A Bi-Directional Deep Learning Interface for Gaze-Controlled Wheelchair Navigation: Overcoming the Midas Touch Problem

Gianni Bremer*
Institute for Psychology
University of Muenster,
Muenster, Germany

Joseph McIntyre
Health Unit, TECNALIA,
Basque Research and Technology Alliance,
San Sebastian, Spain

Je Hyung Jung
Health Unit, TECNALIA,
Basque Research and Technology Alliance,
San Sebastian, Spain

Stefano Ellero
Stam S.r.l., Genova, Italy

Issa Mouawad
Stam S.r.l., Genova, Italy

Davide Di Gloria
Stam S.r.l., Genova, Italy

Markus Lappe
Institute for Psychology
University of Muenster,
Muenster, Germany

ABSTRACT

We present a gaze-based augmented reality control interface for electric wheelchairs, addressing the challenges faced by individuals with mobility impairments. The development transitions through three stages: model training with offline evaluation, Virtual Reality (VR) simulations, and physical deployment.

First, we trained deep learning models, comparing Transformers and LSTMs, to predict locomotion intentions based on gaze data. While gaze predicts steering intentions well, it sometimes diverges from locomotion goals. To tackle this, we classify gaze movements as either indicative of locomotor intention or not. This novel approach addresses the Midas Touch Problem of gaze. Datasets were collected in controlled VR environments featuring different tasks. We find that data sets with tasks that encouraged diverse navigation and gaze behaviors enable strong generalization.

The online VR simulation evaluation phase enabled safe and immersive testing, allowing the assessment of system performance and the integration of feedback for user guidance. Our approach provided smoother navigation control compared to traditional "Where-You-Look-Is-Where-You-Go" methods. Feedback improved user ratings of the system.

In the final stage, the system was deployed on a physical wheelchair equipped with an augmented reality (AR) device to provide feedback about the predictions to the user, allowing real-world evaluation. Despite differences in user behavior between VR and physical environments, the system successfully translated gaze inputs into precise and safe navigation commands. Users were able to steer the wheelchair solely using their eyes while simultaneously being able to look at destinations at the side of the path.

Index Terms: Virtual Reality, Eye Tracking, Eye-Tracking, Locomotion, LSTM, Transformer, Path Prediction, Machine Learning, Deep Learning, Gaze, Wheelchair, Feedback, Augmented Reality, Assistance.

1 INTRODUCTION

Electronic Wheelchairs for patients with limited mobility are usually steered with a joystick or similar handheld device that allows the adjustment of direction and speed. These wheelchairs rely on user input via hand movements but reach their limit when it comes to patients with motor impairments that limit mobility of the upper limbs as well.

The muscles responsible for moving the eye are directly connected to the brain through oculomotor nerves and bypassing the spinal cord [28]. This unique anatomical pathway allows eye movements to remain largely unaffected by motor disabilities that impair other parts of the body. In individuals without disabilities, gaze plays a critical role in locomotion, serving as more than just a tool for focusing on a destination. For instance, when navigating uneven or challenging terrain, people not only fixate on obstacles but also frequently shift their gaze to the ground several steps ahead [15, 14, 46, 36]. This behavior helps identify short-term waypoints, providing essential visual cues that inform immediate movement decisions.

Gaze patterns are also closely tied to changes in navigation direction. When following curved paths or making turns, individuals tend to adjust their gaze inward, aligning it with the curvature of the path [10, 16, 43]. Similarly, gaze behavior is intricately linked to decision-making processes, as people often shift their focus when evaluating options or planning their next move [53, 50, 21]. While these gaze patterns offer valuable insights into upcoming motor actions [27, 11], accurately predicting locomotion based on gaze data remains a complex and unresolved challenge. This difficulty arises from the dynamic and context-dependent nature of gaze behavior, which varies significantly across individuals, situations and tasks. Nevertheless, understanding the relationship between gaze and locomotion continues to be a key area of research, with potential applications in fields ranging from robotics to rehabilitation.

Another potential input signal for motor-impaired patients is brain data. Brain signals that can be gathered through non-invasive methods, such as EEG, often are noisy and provide limited information about locomotion intention. Nevertheless, research connects EEG signals to shifts in walking speed, direction, and obstacle avoidance, shedding light on the neural processes involved in motor planning and execution [9, 22, 19]. Progress has been made in recent years and with consideration for patients who cannot move their hands freely, such methods should be evaluated, not least as an addition to eye tracking.

Effectively utilizing these signals requires advanced predictive methods. One straightforward approach for using gaze to control a wheelchair is the "Where-You-Look-Is-Where-You-Go" method. This technique has been tested in both VR simulations and actual wheelchairs [24, 47, 2]. However, not every eye movement is directed towards a motion target, as our eyes also serve broader visual purposes. This leads to what is known as the "Midas touch problem" [17, 49], where gaze shifts that are directed at objects that are not motion targets can interfere with control. To address this, human computer interfaces need a way to differentiate between gaze movements that can be used for control and those for viewing other objects in the scene. Interfaces based on eye movements can tackle this problem in different ways [42, 33], but the system needs some method to distinguish between these two cases of eye movements.

We propose the use of a predictive model for this distinction. By incorporating such a model, an electronic wheelchair could anticipate a user's intended movements, allowing the system to use eye

*e-mail: gianni.bremer@uni-muenster.com

positions indicative of locomotion intention to autonomously adjust its trajectory in real-time. Recent advances in locomotion prediction have increasingly utilized artificial neural networks. In the field of human motion forecasting, Long Short-Term Memory networks (LSTMs) [13], have emerged as a common approach. Deep learning models that use LSTMs to predict locomotion based on egocentric data have shown increasing success [6, 5, 38]. Gaze tracking, in particular, has proven to be a valuable feature for these models [4, 43, 29, 5], underscoring its importance in accurately forecasting user movement. LSTMs have even been used to address the Midas touch problem of motion prediction from gaze specifically, although not in the context of locomotion [7]. Alternatively, Transformer networks [48, 3] take a different approach by utilizing attention mechanisms to predict human trajectories. They have been effectively used for predicting pedestrian locomotion behavior[52]. Even in the field of eye tracking, transformers deliver promising results [34].

If such a predictive system can accurately determine locomotion intentions from gaze, a further necessity is an interface to collect gaze data and present feedback about decoded intentions to the wheelchair user. The integration of model predictions into Augmented Reality (AR) systems presents a compelling possibility for enhancing control interfaces in assistive devices. AR could project information about intentions predictions directly into the user's field of view, offering immediate visual feedback on the wheelchair's anticipated movement. This feedback loop between user intention, system control, and AR visualization creates a dynamic human-computer interaction: the user's cognitive inputs guide the wheelchair's actions, while AR continuously updates the user on forthcoming movements. Such a system promises to significantly enhance navigation by promoting a transparent and intuitive interaction, thereby improving user autonomy, comfort, and safety in diverse mobility scenarios.

As users observe the projected path of their wheelchair in real time, they gain greater confidence in the system's decisions, fostering trust and a deeper connection with the assistive device. This closed-loop bi-directional system, where user intentions are seamlessly translated into action and visually reinforced by AR, has the potential to optimize assistive wheelchair control. It enhances both mobility and user satisfaction by ensuring that the device responds accurately to the user's cognitive directives, all while providing real-time feedback to ensure safe and efficient navigation. That system feedback in User–AI collaboration can enhance efficient strategies by the user has been shown in other motor tasks [31].

1.1 Related Work

Eye-gaze-controlled wheelchairs have been built since the 2000s [35, 30]. The most straightforward paradigm is a direct action "Where-You-Look-Is-Where-You-Go"-approach, where the movement direction is continuously aligned with the user's gaze [35, 2]. Speed is typically kept constant, although it can be reduced dynamically when users look around or break fixation [35]. Start and stop commands can then be implemented through additional signals, such as head gestures (nodding or shaking) [35], intentionally shutting the eyes for a short duration [47], or by looking towards and away from predefined gaze regions [30, 2, 23]. In some systems, automated obstacle detection via external sensors assists with stopping or avoiding collisions [47, 45]. Backward motion is commonly excluded to reduce user confusion but can be enabled by gazing at specific areas designated for reversing [30, 23, 2].

