Human Trajectory Forecasting in 3D Environments: Navigating Complexity under Low Vision

Franz Franco Gallo, Hui-Yin Wu, Lucile Sassatelli

Human Trajectory Forecasting in 3D Environments: Navigating Complexity under Low Vision

Franz Franco Gallo*
franz.franco-gallo@inria.fr
Université Côte d'Azur, Inria
Sophia-Antipolis, France

Hui-Yin Wu
Université Côte d'Azur, Inria
Sophia-Antipolis, France
hui-yin.wu@inria.fr

Lucile Sassatelli
Université Côte d'Azur, CNRS, I3S.
Institut Universitaire de France
Sophia-Antipolis, France
lucile.sassatelli@univ-cotedazur.fr

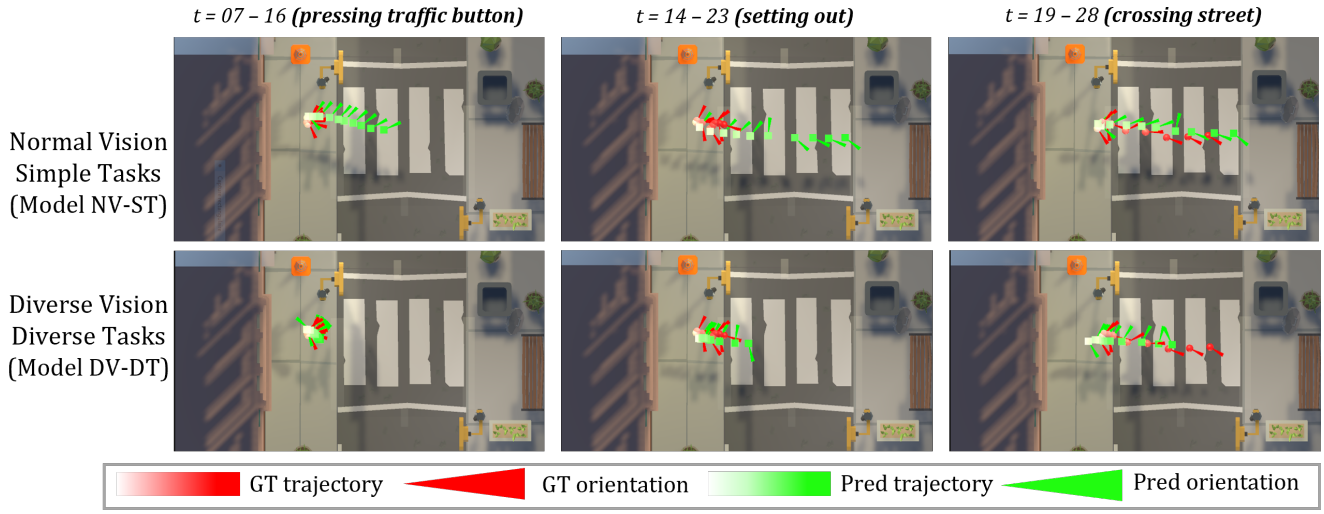

Figure 1: Predicted trajectory and orientation of two GIMO model training variations evaluated in a low-vision, complex task scene. The diverse model takes into account complex interactions (pressing traffic button), and adapts to user walking speed.

ABSTRACT

This work tackles the challenge of predicting human trajectories while carrying out complex tasks in contextually-rich virtual environments. We evaluate the CREATIVE3D multimodal dataset on human interaction and navigation in 3D virtual reality (VR). In the dataset, navigating traffic crossings with simulated visual impairments are used as an example of complex or unpredictable situations. We establish evaluations for a base multi-layer perceptron (MLP) and two state-of-the-art models: TRACK (RNN) and GIMO (transformer), on tasks with varying levels of complexity and visual impairment conditions. Our findings indicate that a model trained on normal visual conditions and simple tasks does not generalize on test data with complex interactions and simulated visual impairments, despite including 3D scene context and user gaze. In comparison, a model trained on diverse visual and task conditions is more robust, with up to 84% decrease in positional error and 9% in orientation error, but with the trade-off of lower accuracy for simpler tasks. We believe this work can benefit real-world applications such as autonomous driving, and enable context-aware computing for diverse scenarios and populations.

CCS CONCEPTS

• **Computing methodologies** → **Neural networks**; **Motion capture**; **Activity recognition and understanding**.

