 2 s in different situations.** Actual trajectory in blue. Trajectory estimation from *RetNet-3D* in red and trajectory estimation from *RetNet-3D-CA-OM* in green. Occupied cells (obstacles) marked in black. Possition of the human-robot centroid marked with a brown cross. *A* - Human making an extra effort to avoid the obstacle in the right. *B* - Human exerting minimal force.

V. Conclusions

In this work, the task of human-robot collaborative transportation has been addressed, focusing on the prediction of two critical variables: the force that the human will exert and the velocity profile that the pair will follow during the next 2 s, thus doubling the temporal horizon of the best models present in the state of the art. To achieve this, two fundamental improvements have been introduced. The first, and lesser, improvement enhances the processing of input information, allowing for better exploitation of the spatial relationships between input data. The second improvement is based on generating a prediction of the environment's evolution, which in turn conditions the generation of the other two predictions.

In this way, the best model obtained achieves acceptable error rates 90.4% of the time when predicting the force that the human will exert within 2 s, and over 93.0% when predicting the velocity of the pair during that period, surpassing the state of the art. These improvements result in a 9.5% reduction in the mean error when attempting to estimate the trajectory followed by the pair, thereby improving the prediction of even less frequent cases.

References

[1] Y. Cheng and M. Tomizuka, "Long-term trajectory prediction of the human hand and duration estimation of the human action," *IEEE Robotics and Automation Letters*, vol. 7, no. 1, pp. 247–254, 2021.

[2] Z. Zhang, G. Peng, W. Wang, Y. Chen, Y. Jia, and S. Liu, "Prediction-based human-robot collaboration in assembly tasks using a learning from demonstration model," *Sensors*, vol. 22, no. 11, p. 4279, 2022.

[3] H. Duan, Y. Yang, D. Li, and P. Wang, "Human–robot object handover: Recent progress and future direction," *Biomimetic Intelligence and Robotics*, vol. 4, no. 1, p. 100145, 2024.

[4] Z. Wang, Z. Liu, N. Ouporov, and S. Song, "Contacthandover: Contact-guided robot-to-human object handover," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 9916–9923.

[5] A. Fusco, V. Modugno, D. Kanoulas, A. Rizzo, and M. Cognetti, "Transformer-Based Prediction of Human Motions and Contact Forces for Physical Human-Robot Interaction," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 3161–3167.

[6] J. Park, Y.-S. Shin, and S. Kim, "Object-aware impedance control for human–robot collaborative task with online object parameter estimation," *IEEE Transactions on Automation Science and Engineering*, 2024.

[7] J. E. Domínguez-Vidal and A. Sanfeliu, "Anticipation and Proactivity. Unraveling Both Concepts in Human-Robot Interaction through a Handover Example," in *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2024, pp. 957–962.

[8] A. J. Sathyamoorthy, J. Liang, U. Patel, T. Guan, R. Chandra, and D. Manocha, "Densecavoid: Real-time navigation in dense crowds using anticipatory behaviors," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 11 345–11 352.

[9] M. Cramer, K. Kellens, and E. Demeester, "Probabilistic decision model for adaptive task planning in human-robot collaborative assembly based on designer and operator intents," *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 7325–7332, 2021.

[10] J. E. Domínguez-Vidal and A. Sanfeliu, "When the inference meets the explicitness or why multimodality can make us forget about the perfect predictor," *International Journal of Social Robotics*, vol. 17, no. 12, pp. 2965–2980, 2025.

[11] A. Bussy, P. Gergondet, A. Kheddar, F. Keith, and A. Crosnier, "Proactive behavior of a humanoid robot in a haptic transportation task with a human partner," in *2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 2012, pp. 962–967.

[12] S. Tarbouriech, B. Navarro, P. Fraisse, A. Crosnier, A. Cherubini, and D. Sallé, "Admittance control for collaborative dual-arm manipulation," in *2019 19th International Conference on Advanced Robotics (ICAR)*. IEEE, 2019, pp. 198–204.