Since these early systems, numerous improvements have been proposed. Various gaze tracking methods have been implemented, ranging from rule-based algorithms [47] to deep learning models [51]. Some systems also integrate additional physiological data, such as EEG signals, to enhance intention recognition and control robustness [8, 25]. In parallel, VR has become a powerful testbed for gaze-based control system design. It allows researchers to simulate diverse environments and rapidly iterate interaction models [44, 23, 2]. However, transfer to real-world scenarios could reveal new mechanical problems and usability issues. In contrast, using AR in real-world gaze-controlled systems opens up promising possibilities for see-through graphical user interfaces [41] and intuitive system feedback. This is particularly relevant to systems like ours, where feedback loops are essential for safety and comfort.

However, few implementations directly address the Midas Touch problem of gaze. Most proposed solutions rely on additional user inputs to confirm relevant fixations. These inputs can come from outside the eye-tracking system, for example, using hand gestures [39]. However, such approaches are unsuitable for users with upper-body motor impairments. Thus, we need a solution that operates solely within the eye-tracking domain. One option is blink classification [32], but the most common strategy is to apply a dwell-time threshold: the wheelchair is activated only after the user intentionally fixates on a target or gaze region for a set duration [1, 44, 41, 2]. While effective in preventing unintended commands, the dwell-time method introduces a noticeable delay, making control less intuitive and responsive.

Araujo et al. [2] compared three gaze-based control approaches in VR simulations, evaluating them in terms of intuitiveness, controllability, and reliability. The study included: (1) a direct continuous control system, (2) a dwell-time method with a complex graphical overlay to define low-level locomotion targets, and (3) a dwell-time method that allowed users to select higher-level waypoints, which were then followed using automated path planning algorithm. They conclude that users slightly prefer gaze-defined waypoints over a direct continuous control system. The low-level movement command dwell-time method with the large overlay was rated poorly. However, in the real wheelchair user evaluation, participants felt uncomfortable with defining waypoints and then having the wheelchair follow that path. All solutions to the Midas Touch problem that require the users to use their eyes to define targets impede a natural way of steering, as their eye movements have to be used in an unnatural way. Users slightly preferred the waypoint-based method over the direct control approach. The low-level dwell-time system with the graphical overlay was rated least favorably. However, in a follow-up evaluation with real wheelchair users, participants reported discomfort using the system with a real head-mounted display. Collectively, solutions to the Midas Touch problem that rely on users explicitly selecting targets with their eyes tend to constrain natural steering behavior, forcing users to employ their gaze in an unnatural or cognitively demanding way.

Recently, predictive modeling with deep learning has been applied to address the Midas Touch problem in gaze-based interfaces. Subramanian et al. [45] proposed a method that uses deep computer vision to identify objects and then used K-Nearest Neighbor and Support Vector Machines to decide whether users want to approach an object that they looked at. Their binary classification approach relies on contextual environmental information rather than eye movement patterns alone. In contrast, Higa et al. [12] focus on decoding user intent directly from gaze dynamics. They define four discrete movement commands—"Stop," "Forward," "Left Turn," and "Right Turn"—and employ Long Short-Term Memory (LSTM) networks to classify these intentions based on temporal sequences of head and eye data. This approach leverages the sequential nature of these signals and represents a more direct method of intention decoding. These works illustrate the growing role of predictive models in enabling more intuitive gaze-based that can reflect user intention implicitly rather than requiring explicit commands.

While the intended user group remains individuals with motor impairments affecting the hands, Maule et al. [37] found only minor differences in driving kinematics when comparing gaze-based control to traditional joystick-based wheelchairs.

1.2 Aim of this Work

This work envisions a wheelchair that can be steered by decoding the user's intended direction of movement from their eye movements. Feedback about the decoded intention is then communicated back to the user through an augmented reality (AR) interface. The paper aims to develop and evaluate such a system in virtual simulations, employing Virtual Reality to gather the gaze data, develop the intention decoder and test the gaze-based locomotion control and interface in a controlled setting. Our approach can be divided into three key steps:

First, we want to create a deep neural network capable of differentiating gaze movements that are directed at a future waypoint from those gaze movements that are not related to the intended locomotion. This is different from the approach of Subramanian et al. [45] as we don't define target objects but locomotion targets that enable continuous locomotion as well. Our approach will suffice with egocentric features instead of detecting objects. Instead of predefined locomotion types [12], we will continue to use the gaze direction as our predictor of movement. Thus, we tackle the Midas Touch problem, without using explicit movement commands, dwell-defined waypoints or target objects. As EEG has been found to be a useful addition to eye-tracking data [8, 25], we aim to evaluate the potential benefits of enhancing eye-tracking-based models for wheelchair control by incorporating EEG data. Additionally, we compare LSTM and transformer architectures to determine which is more effective for predicting intended movement. A useful deep learning network will also need a suitable data set for training. Our locomotion intention decoder will be based on two data sets that we collected: one allowing free locomotion and eye movement, and another contrasting fixed paths and a visual search task. We will also compare the model performance for these two data sets to assess generalizability.

Second, we employ VR as a method to simulate wheelchair locomotion and to test both, the model and user's performance in navigating the simulated wheelchair in the virtual world exclusively by gaze. We believe that providing users with feedback about the model's predictions will increase their trust in the system and result in a more positive and pleasant experience compared to when no feedback is given. Third, we evaluate the system's usability in a real wheelchair with gaze-based navigation and AR feedback.

2 Deep Learning Model

2.1 Data Collection Experiments

To develop our deep learning models, we used datasets gathered from two separate Virtual Reality (VR) experiments, each designed to encompass a diverse range of locomotor behaviors. These experiments, conducted in controlled environments, employed joystick-based navigation. The resulting datasets have been made publicly available to support further research. Researchers can access the Forest Task dataset at https://osf.io/ney6v/ and the Course and Visual Search Double Task dataset at https://osf.io/4g9pw/.

2.1.1 Participants

Twenty individuals without motor impairments (12 females, 8 males) participated in the experiments, ranging in age from 18 to 50 years (M = 24.4, SD = 7.09). All had normal or corrected-to-normal vision, with four participants being left-handed. The participants, who were naive to the experimental objectives, provided informed consent and were compensated either through course credit or a payment of €10 per hour. Ethical approval for the study was granted by the Ethics Committee of the Department of Psychology and Sports Sciences of the University of Muenster.

2.1.2 Material

The VR experiments were conducted using an HTC Vive Pro Eye Head-Mounted Display (HMD), which provided a resolution of 1440×1600 pixels per eye, a refresh rate of 90 Hz, and a field of view of 110 degrees. The VR setup was augmented with six Vive Lighthouses 2.0. The virtual environments were developed using Unity3D, and the system was powered by an Intel Core i9 processor and an NVIDIA RTX2080 graphics card.

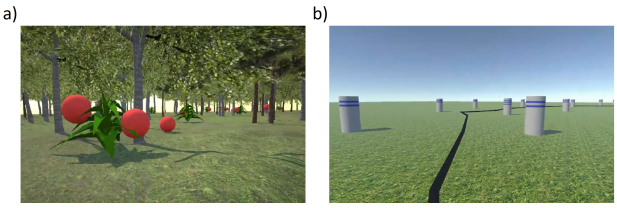

Figure 1: Screenshots of our data collection experiments. a) The forest experiment. Participants had to gather the red balls by approaching them. Participants were free to choose a ball. b) The course experiment. Participants had to follow the black bath while counting the stripes on the objects to the side of the road.

EEG data were recorded using a 32-channel mobile and wireless EEG system produced by Brainproducts, employing an Easycap recording cap with a 10-20 montage. Eye tracking was collected by the Vive Pro Eye, which offers a spatial accuracy ranging from 0.5° to 2° depending on the field of view. Delays in eye tracking data were observed between 50 ms and 60 ms, which, however, did not impede the goals of our study. A critical aspect of the methodology was the synchronization of EEG and VR data streams to ensure coherence between the brain and motor data collected. This was achieved using LabStreamingLayer, a system that enabled real-time data streaming to Python from both the EEG system and the VR environment, including eye tracking data. This setup allowed for accurate and temporally aligned data collection, crucial for developing real-time models.

2.1.3 Procedure

Before beginning the VR tasks, participants received detailed instructions about the tasks and provided demographic information. An EEG cap was fitted to each participant, with impedance checks conducted to ensure signal quality. After fitting the cap, participants donned the HMD, and eye tracking calibration was performed. The quality of the eye-tracking data was assessed using a custom validation tool that prompted the user to fixate on five designated points. Calibration was repeated until the user could accurately fixate on all five points. To minimize motion sickness and ensure stable head positioning, participants were asked to use a chin rest while controlling the joystick with their dominant hand. Additionally, when developing assistive systems for patients with severe motor impairments, a stable head position is more appropriate.