KEYWORDS

human motion prediction, virtual reality, context, RNN, transformer

ACM Reference Format:
Franz Franco Gallo, Hui-Yin Wu, and Lucile Sassatelli. 2024. Human Trajectory Forecasting in 3D Environments: Navigating Complexity under Low Vision. In *16th International workshop on Immersive Mixed and Virtual Environment Systems (MMVE '24), April 15–18, 2024, Bari, Italy.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3652212.3652223

1 INTRODUCTION

Human trajectory forecasting aims to predict future human movements and is of strong interest especially for high-stake scenarios such as pedestrian behavior prediction and understanding in self-driving applications [5]. The complexity of pedestrian behavior,

MMVE '24, April 15–18, 2024, Bari, Italy

ACM ISBN 979-8-4007-0618-9/24/04
https://doi.org/10.1145/3652212.3652223

characterized by their individual intentions, interactions with other pedestrians, vehicles and the environment, present a significant challenge for autonomous systems [16]. Acknowledging these complexities, our work focuses on ensuring that models for autonomous vehicles consider the diverse behaviors of pedestrians, including those with vision inequalities and their interactions with traffic lights, to enhance prediction accuracy and fairness.

Multiple approaches to human motion prediction have been proposed using machine learning and deep learning techniques, with an unsolved challenge of efficiently taking into account scene and social context [1]. In this work, we are the first to approach model performance for pedestrian trajectory and attention prediction under the light of fairness for the visually impaired. We do so thanks to virtual reality (VR) technologies where realistic scenarios can be simulated to investigate human behaviour in-context. One such dataset is the CREATTIVE3D multimodal dataset of user behavior in VR [15][1] which collects user behaviours in complex tasks such as seeking, taking, and transporting objects, and real walking in simulated road intersections [12], including conditions with simulated visual impairments – a virtual scotoma (area in the central visual field with little or no acuity).

We present evaluations of motion prediction models on this newly introduced dataset to identify key weaknesses of existing reference prediction models and research challenges ahead. Specifically, our contributions are:

- We identify how brittle the models are in case of distribution shifts, that is (1) when training on normal-vision data and predicting on low-vision data (up to 90% and 8% error increase in position and attention prediction, respectively), (2) when training on simple tasks and predicting on complex tasks (up to 900% and 44% error increase), and
- We show how a diverse training set with different types of vision conditions and tasks can alleviate the performance unfairness. We reveal that the models exhibit performance trade-offs between the different populations and scenarios when trained on a diverse and balanced dataset (e.g., error increase of up to 25% and 10% for position and attention of normal vision data, 72.3% and 1% for simple tasks), hence exhibiting their inability to properly condition the output based on context. We propose future important research avenues based on the findings.

We first present the related work in Sec. 2. We then introduce our models and testing conditions in Sec. 3 and present the results in Section 4. Finally, we provide a discussion in Sec. 5 and conclusions with the future research avenues Sec. 6.

2 RELATED WORK

Predicting human motion or attention trajectory from multiple modalities has been a long-standing endeavor in various application scenarios (pedestrian trajectory forecast for self-driving, optimization of VR rendering, etc.). We briefly discuss and position our approach within the existing prediction models, and the available datasets with varied contextual conditions (type of vision, type of tasks, physical environment) and representation (unstructured such as point cloud, or structured such as scene graphs).

[1]https://zenodo.org/doi/10.5281/zenodo.8269108

2.1 Models for human motion prediction

The prediction of human motion is adequately approached as a sequence-to-sequence problem, with prior movements providing the basis for forecasting subsequent sequences, and possibly informed by the context. Current models employ a variety of architectures, notably Recurrent Neural Networks (RNN), Graph Convolutional Networks, Generative Adversial Networks , and Transformers. An example of a multimodal RNN-based prediction model is TRACK, which predicts attention in 3 Degrees of Freedom (DoF) VR [13]. Leveraging correlations within a single modality and across several modalities has known substantial progress with Transformers [14], fueling so-called cross-modal Foundation Models that are pre-trained on large-scale datasets [9, 11]. An example of human motion prediction using attention mechanisms to exploit spatial and temporal correlation between joints is STTran[2].

However, transformers are plagued with quadratic complexity in the size of the input, often high-dimensional for images, videos and text, incurring heavy computational costs both in train and test. Several approaches aim to counteract the high complexity, amongst which the family of Perceiver models [8], avoiding computationally-heavy self-attention on a high-dimensional input. Recently, Zheng et al. introduced a Perceiver-based architecture for motion prediction in 6 DoF VR, named GIMO. The GIMO model [17] exploits gaze data from an eponym dataset to improve human motion prediction (center of gravity displacements and positions of joints).

In the present work, we consider TRACK and GIMO as reference representatives of RNN-based and Transformer-based models for motion prediction in VR. To our knowledge, no existing prediction model has neither considered the impact of low vision on prediction accuracy, nor that of different tasks.