[13] D. J. Agravante, A. Cherubini, A. Bussy, P. Gergondet, and A. Kheddar, "Collaborative human-humanoid carrying using vision and haptic sensing," in *2014 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2014, pp. 607–612.

[14] X. Yu, B. Li, W. He, Y. Feng, L. Cheng, and C. Silvestre, "Adaptive-constrained impedance control for human–robot co-transportation," *IEEE transactions on cybernetics*, vol. 52, no. 12, pp. 13 237–13 249, 2021.

[15] X. Yu, W. He, Y. Li, C. Xue, J. Li, J. Zou, and C. Yang, "Bayesian estimation of human impedance and motion intention for human–robot collaboration," *IEEE transactions on cybernetics*, vol. 51, no. 4, pp. 1822–1834, 2019.

[16] C. N. Mavridis, K. Alevizos, C. P. Bechlioulis, and K. J. Kyriakopoulos, "Human-robot collaboration based on robust motion intention estimation with prescribed performance," in *2018 European Control Conference (ECC)*. IEEE, 2018, pp. 249–254.

[17] L. Roveda, J. Maskani, P. Franceschi, A. Abdi, F. Braghin, L. Molinari Tosatti, and N. Pedrocchi, "Model-Based Reinforcement Learning Variable Impedance Control for Human-Robot Collaboration," *Journal of Intelligent & Robotic Systems*, vol. 100, no. 2, pp. 417–433, 2020.

[18] E. Gribovskaya, A. Kheddar, and A. Billard, "Motion learning and adaptive impedance for robot control during physical interaction with humans," in *2011 IEEE International Conference on Robotics and Automation*. IEEE, 2011, pp. 4326–4332.

[19] A. Al-Yacoub, Y. Zhao, W. Eaton, Y. M. Goh, and N. Lohse, "Improving human robot collaboration through force/torque based learning for object manipulation," *Robotics and Computer-Integrated Manufacturing*, vol. 69, p. 102111, 2021.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[21] J. E. Domínguez-Vidal and A. Sanfeliu, "Improving Human-Robot Interaction Effectiveness in Human-Robot Collaborative Object Transportation using Force Prediction," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 7839–7845.

[22] J. E. Domínguez-Vidal and A. Sanfeliu, "Exploring Transformers and Visual Transformers for Force Prediction in Human-Robot Collaborative Transportation Tasks," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 3191–3197.

[23] J. E. Domínguez-Vidal and A. Sanfeliu, "Force and velocity prediction in human-robot collaborative transportation tasks through video retentive networks," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 9307–9313.

[24] Y. Sun, L. Dong, S. Huang, S. Ma, Y. Xia, J. Xue, J. Wang, and F. Wei, "Retentive Network: A Successor to Transformer for Large Language Models," 2023.

[25] H. Yang, Z. Li, Y. Chang, and Y. Wu, "A survey of retentive network," *arXiv preprint arXiv:2506.06708*, 2025.

[26] H. Lin, X. Cheng, X. Wu, and D. Shen, "CAT: Cross attention in vision transformer," in *2022 IEEE international conference on multimedia and expo (ICME)*. IEEE, 2022, pp. 1–6.

[27] C.-F. R. Chen, Q. Fan, and R. Panda, "Crossvit: Cross-attention multi-scale vision transformer for image classification," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 357–366.

[28] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "Film: Visual reasoning with a general conditioning layer," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[29] J. E. Domínguez-Vidal and A. Sanfeliu, "The human intention: a taxonomy attempt and its applications to robotics," *International Journal of Social Robotics*, vol. 17, no. 11, pp. 2479–2499, 2025.

[30] D. Helbing and P. Molnar, "Social Force Model for Pedestrian Dynamics," *Physical review E*, vol. 51, no. 5, p. 4282, 1995.

[31] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[32] L. Deininger, B. Stimpel, A. Yuce, S. Abbasi-Sureshjani, S. Schönenberger, P. Ocampo, K. Korski, and F. Gaire, "A comparative study between vision transformers and cnns in digital pathology," *arXiv preprint arXiv:2206.00389*, 2022.