In the **Forest Task**, participants navigated a procedurally generated virtual forest with the objective of collecting red balls scattered throughout the environment. These balls were presented in four variations: near the ground, high in trees, at mid-levels, or as double balls on plants (see Figure 1 a)). Participants approached a ball until it turned grey, indicating that it could be collected by releasing the joystick. This task allowed participants to engage in a wide range of natural navigational behaviors, such as turning, reversing, and avoiding obstacles. There were no speed or accuracy requirements, allowing participants to move at their own pace and select their own gaze and movement targets, thus allowing us to capture diverse locomotor patterns and decision-making processes.

The **Course and Visual Search Double Task** combined locomotion with a visual search task. Figure 1 b) shows an image from that experiment. Participants followed a black path in a minimalist virtual environment while simultaneously counting grey cylinders with blue stripes that appeared along the path. The paths were gen-

erated using data from the Microsoft Geolife dataset [54, 56, 55], ensuring a varied and realistic navigation experience. Participants were tasked with identifying and counting cylinders with three blue stripes, which were interspersed among cylinders with two stripes. This task required participants to divide their attention between navigating the path and performing the visual search, simulating real-world multitasking scenarios. Each participant completed five different paths, with a marked endpoint indicating the end of each trial. This experiment provided a controlled environment to examine the interplay between locomotion and visual attention.

2.2 Model Training

We developed models to predict locomotion intentions by identifying gaze positions directed at future motion targets, utilizing the two distinct datasets. The datasets were evenly distributed between two experiments: 52.7% from the forest experiment and 47.3% from the course and visual search experiment. Multiple prediction models were created, all following a consistent preprocessing pipeline.

2.2.1 Preprocessing

We wanted to predict whether the current gaze position is directed at a future locomotion target or directed at another point in the world. Thus, our labels were created by marking whether the gaze position at each time point was close to any point along the subsequent 10 seconds of the walking trajectory. We calculated the Euclidean distance between gaze and future path for each time point and standardized it by dividing by the time until that position is reached, as points further into the future are further away and gaze positions will be less accurate. The threshold was set to 0.5.

Motion and eye tracking data were segmented into 100 ms intervals, using the past 2 seconds for prediction. Positions were transformed into a unified reference frame, resetting the origin and aligning the forward axis with yaw orientation [4]. Velocities were calculated separately for each axis, and eye tracking remained head-centered. Velocities, eye tracking data (head-centered), and EEG features were included in our models. EEG data was downsampled to 250 Hz, high-pass filtered (1 Hz), low-pass filtered (80 Hz), and analyzed for frequencies between 1–30 Hz using multitaper spectral density estimation.

In total, the available input features thus consisted of the two-dimensional velocity from the preceding 20 time steps, yaw and pitch eye tracking data, and optionally preprocessed EEG data. All data were z-score normalized. After excluding data with missing values, 166,685 input-output pairs were obtained.

2.2.2 Model Design

We implemented two main model architectures, LSTMs and Transformer models, to capture the temporal patterns in the data. For the models using EEG data, two one-dimensional convolutional layers with ReLU activations were applied to the EEG data before being fed into a model. The LSTM had 16 hidden units and a dropout rate of 0.1. The Transformer model included a linear embedding layer, positional encoding, four attention heads, two encoder and two decoder layers. A linear dense layer produced a single output, which was passed through a sigmoid function to create the final model output, allowing for rounding. During training, the sigmoid activation and binary cross-entropy loss function were integrated into a single layer. The models were optimized using Adam [20] with a learning rate of $1e-4$. Training occurred over 25 epochs with a batch size of 128, and the model with the lowest validation error was selected.

2.3 Results

Our primary objective was to assess the performance of various classification architectures and features. To facilitate a fair comparison, we employed leave-one-out cross-validation. Before training, both features and labels underwent z-score normalization. Prediction errors were calculated for each participant, and a paired t-test, adjusted with the Nadeau and Bengio correction [40], was used to determine whether one model significantly outperformed another ($\alpha = 0.05$). The Benjamini-Hochberg correction was applied to control for multiple comparisons across objectives.

Table 1: The performance of different models for classifying gaze movements indicative of locomotion intention. Weighted averages are given for sensitivity and specificity. Between-Subject Standard Deviations are given in brackets.

Model	Accuracy	Sensitivity	Specificity
Transformer	78 % [3 %]	77 % [17 %]	78 % [10 %]
LSTM	76 % [3 %]	77 % [8 %]	74 % [14 %]
Log. Regression	71 % [3 %]	63 % [23 %]	78 % [23 %]

The transformer architecture demonstrated the highest average prediction accuracy (77.67 %), outperforming the LSTM model (accuracy = 75.49 %). Figure 2 shows a visualization of these results, which indicated that the transformer model achieved a significantly higher average prediction accuracy than the LSTM model ($T = 4.94$, $p < 0.001$). While the transformer model required fewer floating-point operations than the LSTM (14 thousand vs. 77 thousand), the LSTM achieved faster inference times on an NVIDIA RTX 3080 laptop GPU (2.0 ms vs. 0.2 ms). Both models are lightweight enough to run effortlessly in real-time on older consumer-grade hardware. To further contextualize the performance, we also report the results of a simple comparative model that uses a logistic regression. As summarized in Table 1, logistic regression performed notably worse than both the transformer ($T = 5.46$, $p < 0.001$) and LSTM models ($T = 4.14$, $p < 0.001$). This logistic regression fit can be further simplified by only using the pitch values of the eye tracking system as the single input feature. A model that classifies glances to the ground as a movement target based on pitch direction and separates other gaze positions using that method reaches a very similar accuracy of 71.34 %. As looking to the ground is indicative of movement intention [14, 36], this method serves as a baseline approach.

To assess the contribution of different input features, we used the transformer architecture as it reached the best overall performance. Models using gaze as well as positional features reached the highest accuracy (Transformer with 77.67 % accuracy). With an accuracy of 75.78 % the model using only gaze came close to the best model. However, this small difference was statistically significant ($T = 5.65$, $p < 0.001$). The model using only positional data showed lower prediction accuracy (57.24 %). As Figure 2 shows occasional predictions below chance level, overfitting might have been a minor issue. Adding EEG data did not further improve prediction accuracy (average accuracy = 75.79 %). The positional model was significantly worse than both the model using both features ($T = 9.74$, $p < 0.001$) and the model only using eye data ($T = 8.56$, $p < 0.001$).

Figure 3 a) shows the precision recall curve for our main transformer model. The curve indicates a favorable trade-off between precision and recall. Even at higher recall levels, precision remains above 70%, highlighting the model's robustness in identifying true positives without sacrificing much precision.

Generalization was tested by comparing errors in two distinct datasets: the forest and the course tasks. The accuracy of the models was nearly identical, when training on all data (see Figure 3 b)). When models were trained on one dataset and tested on the other, the model trained on the course dataset performed worse in the forest task (accuracy 54.91 %) than the model trained on the forest dataset, which achieved an accuracy of 61.79 % in the course task. The variances were just small enough for this difference to become statistically significant ($T = 2.17$, $p = 0.022$).

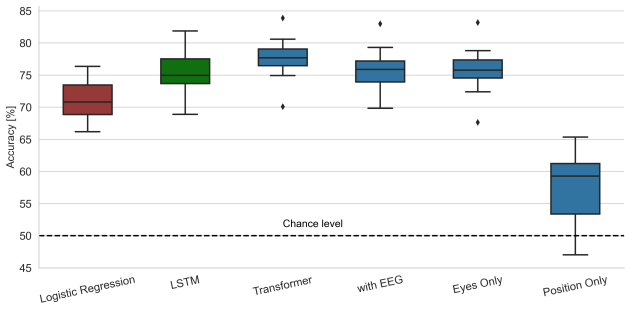

Figure 2: The prediction accuracy of different models. Blue models are based on the transformer architecture, green is an LSTM and brown is a logistic regression model.