2.2 Datasets for human motion prediction

The context in which actions are carried out by people, including the vision condition, environment and tasks, can be represented in three forms: images (2D), point clouds (3D), and scene graphs. The endeavor to accurately model human motion is extensively pursued through the utilization of high-caliber motion capture datasets. These range from the more compact CMU Graphics Lab motion capture database[4] to large collections like AMASS [10] and Human3.6M dataset [7]. The latter is distinguished by its high-quality motion capture with a multi-view camera system, establishing itself as a benchmark for motion prediction and 3D pose estimation. For rich contexts, datasets such as GIMO [17] and CIRCLE [3] have emerged taking advantage of virtual and augmented reality technologies, concentrating on simple tasks like reaching for an object or navigating to a location.

Nevertheless, these datasets do not portray realistic interactions with the environment that are often chained and overlapping. The recent CREATTIVE3D dataset [15] addresses this gap having key interesting features to address our objective. Indeed, it is the largest dataset of human motion in context (over 2.6 million poses), it is captured in fully annotated and dynamic 3D scenes with multivariate – gaze, physiology, and motion – data, and it investigates the impact of simulated low-vision conditions using dynamic eye tracking under real walking and simulated walking conditions. It therefore allows the analysis of predicted pedestrian behavior disaggregated

over simple and complex tasks, such as interacting with the traffic light before crossing, as depicted in Fig. 1, and over normal and simulated low-vision conditions. It also provides point clouds of the environments, which can be processed as input and incorporated into existing models such as TRACK and GIMO.

3 METHODS

We investigate how predictive models trained on normal-vision and simple navigation tasks perform on simulated low vision and higher task complexity at inference time. We introduce the dataset for this analysis and the models chosen for benchmarking.

3.1 Problem Definition

We consider predicting the future trajectory of a human, modeled by the head position and orientation in 3D space from past position and possibly context (depending on the models). The human model comprises of, at any given time t (in frames), the head **position** $\mathbf{p}_t \in \mathbb{R}^3$, each component in meters, and head **orientation** $\mathbf{r}_t \in \mathbb{R}^3$ in Euler angles (roll, pitch, yaw). Head position represents the user's absolute position in the scene where they walk physically with a 1:1 ratio between the real and virtual distance in the 10 by 4 meters tracked space. Head orientation corresponds to the direction of the center of the headset field of view.

The problem consists in predicting a full motion sequence over a future horizon H, represented as $\mathbf{M}_{t+1:t+H} = \{(\hat{\mathbf{p}}_{t+1}, \hat{\mathbf{r}}_{t+1}), \ldots, (\hat{\mathbf{p}}_{t+H}, \hat{\mathbf{r}}_{t+H})\}$ from a given time t. We employ a sampling rate of 2 fps, utilize 3 seconds of past motion and gaze data for input, and aim to forecast motion for the subsequent 5 seconds. Specifically, for any time-stamp t, our prediction spans $\{\mathbf{M}_{t+s}\}_{s=1}^{H}$ for each time-step s across a horizon of $H = 10$. The model accounts for a past motion history $\mathbf{M}_{t-L+1:t}$ with $L = 6$.

3.2 The CREATTIVE3D multimodal dataset

We take advantage of our newly released CREATTIVE3D dataset [15] of human interactions and navigation in VR, specifically scenes of road crossings. The CREATTIVE3D dataset includes an extensive collection of simulated pedestrian behaviors, designed to capture a wide range of human activities, from basic motion to complex interactions with objects and urban infrastructure. Its richness lies in the multimodal data collected allowing for an in-depth analysis of pedestrian dynamics under varying conditions.

This dataset stands out due to its comprehensive multivariate data including gaze, physiology, and motion in fully-annotated dynamic 3D scenes. It explores the impact of simulated low-vision conditions, incorporating real-time eye tracking to simulate visual impairments. The dataset supports a broad spectrum of research, from cognitive studies to computational modeling for understanding human behavior in VR. The dataset includes 6 scenarios of two task complexities: simple tasks (ST) with only navigation, and complex tasks (CT) with simultaneous navigation and interaction. An example of a complex tasks consisting of interacting with the traffic light then crossing is shown in Fig. 1. Each scenario is further observed under two visual conditions: Normal Vision (NV) and simulated Low Vision (LV).

Training. We consider 4 types of models: trained on normal vision and simple tasks, as well as a combinations of low vision or complex task. Specifically, we designed training and validation datasets that (1) for the scenario, comprise of either simple tasks only (ST) or diverse simple and complex tasks (DT), and (2) either normal vision only (NV) or diverse normal and low vision (DV). The resulting four training and validation sets, along with the number of samples per scenario/visual condition are summarized in Table 1. Note that each sample across all models is unique to ensure a diverse and comprehensive dataset for model training and validation.