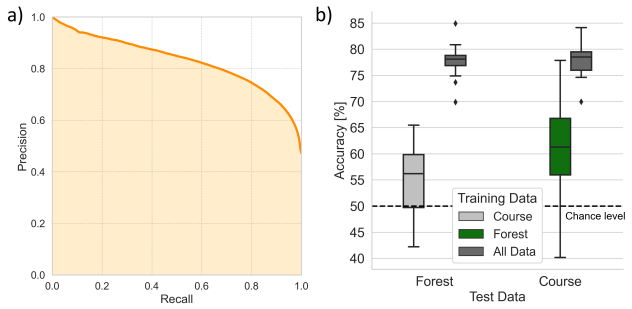

Figure 3: a) A precision recall curve of our model predictions. Even at high recall values, precision mostly remains above 70 %. b) Boxplots for models evaluated in different data sets. The color indicates the training data. Models evaluated on the forest task are shown on the left, whereas models tested in the course experiment are shown on the right.

2.4 Discussion

With our model, we showed that it is possible to accurately classify gaze movements indicative of locomotor intention and eye movements directed at different targets. We evaluated multiple architectures. The results of our study underscore the efficacy of the transformer architecture in accurately predicting locomotion intentions. Notably, the transformer model demonstrated superior performance when compared to both the LSTM model and a traditional logistic regression baseline. This performance advantage can be attributed to the transformer's unique ability to leverage its self-attention mechanisms, which excel at capturing complex patterns and relationships within sequential data. Specifically, the architecture's strength lies in its capacity to process and interpret dynamic features such as eye tracking data, which often contain rapid fluctuations during saccadic movements. Unlike traditional models that may struggle with such abrupt transitions, the transformer's attention-based approach allows it to effectively weigh and prioritize relevant temporal information, making it particularly well-suited for tasks involving erratic or non-linear data patterns characteristic of biological signals.

Our findings identify the history of gaze behaviour as the central feature in predicting whether our gaze is aligned with a locomotion target or future waypoint. Models that integrated eye tracking information with other features like the history of positional data yielded a higher accuracy. This finding aligns with existing research, which emphasizes the predictive power of gaze data in locomotion tasks [4, 29], as human gaze typically precedes movement targets and encodes directional information before actions are executed [27, 11, 26]. Incorporating EEG data alongside eye tracking did not significantly improve the prediction accuracy, likely because eye tracking alone captures sufficient information regarding gaze direction and locomotor intention. Future work could explore optimizing EEG live preprocessing techniques to better handle artifacts and assess its potential in scenarios where eye tracking data is unavailable or unreliable

We also examined the generalization capabilities of our classification model using two distinct datasets: one derived from a forest navigation task and the other from a course-following and visual search task. Despite the differing constraints on locomotion and gaze behavior in these tasks, both datasets demonstrated similar predictive accuracy. However, the model trained on the forest dataset exhibited superior generalization when applied to the course task, compared to the reverse scenario. This suggests that the more varied and naturalistic movements in the forest task provided a richer set of training data, enabling better generalization to different locomotor scenarios.

3 Virtual Reality Evaluation Experiment

To test the performance of the trained model and three possible control systems for steering, a real-time evaluation experiment was conducted. Subjects were asked to steer through a virtual environment using their gaze and three possible control systems. Two of these systems used our trained model as an intention decoder. The environment matched the scenario of our Course and Visual Search Double Task experiment.

3.1 Method

3.1.1 Participants

The evaluation study included 18 volunteers with no motor impairments (15 women, 3 men) between the ages of 19 and 27 (M = 21.84, SD = 1.98), four of whom were left-handed. To prevent bias, participants were unaware of the study's purpose. Prior to taking part, all provided informed consent. They were compensated with either course credit or a payment of €10 per hour. Data from one additional participant had to be discarded due to severe motion sickness. The study was approved by the Ethics Committee of the Department of Psychology and Sports Sciences of the University of Muenster.

3.1.2 Material

This VR experiment utilized an HTC Vive Pro Eye HMD as well. Two Vive Lighthouses were used to track the participants. The virtual environment was created using Unity3D, and the system ran on an NVIDIA RTX4090 graphics card. The deep learning model was running in python. Unity and python communicated with a TCP-connection with each other. Latencies were below the 90 Hz refresh rate of the HMD. Similar to the data collection experiment, the subjects were asked to position their head straight on a chin rest.

A questionnaire was completed for each control system in order to obtain the opinions of the participants. Here, the participants were able to rate the control system on a scale of 1 to 10 and assess trust in the system and their own experience of control.

3.1.3 Procedure

While the translation speed was fixed at 1.5 meter per second, three locomotion control systems were used by the participants, which we wanted to test and compare:

1. The first system followed a "Where-You-Look-Is-Where-You-Go"-approach. Whenever a participant was looking to the left or to the right, a respective angular velocity was applied. This way it was possible to directly control the direction of locomotion. The angular velocity was proportional to the yaw angle of the eye tracking. We refer to this system as "direct action" control system.

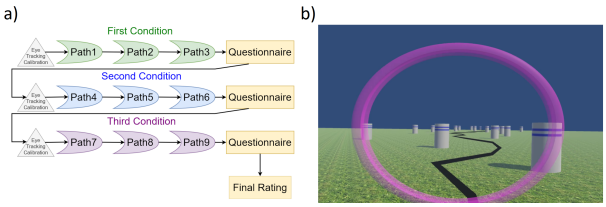

Figure 4: a) Our experimental procedure. Participants would complete three runs with each system before answering the questionnaire. b) A screenshot of the evaluation experiment with feedback shown as a pink circle. The pink circle would only show up, when the system was positive that gaze was on a locomotion target. We reused the visual setup from the course experiment.

2. The second control system used our trained model to decode user intention, i.e. differentiate between gaze movements to a locomotion target and gaze movements to a different target (e. g. one of the objects to count the stripes). Angular velocity was applied only when the model confidently identified the user as fixating on a future locomotion target. To operationalize this, a prediction threshold of 53% was set for identifying locomotion targets, while a separate threshold of 46% was then used to classify gaze toward non-locomotion targets. This 7% gap (i.e., a confidence buffer) was introduced to avoid rapid toggling of classifications near the 50% mark, thus ensuring more stable control.

   To prevent abrupt halts in movement, a smoothing mechanism was applied when the model detected a saccade to a different target. Specifically, the angular velocity ω was exponentially reduced over time using the following decay function:

$$\omega_i = 0.95^i \omega_0 \tag{1}$$

   where ω_0 is the angular velocity at the moment the model stopped detecting a valid fixation, and i is the number of frames since that moment. At a frame rate of 90 Hz, this results in a rapid but smooth decay of velocity.

3. The final system also employed the trained model and functioned identically to the second system. However, when the model detected a gaze position indicative of locomotor intention and adjusted the virtual direction accordingly, it additionally displayed a pink circle in the peripheral visual field (see Figure 4b)). This served as feedback to communicate the system's current classification state to the user.

Participants were instructed on the tasks and provided demographic information. The experiment was divided into three blocks, each corresponding to a different locomotion control system. Each block consisted of three paths that had to be completed with the respective locomotion control system. Thus, every participant completed a total of nine paths during the experiment (see Figure 4 a)). After each block, participants completed a questionnaire about that particular control system. The order of the three locomotion control systems was counterbalanced across participants to mitigate order effects. Additionally, the same set of randomly generated paths was reused across participants, with each path assigned to a different control system, ensuring that path variability did not bias the results.

The objectives and visual setup were similar to those in the course experiment from the initial data collection. The task was always to follow a black path and at the same time count the number of objects with exactly three stripes next to the path (see Figure 1 b)). Participants were instructed to keep the distance to the black line as small as possible while counting the objects with three stripes. As before, paths were randomly generated from Geolife [54, 56, 55].

The speed of movement was constant. Each system used the participant's gaze position to determine the direction of movement. Participants were informed that intentional blinking or unnatural eye movements would not improve the system's performance.

3.2 Results

During the evaluation experiment, we employed our intention decoder as a real-time motion controller driven by gaze input. Participants navigated a virtual environment solely by directing their gaze, while our trained model predicted whether the current gaze direction was indicative of locomotion intention and adjusted movement accordingly. We benchmarked its performance against a straightforward direct action system ("Where-You-Look-Is-Where-You-Go"-approach), which directly translated gaze direction into movement.

Regardless of the navigation system used, in every run all participants successfully arrived at the final locomotion target. Using the model with feedback, participants' eye movements exhibited by the lowest average absolute yaw values of 8.98 degree (without feedback 11.67 degree, direct action 9.02 degree) and the lowest average pitch values of -8.69 degree (without feedback -6.50 degree, direct action -5.45 degree). The model gave similar predictions in the three conditions (70.77 % of gazes on path for the model with feedback, 63.43 % for the model without feedback and 62.65 % for direct action), although in the direct action system these were not applied.

3.2.1 Steering Performance

We used three metrics for the steering performance: first, the closeness to the path, second the difference between participants orientation and the direction of the given path ahead and, third, the smoothness of the steering. Measures like this are known to vary with gaze control during car driving [18].

We tested the hypothesis that our intention decoder performs better than the direct action control steering system by using both versions of the decoder: the intention decoder without feedback and the intention decoder with feedback (where a pink circle indicated the classification result). We compared the results from these two steering systems to those of the direct action system using three paired t-tests ($\alpha = 0.05$). The assumption of normality was assessed using Shapiro-Wilk tests and could not be rejected for any of the comparisons. To account for multiple comparisons, we applied the Benjamini-Hochberg procedure to control the false discovery rate.

To assess the closeness to the path, we excluded the first and last 15 seconds of each run to minimize artifacts. We then measured the median distance of the virtual wheelchair to the black path. With 14.07 cm on average for the feedback condition and with 12.15 cm on average for the intention decoder without feedback, the use of our intention decoder resulted in lower distances than the direct action control system (average distance 15.07 cm). Although this difference between our intention decoder and the direct action system did not reach statistical significance (T = 1.2, p = 0.14), most participants performed better with our model (see Figure 5 a)).

The difference between the virtual wheelchair's orientation and the optimal wheelchair orientation defined by the direction of the black path was also lower for our intention decoder (6.20° without feedback and 6.59° with feedback) compared to the 7.81° of the direct action system (see Figure 5 b). This was statistically significant (T = 2.60, p = 0.014).

To assess steering smoothness, we analyzed the adjustments required to stay on the intended path. We examined changes in steering angles over a 500 ms time frame. The inverted angle in serves as our measurement of smoothness. The smoothness scores were

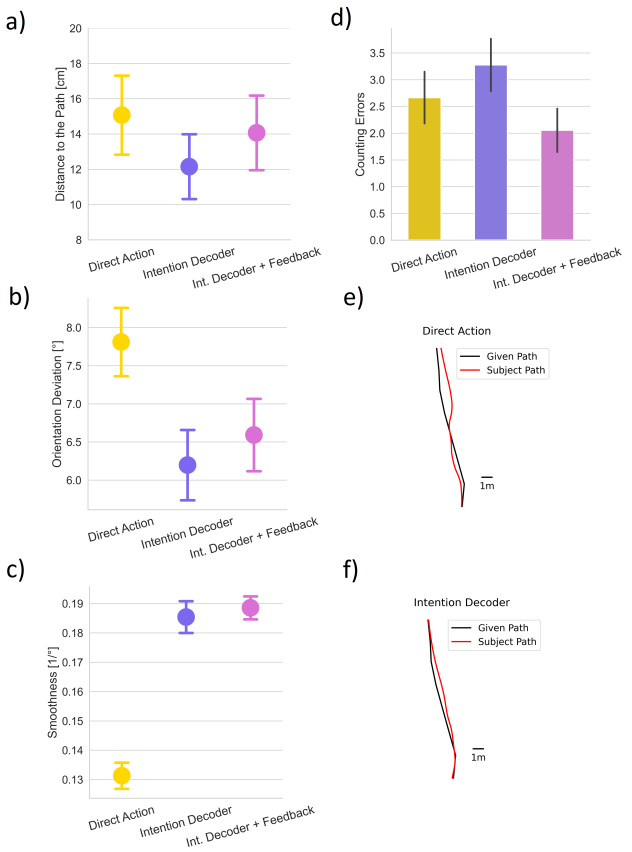

Figure 5: Participants performance in the simulation experiment. Error bars denote standard errors. a) The average distances to the path for all three conditions. b) The average difference between the wheelchair's orientation and the path direction observed in the different conditions. c) The smoothness for all three conditions d) A bar plot showing the number of difference between the true number of target objects and the number of target objects as counted by the participants. e) and f) real examples of paths with the two navigation systems. A birds eye view is depicted. e) shows the back-and-forth-steering adjustments with the direct action system.

0.202 deg^{-1} (σ = 0.053 deg^{-1}) without feedback and 0.205 deg^{-1} (σ = 0.043 deg^{-1}) with feedback, compared to the much lower 0.131 deg^{-1} (σ = 0.028 deg^{-1}) smoothness result of the direct action approach. The difference between our intention decoder and the direct action model was significant (T = 12.65, p $<$ 0.001). The smoothness produced by our intention decoder also came closer to the smoothness score produced by the joystick data from the Course and Visual Search Double Task data collection experiment (0.189 deg^{-1} (σ = 0.046 deg^{-1})). On average, participants needed an absolute steering angle of 5.19° (σ = 0.96°) when using our intention decoder, allowing for more subtle directional changes compared to the 7.97° (σ = 1.68°) observed with the direct action system. Our intention decoder produced smoother, smaller turning angles, whereas the direct action system required sharper, more abrupt turns. Figure 5 c) illustrates that difference.

3.2.2 Counting Task

Additionally, we measured the counting task performance. We determined the difference between the number of objects the participant reported with the actual number of target objects in all three runs (see Figure 5 d)). Participants made the least errors in the feedback condition, with 2.06 errors on average (σ = 1.68). The direct action control condition came in second with 2.67 errors on average (σ = 2.00) and participants made the most errors in the model without feedback condition with 3.28 errors on average (σ = 2.02).

We investigated these differences using a repeated measures ANOVA (α = 0.05). The assumption of normality was assessed with Shapiro-Wilk tests and was not violated. The main effect of the control system narrowly reached statistical significance (F = 3.95, p = 0.043). Post-hoc comparisons were performed using two-sided paired t-tests with Benjamini-Hochberg correction for multiple comparisons. The differences between the direct action condition and the intention decoder with feedback (T = 1.38, p = 0.23), as well as the direct action condition and the intention decoder without feedback (T = -1.26, p = 0.23), were not statistically significant. While the difference in mean values was the largest, the comparison between the decoders without feedback and with feedback narrowly missed significance (T = 2.65, p = 0.050).

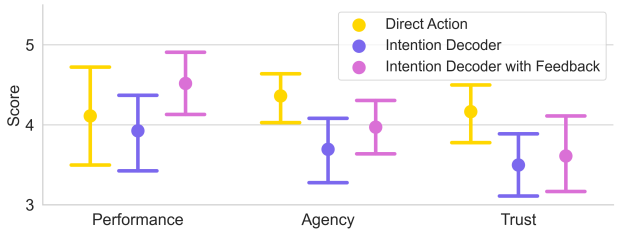

Figure 6: Participants ratings of different models. We show average scores of the answers to our questionnaire regarding general performance, agency and trust. Most questions were asked on a 5-point Likert Scale. The error bars show confidence intervals.

3.2.3 Ratings

To gain additional exploratory insight, we asked our participants to rate each model and their respective experience. Figure 6 shows scores for the different models. Given the large variances typically seen in questionnaire data of this sample size, no definitive statistical conclusions can be drawn, as we lack statistical power. Nevertheless, we can observe some general trends. Additional statistical analyses of the questionnaire data are provided in the supplementary materials. First, the score for general performance was calculated by averaging over a 1-to-10 rating of the system, the 5-point Likert Scale item "The control system made it easy to solve the tasks." and a doubled inverted relative rating in which the participants could assign placements to the different models. The model with feedback received the best rating of 4.52 points, the model without only received 3.93 points, whereas the direct-action-approach got 4.11 points. Second, agency was measured through the items "I had control over the movement" and "The direction of movement was caused by me". The direct action system came in first with 4.36 points, the model with feedback received 3.97 points and the model without received 3.69 points Finally, trust was measured by rating "I could trust the control system". The respective average ratings for the direct action system, the intention decoder with and the intention decoder without feedback were 4.17, 3.61 and 3.50.