Table 1: Summary of models: training and validation sets

Model	Training Samples	Validation Samples
NV-ST	NV-ST: 251	NV-ST: 63
NV-DT	NV-ST: 139 NV-CT: 139	NV-ST: 35 NV-CT: 35
DV-ST	NV-ST: 139 LV-ST: 138	NV-ST: 35 LV-ST: 35
DV-DT	NV-ST: 139 NV-CT: 139 LV-CT: 138	NV-ST: 35 NV-CT: 35 LV-CT: 35

Test. To investigate the robustness of reference models to distribution shifts over vision conditions and task complexities, we consider 4 test sets (with number of samples) to assess their performance across the spectrum of tasks and visual conditions: Test NV-ST (78), Test NV-CT (44), Test LV-ST (42), Test LV-CT (43)

3.3 Prediction models

We evaluate three reference models for human motion prediction: MLP, TRACK, and GIMO, as depicted in Fig. 2 on their accuracy and robustness across various vision conditions and task complexities. The models take as input different feature vectors processed from scene point cloud, gaze point cloud, and pose data. As shown in top of Figure 2 using PointNet++ for feature extraction, we obtain a per-point feature map (F_P) and a global scene descriptor (F_O).

MLP baseline model [6]. includes fully connected layers, transpose operations, and layer normalization to merge information across frames effectively. Each MLP block has a fully connected layer and layer normalization, iteratively applied to capture the temporal dynamics in the motion sequence, as shown in Figure 2-(a). Our adaptation uses 4 MLP blocks, leveraging its ability to effectively model temporal dependencies for improved accuracy.

TRACK [13]. based on RNN, a sequence-to-sequence model where the past ego motion and scene features are each processed by individual LSTMs, before being fused by a third LSTM. As shown in Figure 2-(b), the scene context is the gaze-interpolated feature f_g. Given the per-point feature F_P, the gaze point feature f_g is computed through inverse distance-weighted interpolation [17], this interpolated gaze feature thus encapsulates relevant scene information, offering clues to deduce the subject's intention.

GIMO model [17]. a tranformer model composed of three cross-attention modules, where self-attention is first applied to the key-value modality, followed by cross-attention with the query modality

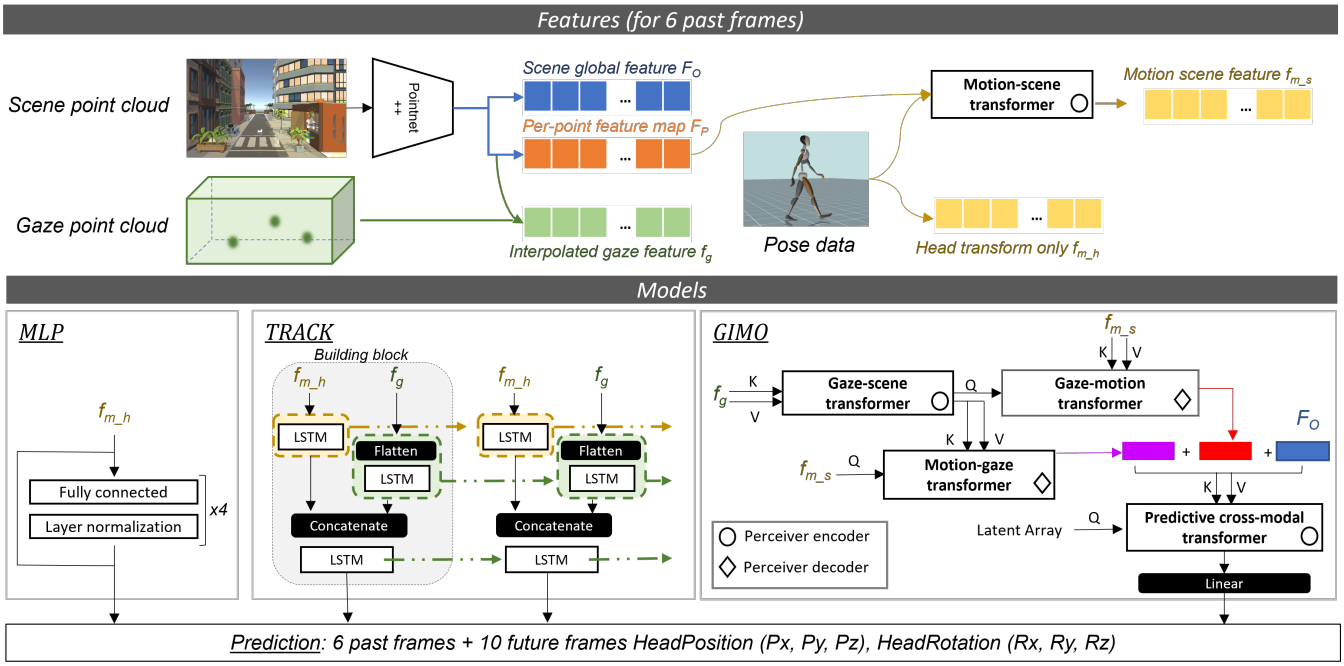

Figure 2: Our workflow takes into account scene, gaze, and human motion data, building different feature vectors. We evaluate the dataset on three models: a baseline MLP, TRACK (LSTM) and GIMO (transformer).