3.3 Discussion

Our experiment compared the intention-decoding model to the direct action "Where-You-Look-Is-Where-You-Go" approach. While both enabled successful navigation, our intention decoder provided smoother movement with fewer corrections. This outcome indicates that our intention decoder enabled users to complete search tasks more efficiently, avoiding the common pitfall of inadvertently steering the virtual wheelchair toward peripheral objects, which typically necessitates additional turns to realign with the

path. Participants were more aligned with the path using our intention decoder. By reducing the frequency of such errors, our intention decoder significantly diminished the need for participants to make repetitive back-and-forth steering adjustments. These results demonstrate the intention decoder's potential for improving user experience and efficiency in locomotion control. While choice of system showed a significant difference in the counting task, the result should be interpreted with caution due to high variances, marginal p-value and non significant post-hoc tests. The intention decoder with feedback resulted in the least errors, while the intention decoder without feedback resulted in the most errors. If not spurious, however, this finding would highlight a potential benefit of providing real-time system feedback.

The integration of our intention decoder with an AR interface could provide real-time feedback to users, creating a symbiotic relationship between the system and the user, facilitating bi-directional communication. We asked participants to rate their subjective sense of successful navigation, sense of agency and trust in the system. When comparing our intention decoder without feedback with the direct action system, participants rated the direct action system better for every category. However, feedback to the user made a difference. The intention decoder with feedback was rated better than the model without feedback for every category. For the general performance ratings, the model with feedback could even beat the direct action model. However, no statistically significant conclusions can be drawn from the questionnaire data as we lack the statistical power to handle questionnaire data. More research with a much larger sample sizes is needed here to contextualize these initial exploratory numbers.

4 Wheelchair Demonstration

To evaluate our intention decoder in a real physical environment, we developed a hardware prototype consisting of a motorized wheelchair whose speed and direction could be adjusted from a computer. For eye tracking and to be able to display feedback signals to the user, an augmented reality (AR) display, the Microsoft HoloLens 2 was connected to the system.

The hardware has been built from off-the-shelf components but is equipped with computing hardware and device drivers to allow the chair to be steered using our model. To control the robotic prototype, ROS (Robot operating system) has been adopted as a software architecture. Whenever our intention decoder predicted that the current gaze direction aligns with the intended movement direction, an angular velocity was applied until gaze direction and wheelchair orientation aligned. Figure 7 b) shows the wheelchair.

4.1 Electrical and mechanical modifications to the Wheelchair

A commercially available electric powerchair was modified extensively to meet the requirements of the proposed use case. The original control system was removed, and the brushed DC motors were replaced with open-loop stepper motors and their respective controllers. An electrical cabinet was installed on the backrest to house the computing platform (NVIDIA Jetson Orin Nano), a 24V-to-12V DC/DC converter, and the electrical system. Additionally, an emergency stop button was integrated to cut power to the motors, enhancing safety. A LiDAR scanner and an overhead RGB camera were also added, alongside a robot base frame securely attached to the wheelchair's main frame. These modifications allow the system to operate for several hours using its onboard sealed lead-acid (SLA) batteries.The onboard batteries are rechargeable with the wheelchair's original charging system.

A custom mechanical interface was designed and constructed from aluminum and steel to support the equipment. It included steel attachment brackets and a 10 mm thick aluminum mounting plate with a 9×9 grid of 9 mm diameter holes spaced 20 mm apart, offering flexibility in mounting configurations. A robotic arm equipped with an anthropomorphic hand (7-DOF Franka Emika Panda) was attached to the wheelchair via an aluminum frame made from standard 40×40 MISUMI profiles. However, for this work we did not use this feature and focused on locomotion entirely.

4.2 System Testing

The user interface incorporates an augmented reality (AR) headset. A custom software solution was developed to enable TCP communication between the HoloLens 2 AR headset, the Python-based model, and the wheelchair's ROS (Robot Operating System) framework. Predicted system outputs were visualized in real-time on the HoloLens 2 display. When the model predicted that the current gaze direction corresponded with the users movement intention, a pink circle was presented on the AR device while an angular velocity was applied to align the wheelchair orientation. This pink circle was shown in the visual periphery and resembled the pink circle that was used in our VR simulation experiment (see Figure 4 b)).

To align the wheelchair orientation, we added the head-centered yaw direction of the eye tracking system integrated in the HoloLens 2 to the wheelchair-centered head angle. The wheelchair-centered head angle could be determined with the overhead RGB camera and an ArUco glued on top of the HoloLens 2.

4.2.1 Participants

Eleven participants without motor impairments (4 female, 7 male) provided informed consent to evaluate the wheelchair system. The mean age was 43.18 years (SD = 11.05). Nine participants had no prior experience with the wheelchair. All had normal or corrected-to-normal vision. The study was approved by the Ethics Committee of the Department of Psychology and Sports Sciences of the University of Muenster.

4.2.2 Task

The evaluation task required participants to navigate the wheelchair in circular paths around minor obstacles (see Figure 7a)). These obstacles were placed to increase the difficulty of accurately steering along the intended path. Before starting the task, participants were told that they could take small detours within the circle or turn around and drive in the opposite direction of the circle if they wanted. This flexibility was intended to encourage more natural and varied driving behavior and to evaluate the system's robustness under varied conditions. However, they were instructed not to leave the testing area. They were also encouraged to look around the room in a natural way, rather than keeping their gaze fixed on the path, to assess how the system handled varying gaze patterns. Each participant drove the wheelchair for 4 minutes.

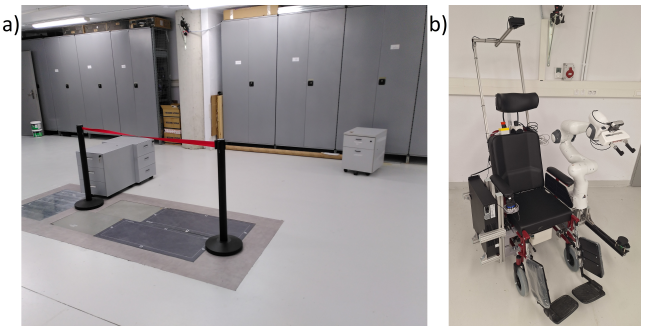

Figure 7: a) The room in which participants drove the wheelchair around the red barrier in circular paths. The small cabinets were obstacles that had to be avoided. b) The wheelchair with the robotic arm attached.

Table 2: Different Metrics for the three stages of development.

Measurement	Experiment		Evaluation
	Joystick	VR Sim.	Wheelchair
Errors per Minute	None	0.01	0.02
Subjects without Errors	100 %	78 %	91 %
Smoothness	0.19 deg^{-1}	0.20 deg^{-1}	0.18 deg^{-1}
Eye Pitch	-4.90°	-9.04°	-30.48°
Abs. Eye Yaw	9.81°	7.41°	6.47°
Pred. Locomotion Int.	50.82 %*	66.05 %	91.38 %

*The true amount of eye positions with locomotion intention in the training data was 52.46 %.

4.3 Results

For safety, the wheelchair operated at a reduced maximum speed of approximately 0.2 m/s. The highest recorded speed was 0.228 m/s, while the average participant speed was 0.144 m/s.

We used three measurements to quantify the wheelchair performance: First, we count the number of collision or near collision incidents where we had to stop the wheelchair. Second, we count the number of navigation errors produced by the system. Navigation errors are defined as straying from the given path in a significant way. Given a somewhat consistent smoothness, this can be defined as 4 standard deviations. Third, we measure the smoothness. From our experimental data, we measured an average smoothness of 0.19 deg^{-1} in the joystick data and 0.20 deg^{-1} in the VR simulation driving. While this evaluation is not an experiment, we would expect the smoothness to be similar.

Across all trials, a near-collision incident occurred only once, during which the wheelchair was manually stopped. The subject involved in this case exhibited highly exploratory behavior. Second, while there was no significant straying in the wheelchair evaluation, we count the one near collision incident as a navigation error, as it might have led to a significant deviation from the intended path (and possibly a collision) if we had not intervened. Third, we measured an average smoothness of 0.18 deg^{-1}.

Table 2 compares measurements in the three stages of system development regarding gaze behavior, smoothness, error rate and how often our system predicted a gaze position indicative of locomotion intention.

4.4 Discussion

Our system, trained exclusively on VR simulation data, demonstrated successful transfer to a real wheelchair. While steering was easy, participant behavior differed from both the VR simulation and prior data collection. Participants looked at the ground more often, showing eye movements linked to stronger locomotor intention. Unlike in VR, there was no visual search task, reducing upward glances, as we prioritized safety. This could also reflect greater caution in the real world, where users may have feared collisions. Additionally, participants had limited time to adapt to the wheelchair, which may have influenced their behavior. Future tests should allow more practice.

Despite these behavioral differences, the system maintained high control fidelity and generalization without fine-tuning the weights. Users navigated through the obstacle course while being able to look at other objects without the wheelchair turning in the wrong direction, demonstrating strong real-world transfer.

5 General Discussion

We introduce a three-stage framework for gaze-based wheelchair navigation — offline evaluation, VR simulation, and real-world deployment — integrating machine learning, immersive simulation, and assistive technology. It systematically bridges research and real-world application to improve mobility control.

Our intention decoder acts as a filter for gaze inputs, ensuring only eye movements indicative of locomotion intention adjust the wheelchair's direction. Unlike previous deep learning approaches to the Midas Touch problem [45, 12], our method preserves raw gaze direction as a movement cursor, allowing for intuitive control.

The framework also highlights the importance of training models with diverse and naturalistic datasets to ensure generalizability across different environments. The combination of predictive modeling with intuitive user interfaces holds promise not only for assistive mobility but also for broader applications in human-machine interaction.

A key finding was the critical role of AR feedback in enhancing user trust and scene awareness, demonstrating that bi-directional interaction between the system and user is crucial for successful adoption. Displaying system feedback to the user led to fewer counting errors and more favorable ratings when surveying our participants.

5.1 Limitations

A key limitation of this study is that all participants were able-bodied individuals without motor impairments. While this population is appropriate for initial prototyping and feasibility testing, it does not fully reflect the end users for whom such gaze-based wheelchair control systems are intended. Individuals with motor impairments may exhibit different gaze patterns, attention dynamics, or cognitive strategies, particularly in real-world environments where fatigue, involuntary movements, or co-occurring conditions could influence system usability and performance. As such, the generalizability of our findings to the target population remains limited. Future studies should include participants with motor impairments to assess the system's practical applicability, adaptability, and robustness in real-world conditions. Another limitation lies in the statistical power of the questionnaire-based data. Given the relatively small sample size, the resulting variance is high, and definitive statistical conclusions cannot be drawn from the subjective ratings. While our analysis revealed trends that may inform future design decisions, these results should be viewed as exploratory and hypothesis-generating rather than confirmatory. Larger-scale studies are required to validate these observations. Additionally, the focus on egocentric features, those related to the user's body and gaze movements, was a deliberate choice to ensure that the models could generalize across different environments. However, incorporating additional data, such as optic flow or environmental context, could further improve model performance. While the current study used standard implementations of transformer and LSTM models, further refinements and the inclusion of more diverse environments could enhance prediction accuracy.

6 Conclusion

This study presents a novel interface for gaze-based wheelchair navigation, achieving seamless learning transfer from offline evaluation to VR simulation to physical implementation. The use of advanced predictive models, AR feedback, and a stepwise development approach ensured robust performance and user satisfaction. Deep neural networks are able to distinguish gazes to future waypoints from gazes directed at other targets in the scene. Wheelchair simulations show that a system like this enables accurate and smooth steering control without expendable course corrections.

Acknowlegments

Special thanks to Josefina Dreiling and Malena Raabe for their help in running experiment 2. This work was supported by the German Research Foundation (LA 952/11-1) and has received funding from the European Union's Horizon 2020 and Horizon Europe research programs under grant agreements No 951910 and No 101086206.

REFERENCES

[1] Y. Adachi, H. Tsunenari, Y. Matsumoto, and T. Ogasawara. Guide robot's navigation based on attention estimation using gaze information. In *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)(IEEE Cat. No. 04CH37566)*, vol. 1, pp. 540–545. IEEE, 2004. 2

[2] J. M. Araujo, G. Zhang, J. P. P. Hansen, and S. Puthusserypady. Exploring eye-gaze wheelchair control. In *ACM symposium on eye tracking research and applications*, pp. 1–8. ACM, Stuttgart, Germany, 2020. 1, 2

[3] G. Bremer and M. Lappe. Predicting locomotion intention using eye movements and eeg with lstm and transformers. In *2024 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 21–30. IEEE, 2024. 2

[4] G. Bremer, N. Stein, and M. Lappe. Predicting future position from natural walking and eye movements with machine learning. In *2021 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*, pp. 19–28. IEEE, IEEE, Taichung, Tawian, 2021. doi: 10.1109/AIVR52153.2021.00013 2, 4, 5

[5] G. Bremer, N. Stein, and M. Lappe. Machine learning prediction of locomotion intention from walking and gaze data. *International Journal of Semantic Computing*, 17(01):119–142, 2023. doi: 10.1142/S1793351X22490010 2

[6] Y.-H. Cho, D.-Y. Lee, and I.-K. Lee. Path prediction using lstm network for redirected walking. In *2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 527–528. IEEE, IEEE, Reutlingen, Germany, 2018. doi: 10.1109/VR.2018.8446442 2

[7] P. Festor, A. Shafti, A. Harston, M. Li, P. Orlov, and A. A. Faisal. Midas: Deep learning human action intention prediction from natural eye movement patterns, 2022. 2

[8] M. Gneo, G. Severini, S. Conforto, M. Schmid, and T. D'Alessio. Towards a brain-activated and eye-controlled wheelchair. *international Journal of Bioelectromagnetism*, 13(1):44–45, 2011. 2, 3

[9] K. Gramann, J. T. Gwin, D. P. Ferris, K. Oie, T.-P. Jung, C.-T. Lin, L.-D. Liao, and S. Makeig. Cognition in action: imaging brain/body dynamics in mobile humans. *Reviews in the neurosciences*, 22:593–608, 2011. doi: 10.1515/RNS.2011.047 1

[10] R. Grasso, P. Prévost, Y. P. Ivanenko, and A. Berthoz. Eye-head coordination for the steering of locomotion in humans: an anticipatory synergy. *Neuroscience Letters*, 253(2):115–118, 1998. doi: 10.1016/s0304-3940(98)00625-9 1

[11] M. Hayhoe and D. Ballard. Eye movements in natural behavior. *Trends in Cognitive Sciences*, 9(4):188–194, 2005. doi: 10.1016/j.tics.2005.02.009 1, 5

[12] S. Higa, K. Yamada, and S. Kamisato. Intelligent eye-controlled electric wheelchair based on estimating visual intentions using one-dimensional convolutional neural network and long short-term memory. *Sensors*, 23(8):4028, 2023. 2, 3, 9

[13] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735 2

[14] M. A. Hollands and D. E. Marple-Horvat. Visually guided stepping under conditions of step cycle-related denial of visual information. *Experimental Brain Research*, 109(2):343–356, 1996. doi: 10.1007/BF00231792 1, 4

[15] M. A. Hollands, D. E. Marple-Horvat, S. Henkes, and A. K. Rowan. Human eye movements during visually guided stepping. *Journal of Motor Behavior*, 27(2):155–163, 1995. doi: 10.1080/00222895.1995.9941707 1

[16] T. Imai, S. T. Moore, T. Raphan, and B. Cohen. Interaction of the body, head, and eyes during walking and turning. *Experimental Brain Research*, 136(1):1–18, 2001. doi: 10.1007/s002210000533 1

[17] R. J. Jacob. The use of eye movements in human-computer interaction techniques: what you look at is what you get. *ACM Transactions on Information Systems (TOIS)*, 9(2):152–169, 1991. 1

[18] F. I. Kandil, A. Rotter, and M. Lappe. Driving is smoother and more stable when using the tangent point. *Journal of vision*, 9(1):11–11, 2009. 6

[19] A. Khajuria, R. Sharma, and D. Joshi. Eeg dynamics of locomotion and balancing: Solution to neuro-rehabilitation. *Clinical EEG and Neuroscience*, 55(1):143–163, 2024. doi: 10.1177/15500594221123690 1

[20] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization, 2017. 4

[21] D. Kit, L. Katz, B. Sullivan, K. Snyder, D. Ballard, and M. Hayhoe. Eye movements, visual search and scene memory, in an immersive virtual environment. *PLOS ONE*, 9(4):1–11, 04 2014. doi: 10.1371/journal.pone.0094362 1

[22] J. E. Kline, H. J. Huang, K. L. Snyder, and D. P. Ferris. Isolating gait-related movement artifacts in electroencephalography during human walking. *Journal of neural engineering*, 12(4):046022, 2015. doi: 10.1088/1741-2560/12/4/046022 1