(Figure 2-(c)). The modalities attending to each other are the position, the scene context around the body, and the scene context around the gaze target. All three latent vectors are then combined in a last cross-modal transformer, producing estimates of the future positions and orientations over a prediction horizon of 5 seconds with a history length of 3 seconds. The hyper-parameters are kept as in the original GIMO article.

Each of the three architectures is trained on the four training and validation sets detailed in the previous section. Each of the three architectures undergoes training on the four training and validation sets outlined in the preceding section. Each of the three architectures is subjected to learning processes on the four training and validation sets outlined in the preceding section.

3.4 Evaluation Metrics

For assessing prediction accuracy on position and head orientation, we use the following metrics:

Position error. We measure prediction error on position with the Mean Squared Error (MSE). MSE calculates the average distance, in square meters, between ground truth and predicted trajectory position across all time steps in the future horizon H (5 seconds).

Orientation error. The error prediction on head orientation is measured with the Orthodromic Distance (OD). OD quantifies the average angular distance, in radians, between ground truth and predicted orientations across all time steps in the future horizon H. The OD is defined as:

$$OD = \frac{1}{H}\sum_{i=1}^{H} 2\arccos(|\mathbf{r}_{t+i} \cdot \hat{\mathbf{r}}_{t+i}|) \quad (1)$$

where $H = 10$ is the total number of predictions, $\mathbf{r}_{t+i}$ and $\hat{\mathbf{r}}_{t+i}$ are the unit quaternions representing the ground truth and predicted orientations for the t+ith prediction.

4 EXPERIMENTAL EVALUATION

In this section, we address the following research questions:

- RQ1 How do the models compare to each other in different train-test configurations, and can we identify a superior model?
- RQ2 To what extent can models trained on normal vision tasks maintain accuracy in low vision scenarios? Does refining the training dataset to reflect low vision test conditions optimize predictions, and what inherent model limitations does this approach reveal?
- RQ3 How well do models designed for tasks under normal vision adapt to more complex challenges? Is the accuracy of predictions enhanced by aligning the training data with the complexities of the test environment, or does this strategy expose fundamental flaws in the models?

4.1 Global analysis

Table 2 shows the median values of MSE and OD for the position and orientation predictions respectively. GIMO has the lowest MSE values in all tests except LV-ST. GIMO's architecture under simple tasks (NV-ST, LV-ST) has the lowest OD, with relatively consistent

performance across different conditions. TRACK and MLP architectures on the other hand, seem sensitive to test conditions, as evidenced by fluctuating OD values.

Table 2: MSE and OD values for position and orientation on (1) three architectures (MLP, TRACK and GIMO), (2) four model variations (Table 1), and (3) on the four test sets.

	Tests	NV-ST		NV-CT		LV-ST		LV-CT	
Arch	Model	MSE	OD	MSE	OD	MSE	OD	MSE	OD
MLP	NV-ST	0.141	0.838	0.854	0.889	0.160	0.725	0.761	0.936
	NV-DT	0.718	0.805	0.267	0.877	0.285	0.704	0.214	0.819
	DV-ST	0.152	0.862	0.668	0.901	**0.142**	0.774	0.607	0.939
	DV-DT	0.744	0.797	0.261	0.854	0.520	0.691	0.252	0.833
TRACK	NV-ST	0.088	0.864	0.845	0.913	0.174	0.790	0.767	1.066
	NV-DT	0.284	0.669	0.250	0.697	0.260	0.709	0.201	0.777
	DV-ST	0.136	0.929	0.527	0.861	0.163	0.832	0.598	0.917
	DV-DT	0.260	0.631	0.193	0.685	0.184	0.616	0.188	0.732
GIMO	NV-ST	**0.083**	**0.557**	0.829	0.804	0.157	0.597	0.832	0.714
	NV-DT	0.143	0.562	**0.167**	0.688	0.225	**0.590**	0.186	0.679
	DV-ST	0.104	0.618	0.855	0.806	0.144	0.646	0.879	0.762
	DV-DT	0.223	0.649	**0.167**	**0.640**	0.180	0.646	**0.137**	**0.647**

Answer to RQ1: Under the NV-ST test for simple tasks, the MLP model displays moderate to high MSE values, achieving its best performance at $0.141m^2$ for the NV-ST model. TRACK demonstrates an improvement over MLP, while GIMO surpasses both MLP and TRACK in the NV-ST configuration by recording the lowest MSE of $0.083m^2$. Upon including low vision and complex tasks into the test, the performances of MLP, TRACK, and GIMO vary, with each showing their best results in DV-DT model. GIMO's architecture outperforms with both NV-ST and DV-DT models, offering the most accurate predictions. GIMO consistently exhibits lower OD values, which confirms it as the superior model across all tests.