[23] S. I. Ktena, W. Abbott, and A. A. Faisal. A virtual reality platform for safe evaluation and training of natural gaze-based wheelchair driving. In *2015 7th International IEEE/EMBS Conference on Neural Engineering (NER)*, pp. 236–239. IEEE, 2015. 2

[24] Y. Kuno, N. Shimada, and Y. Shirai. Look where you're going [robotic wheelchair]. *IEEE Robotics & Automation Magazine*, 10(1):26–34, 2003. 1

[25] H. A. Lamti, M. M. Ben Khelifa, P. Gorce, and A. M. Alimi. A brain and gaze-controlled wheelchair. *Computer Methods in Biomechanics and Biomedical Engineering*, 16(sup1):128–129, 2013. 2, 3

[26] M. Land and B. Tatler. *Locomotion on foot*, pp. 100–115. Oxford University Press, New York, USA, 07 2009. doi: 10.1093/acprof:oso/9780198570943.003.0006 5

[27] M. F. Land and M. Hayhoe. In what ways do eye movements contribute to everyday activities? *Vision Research*, 41(25-26):3559–3565, 2001. doi: 10.1016/S0042-6989(01)00102-X 1, 5

[28] R. J. Leigh and D. S. Zee. *The neurology of eye movements*. Oxford University Press, USA, New York, USA, 2015. 1

[29] M. Li, B. Zhong, E. Lobaton, and H. Huang. Fusion of human gaze and machine vision for predicting intended locomotion mode. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 30:1103–1112, 2022. doi: 10.1109/TNSRE.2022.3168796 2, 5

[30] C.-S. Lin, C.-W. Ho, W.-C. Chen, C.-C. Chiu, and M.-S. Yeh. Powered wheelchair controlled by eye-tracking system. *optica applicata*, 36, 2006. 2

[31] O. Lukashova-Sanz, M. Dechant, and S. Wahl. The influence of disclosing the ai potential error to the user on the efficiency of user–ai collaboration. *Applied Sciences*, 13(6):3572, 2023. 2

[32] W. Luo, J. Cao, K. Ishikawa, and D. Ju. A human-computer control system based on intelligent recognition of eye movements and its application in wheelchair driving. *Multimodal Technologies and Interaction*, 5(9):50, 2021. 2

[33] P. Majaranta, K.-J. Räihä, A. Hyrskykari, and O. Špakov. Eye movements and human-computer interaction. *Eye movement research: An introduction to its scientific foundations and applications*, 1:971–1015, 2019. 1

[34] A. A. Masaoodi, H. H. Abbas, and H. I. Shahadi. A comparative study of traditional and transformer-based deep learning models for multi-class eye movement recognition using collected dataset. In *2023 International Conference on Advanced Mechatronics, Intelligent Manufacture and Industrial Automation (ICAMIMIA)*, pp. 624–630. IEEE, IEEE, Mataram City, Indonesia, 2023. 2

[35] Y. Matsumotot, T. Ino, and T. Ogsawara. Development of intelligent wheelchair system with face and gaze based interface. In *Proceedings 10th IEEE International Workshop on Robot and Human Interactive Communication. ROMAN 2001 (Cat. No. 01TH8591)*, pp. 262–267. IEEE, 2001. 2

[36] J. S. Matthis, J. L. Yates, and M. M. Hayhoe. Gaze and the control of foot placement when walking in natural terrain. *Current Biology*, 28(8):1224–1233, 2018. doi: 10.1016/j.cub.2018.03.008 1, 4

[37] L. Maule, M. Zanetti, A. Luchetti, P. Tomasin, M. Dallapiccola, N. Covre, G. Guandalini, and M. De Cecco. Wheelchair driving strategies: A comparison between standard joystick and gaze-based control. *Assistive Technology*, 35(2):180–192, 2023. 2

[38] J. Mayor, P. Calleja, and F. Fuentes-Hurtado. Long short-term memory prediction of user's locomotion in virtual reality. *Virtual Reality*, 28(1):1–12, 2024. doi: 10.1007/s10055-024-00962-9 2

[39] Y. K. Meena, H. Cecotti, K. Wong-Lin, and G. Prasad. A multimodal interface to resolve the midas-touch problem in gaze controlled wheelchair. In *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 905–908. IEEE, 2017. 2

[40] C. Nadeau and Y. Bengio. Inference for the generalization error. *Machine Learning*, 52(3):239–281, 2003. doi: 10.1023/A:1024068626366 4

[41] C. C. Singer and B. Hartmann. See-thru: Towards minimally obstructive eye-controlled wheelchair interfaces. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility*, pp. 459–469, 2019. 2

[42] S. Soltani and A. Mahnam. A practical efficient human computer interface based on saccadic eye movements for people with disabilities. *Computers in biology and medicine*, 70:163–173, 2016. 1

[43] N. Stein, G. Bremer, and M. Lappe. Eye tracking-based lstm for locomotion prediction in vr. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 493–503. IEEE, IEEE, Christchurch, New Zealand, 2022. doi: 10.1109/VR51125.2022.00069 1, 2

[44] S. Stellmach and R. Dachselt. Designing gaze-based user interfaces for steering in virtual environments. In *Proceedings of the symposium on eye tracking research and applications*, pp. 131–138, 2012. 2

[45] M. Subramanian, S. Park, P. Orlov, A. Shafti, and A. A. Faisal. Gaze-contingent decoding of human navigation intention on an autonomous wheelchair platform. In *2021 10th International IEEE/EMBS Conference on Neural Engineering (NER)*, pp. 335–338. IEEE, 2021. 2, 3, 9

[46] B. M. 't Hart and W. Einhauser. Mind the step: complementary effects of an implicit task on eye and head movements in real-life gaze allocation. *Experimental Brain Research*, 223(2):233–249, 2012. doi: 10.1007/s00221-012-3254-x 1

[47] J. Thota, P. Vangali, and X. Yang. Prototyping an autonomous eye-controlled system (aecs) using raspberry-pi on wheelchairs. *International Journal of Computer Applications*, 158(8):1–7, 2017. 1, 2

[48] 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*, 30:2–10, 2017. doi: 10.48550/arXiv.1706.03762 2

[49] B. Velichkovsky, A. Sprenger, and P. Unema. Towards gaze-mediated interaction: Collecting solutions of the "midas touch problem". In *Human-Computer Interaction INTERACT'97: IFIP TC13 International Conference on Human-Computer Interaction, 14th–18th July 1997, Sydney, Australia*, pp. 509–516. Springer, Springer, Sydney, NSW, Australia, 1997. 1

[50] J. Wiener, O. De Condappa, and C. Holscher. Do you have to look where you go? gaze behaviour during spatial decision making. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 33, pp. 1583–1588. Cognitive Science Society, Boston, MA, USA, 2011. 1

[51] J. Xu, Z. Huang, L. Liu, X. Li, and K. Wei. Eye-gaze controlled wheelchair based on deep learning. *Sensors*, 23(13):6239, 2023. 2

[52] C. Yu, X. Ma, J. Ren, H. Zhao, and S. Yi. Spatio-temporal graph transformer networks for pedestrian trajectory prediction. In *European Conference on Computer Vision*, pp. 507–523. Springer, Springer, Glasgow, United Kingdom, 2020. 2

[53] M. Zank and A. Kunz. Eye tracking for locomotion prediction in redirected walking. In *2016 IEEE Symposium on 3D User Interfaces (3DUI)*, pp. 49–58. IEEE, IEEE, Greenville, SC, USA, 2016. 1

[54] Y. Zheng, Q. Li, Y. Chen, X. Xie, and W.-Y. Ma. Understanding mobility based on gps data. In *Proceedings of the 10th international conference on Ubiquitous computing*, pp. 312–321. Association for Computing Machinery, Seoul, Korea, 2008. doi: 10.1145/1409635.1409677 4, 6

[55] Y. Zheng, X. Xie, W.-Y. Ma, et al. Geolife: A collaborative social networking service among user, location and trajectory. *IEEE Data Eng. Bull.*, 33(2):32–39, 2010. 4, 6

[56] Y. Zheng, L. Zhang, X. Xie, and W.-Y. Ma. Mining interesting locations and travel sequences from gps trajectories. In *Proceedings of the 18th international conference on World wide web*, pp. 791–800. Association for Computing Machinery, Madrid, Spain, 2009. doi: 10.1145/1526709.1526816 4, 6