4.2 GIMO analysis

We conduct a detailed examination of the GIMO architecture's performance. The whisker plots in Figure 3 show the MSE and OD distribution along the prediction horizon and the median estimation using the sliding window along the task sequence length, computed between the ground truth and predicted future motion across the four different training and validation sets for GIMO.

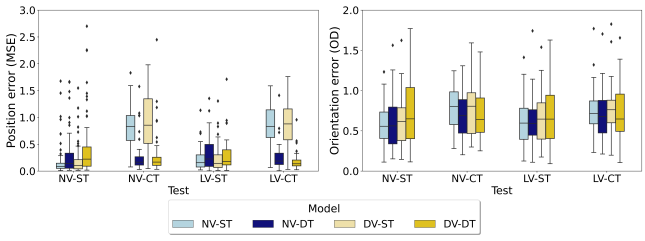

Figure 3: Comparative Analysis of Position and Orientation Errors in the GIMO Architecture: (Left) MSE Box plots for position estimation and (Right) OD for orientation estimation.

Model NV-ST stands out for its consistency in NV-ST test as depicted with Figure 3 (left), demonstrated by tight interquartile range (IQRs) and low median MSE value. However, wider IQRs in the NV-CT and LV-CT tests indicate significant prediction errors under complex conditions. Model NV-DT shows wider IQRs in NV-ST and LV-ST, reflecting greater MSE variability for simple tasks than NV-ST, however reducing the IQRs in NV-CT and LV-CT. Model DV-ST maintains narrow IQRs in NV-ST and LV-ST tests, indicating stable performance, but struggles with increased variability in NV-CT and LV-CT. Model DV-DT exhibits similar trends, with variable median MSE values and considerable outliers in NV-ST, NV-CT, and LV-CT, underscoring challenges in complex and low vision conditions. Overall, while NV-DT and DV-DT models offer accuracy and consistency, NV-ST and DV-ST highlight increased variability and occasional large errors, especially in complex task scenarios.

Across all models, the transition from simple to complex tasks tends to result in a slight increase in orientation error under both normal and low vision conditions. The variability of OD, as shown by the IQR and outliers in Figure 3 (right), does not change drastically, which could mean that the models maintain a similar level of consistency in orientation prediction despite task complexity. In the following sections we extend our evaluation to focus on the impact of vision conditions and task complexity using Table 2.

4.2.1 Impact of vision condition. Comparing the NV-ST model's performance across tests sets reveals a significant shift when going from normal to low vision: position MSE increases by 89.16% ($+0.074\,m^2$) and orientation OD by 7.18%. Training on diverse vision (DV-ST) introduces a 8.2% ($-0.013\,m^2$) decrease in position MSE over the base model (NV-ST), but an increase (also 8.2%) in orientation OD. Meanwhile, The MSE and OD for the NV-ST test condition also increase by 25.30% ($+0.021\,m^2$) and 10.95% respectively, further reinforcing the notion of a trade-off in model performance.

Answer to RQ2: The Model NV-ST trained on normal vision and simple tasks exhibit poor robustness when predicting on low vision with increase in position (MSE) and orientation (OD) error. Modifying the training set to include diverse vision conditions (model DV-ST) improves the position error but worsens the orientation error, and even more increases the position and orientation error for the simple task test NV-ST. While the model becomes more adaptable to low-vision trajectories, its performance slightly degrades in normal-vision conditions.

4.2.2 Impact of task complexity. Comparing the NV-ST model performance on various test sets, we notice a 898.80% ($+0.746\,m^2$) increase in position MSE and 44.34% increase in orientation OD when going from simple to complex tasks. The substantial increase in both MSE and OD under the complex task condition reflects that task complexity has a more profound impact on the model's performance than changes in vision conditions.

In contrast, the model trained on diverse tasks (NV-DT) outperforms the NV-ST model on complex tasks, with 79.98% ($-0.662\,m^2$) decrease in position MSE, and 14.43% decrease in OD. However, this is at the expense of accuracy for NV-ST test, with an 72.29% ($+0.06\,m^2$) increase for position MSE. The orientation OD is less

impacted, only resulting in a 0.9% increase. This reflects that training on diverse tasks (NV-DT) improves the model's ability to tackle complex challenges, enhancing both positional accuracy and orientation precision. However, this focused improvement on complex tasks can potentially lead to a reduction in performance on simpler, baseline tasks (NV-ST).

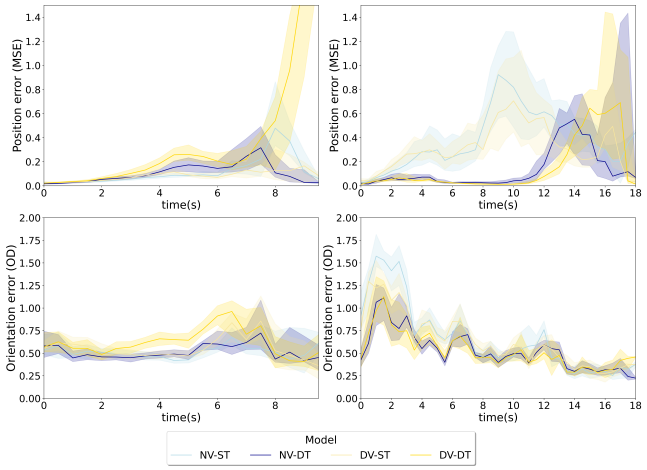

Figure 4: Disaggregated results for MSE and OD along the task duration, under NV-ST (left) and LV-CT (right) test conditions.

Finally, if we evaluate the performance of the DV-DT model against the baseline model NV-ST across the four different test conditions, we observe that the DV-DT model shows a remarkable improvement in handling position prediction in complex tasks and scenarios involving low vision (-83.53% on position MSE $-0.695\,m^2$), with also a moderate improvement to orientation OD (-9.38%). However, when evaluated under low vision conditions with simple tasks (LV-ST test), the model's performance slightly deteriorates. Significant concern arises from the model's performance in standard test conditions (NV-ST test), where the MSE position error increases by 168.67% ($+0.14\,m^2$), with also a notable increase in OD orientation error (+16.52%).

Answer to RQ3: The Model NV-ST trained on normal vision and simple tasks exhibit poor robustness when predicting on complex tasks. Modifying the training set to include diverse tasks, generally improves model performance in those specific conditions. However, this focused improvement comes with compromises, on the baseline tasks (NV-ST). The quantified data reveal that training on a broad spectrum of conditions significantly improves performance, evidenced by up to an 89% reduction in positional error and 20% in orientation error, but introduces a trade-off, resulting in reduced accuracy for simpler baseline tasks.

The disaggregated plots in Figure 4 shows MSE and OD across models under NV-ST (right) and LV-CT (left) test. The model NV-ST demonstrates impressive accuracy, contrasting with the model DV-DT, where a pronounced increase in MSE is observed towards the task's end. In the LV-CT test, models NV-ST and DV-ST exhibit heightened MSE at the task's onset due to their training void of complex tasks, whereas NV-DT and DV-DT models show initial MSE reductions, only to rise again as tasks progress. This trend is paralleled by escalating OD errors from the outset, particularly when individuals engage with traffic lights, highlighting increased positional uncertainty. The influence of low vision introduces amplified uncertainty in tracking ground truth positions and orientations, notably exacerbating as tasks conclude. Moreover, the onset of complex tasks elevates orientation errors, especially during initial traffic light interactions, leading to escalated positional errors by the task's end.

5 DISCUSSIONS

The experimental evaluation described in Section 4 details the impact of low-vision and task complexity conditions on human motion prediction.

Our findings reveal GIMO as the superior model, taking advantage of the extra contextual information in this model, consistently outperforming MLP and TRACK in prediction for both position and orientation, especially in the NV-ST and DV-DT tests. However, performing a deeper analysis on the models trained with GIMO, reveals issues in the NV-ST model's ability to predict tasks with low vision and complex task conditions. And even if we refine the training data set to include diversity of conditions (NV-DT, DV-ST and DV-DT) marginally improved position and orientation prediction errors, but at the cost of increased prediction errors in tasks with normal conditions.

Our study specifically addresses the challenges faced by individuals with scotoma, a condition resulting in partial vision loss, in the context of human motion forecasting but also highlights the urgent need for a more inclusive approach in subsequent research efforts.

6 CONCLUSIONS

Our study provides a foundational understanding of model performance in predicting human motion in immersive environments under low visual conditions and complex tasks. We found that models trained with a diverse range of task conditions stands out for its robustness in reducing position and orientation prediction errors by 84% and 9%, respectively. Nonetheless, the subtle trade-offs observed across normal vision and simple task conditions highlight the complexity of designing predictive models. Future work could explore more training strategies or architectural improvements to enhance performance under these challenging conditions, a promising direction involves using context annotations from the CREATIVE3D dataset to deepen our understanding of human behavior during task execution, particularly in complex scenarios.

7 ACKNOWLEDGEMENTS

This work has been partially supported by the French National Research Agency through the ANR CREATTIVE3D project ANR-21-CE33-0001 and UCA^{JEDI} Investissements d'Avenir ANR-15-IDEX-01 (IDEX reference center for extended reality XR^{2}C^2). This work was granted access to the HPC resources of IDRIS under the allocation 2024-AD011014115R1 made by GENCI.

REFERENCES

[1] Vida Adeli, Ehsan Adeli, Ian Reid, Juan Carlos Niebles, and Hamid Rezatofighi. 2020. Socially and Contextually Aware Human Motion and Pose Forecasting. *IEEE Robotics and Automation Letters* 5, 4 (2020), 6033–6040. https://doi.org/10.1109/LRA.2020.3010742 Number: 4.

[2] Emre Aksan, Manuel Kaufmann, Peng Cao, and Otmar Hilliges. 2021. A spatio-temporal transformer for 3d human motion prediction. In *2021 International Conference on 3D Vision (3DV)*. IEEE, 565–574.

[3] Joao Pedro Araújo, Jiaman Li, Karthik Vetrivel, Rishi Agarwal, Jiajun Wu, Deepak Gopinath, Alexander William Clegg, and Karen Liu. 2023. CIRCLE: Capture In Rich Contextual Environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 21211–21221.

[4] Fernando De la Torre, Jessica Hodgins, Adam Bargteil, Xavier Martin, Justin Macey, Alex Collado, and Pep Beltran. 2009. Guide to the carnegie mellon university multimodal activity (cmu-mmac) database. (2009).

[5] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. 2013. Vision meets robotics: The kitti dataset. *Int. J. of Robotics Research* 32, 11 (2013), 1231–37.

[6] Wen Guo, Yuming Du, Xi Shen, Vincent Lepetit, Xavier Alameda-Pineda, and Francesc Moreno-Noguer. 2023. Back to mlp: A simple baseline for human motion prediction. In *Proc. IEEE Winter Conf. on Appl. of Computer Vision*. 4809–19.

[7] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. 2013. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence* 36, 7 (2013), 1325–1339.

[8] Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, et al. 2021. Perceiver IO: A General Architecture for Structured Inputs & Outputs. In *International Conference on Learning Representations*.

[9] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*. PMLR, 4904–4916.

[10] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. 2019. AMASS: Archive of motion capture as surface shapes. In *Proc. IEEE/CVF international conference on computer vision*. 5442–5451.

[11] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

[12] Florent Robert, Hui-Yin Wu, Lucile Sassatelli, Stephen Ramanoël, Auriane Gros, and Marco Winckler. 2023. An Integrated Framework for Understanding Multimodal Embodied Experiences in Interactive Virtual Reality. In *Proc. 2023 ACM International Conference on Interactive Media Experiences* (Nantes, France). Association for Computing Machinery, New York, NY, USA, 14–26.

[13] Miguel Fabián Romero Rondón, Lucile Sassatelli, Ramón Aparicio-Pardo, and Frédéric Precioso. 2021. TRACK: A New Method From a Re-Examination of Deep Architectures for Head Motion Prediction in 360^o Videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 9 (2021), 5681–5699.

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[15] Hui-Yin Wu, Florent Alain Sauveur Robert, Franz Franco Gallo, Kateryna Pirkovets, Clément Quere, Johanna Delachambre, Stephen Ramanoël, Auriane Gros, Marco Winckler, Lucile Sassatelli, Meggy Hayotte, Aline Menin, and Pierre Kornprobst. 2023. Exploring, walking, and interacting in virtual reality with simulated low vision: a living contextual dataset. (2023). https://inria.hal.science/hal-04429351 preprint.

[16] Chi Zhang and Christian Berger. 2023. Pedestrian Behavior Prediction Using Deep Learning Methods for Urban Scenarios: A Review. *IEEE Transactions on Intelligent Transportation Systems* 24, 10 (2023), 10279–10301. https://doi.org/10.1109/TITS.2023.3281393

[17] Yang Zheng, Yanchao Yang, Kaichun Mo, Jiaman Li, Tao Yu, Yebin Liu, Karen Liu, and Leonidas J Guibas. 2022. GIMO: Gaze-Informed Human Motion Prediction in Context. *arXiv preprint arXiv:2204.09443* (2022). arXiv:2204.09443

Received 07 February 2024; revised 08 March 2024; accepted xx March 2024