Modeling Driving and Visual Perception through Computational Rationality

Alexander Lingler
Intelligent User Interfaces Group
Interdisciplinary Transformation University
Linz, Austria
alexander.lingler@it-u.at

Martin Lorenz
ScaDS.AI
Leipzig University
Leipzig, Germany
lorenz@cs.uni-leipzig.de

Arezoo Sedghi
GeoSocial Artificial Intelligence Group
Interdisciplinary Transformation University
Linz, Austria
arezoo.sedghi@it-u.at

Helena Anna Frijns
Intelligent User Interfaces Group
Interdisciplinary Transformation University
Linz, Austria
helena-anna.frijns@it-u.at

Bernd Resch
GeoSocial Artificial Intelligence Group
Interdisciplinary Transformation University
Linz, Austria
bernd.resch@it-u.at

Patrick Ebel
Computational Interaction Group
Hasso Plattner Institute, University of Potsdam
Potsdam, Germany
patrick.ebel@hpi.de

Philipp Wintersberger
Intelligent User Interfaces Group
Interdisciplinary Transformation University
Linz, Austria
philipp.wintersberger@it-u.at

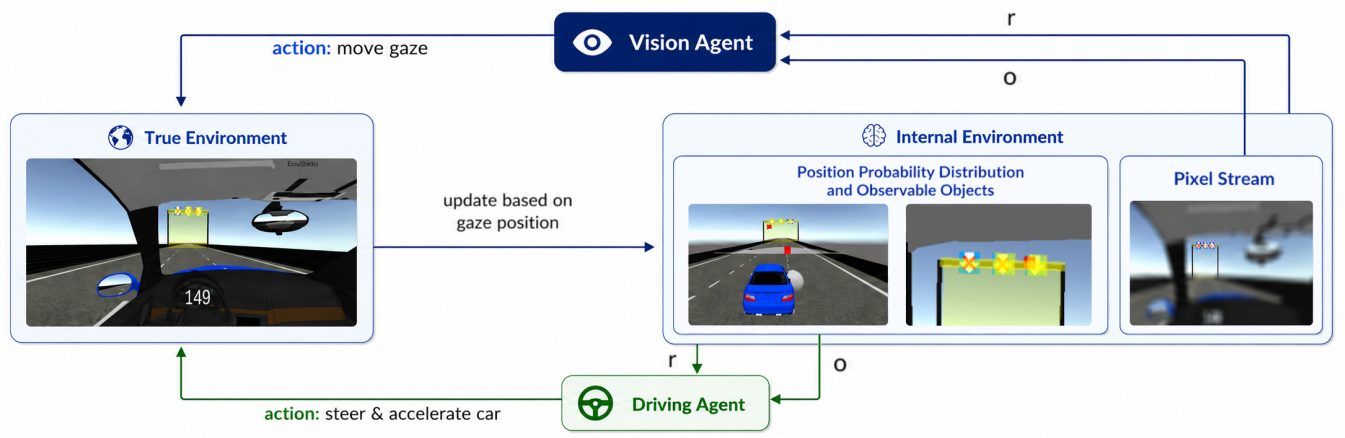


Figure 1: We propose a computational-rational dual-agent model with both a driving and a vision agent, implemented in a dynamic Unity scene. The vision agent can observe the true environment, and updates the probability distribution over the car's future position, which is used by the driving agent. Moreover, the vision agent updates the target lane and target speed based on observation. For the vision agent, the pixels in the foveal region are sharp, while the pixels outside are blurred.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

AutomotiveUI '26, Gothenburg, Sweden

© 2026 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-2814-3/26/09
https://doi.org/10.1145/3828157.3828783

Abstract

Drivers' attention allocation plays a critical role in traffic safety, as driving requires the coordination of attention and control under uncertainty. Thus, predicting driving and gaze behavior is essential for automotive interfaces and intelligent vehicle systems. However, prior models typically treat gaze and driving behavior in isolation,

failing to capture their interdependence. Consequently, we present a computational-rational model that jointly captures visual scanning and driving. A vision agent allocates gaze to update a belief state, which informs a driving agent controlling driver behavior. The model is based on first principles, modeling vision and eye-movement timing. We evaluated the model in a lane-change task with eye tracking, comparing agent and human behavior. Results show that the model attends to task-relevant regions and shows driving behavior closer to human trajectories when relevant information is sampled. This work contributes a unified, interpretable framework for modeling perception-action coupling in driving under partial observability.

CCS Concepts

• **Human-centered computing** → **HCI theory, concepts and models**; *Laboratory experiments*; • **Computing methodologies** → *Agent / discrete models*; • **Theory of computation** → *Reinforcement learning*.

Keywords

Machine Learning, Computational rationality, Driving, Visual attention, Eye tracking, Lab study

ACM Reference Format:
Alexander Lingler, Martin Lorenz, Arezoo Sedghi, Helena Anna Frijns, Bernd Resch, Patrick Ebel, and Philipp Wintersberger. 2026. Modeling Driving and Visual Perception through Computational Rationality. In *18th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '26), September 20–23, 2026, Gothenburg, Sweden.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3828157.3828783

1 Introduction

Drivers' visual attention allocation plays a critical role in traffic safety. Research has shown that while engaging in cognitively demanding secondary tasks, drivers also adjust their gaze behavior [32]. However, taking the eyes off the road while driving significantly increases crash risk [4, 17]. Furthermore, visual perception can be seen as the primary gateway through which situation awareness [7] is formed and maintained. In driving, most task-relevant information is acquired visually, making perception the foundation for understanding the current situation and anticipating how it will evolve. Accordingly, the ability to model and predict both drivers' gaze and their resulting driving behavior presents various opportunities for (1) better understanding driver behavior and for (2) improving automotive user interfaces and intelligent vehicle systems, making them not only safer but also more usable [5].

First, predictive models can fundamentally change how automotive user interfaces are designed and evaluated. Today, new interface concepts are predominantly assessed through empirical user studies, which are costly, time-consuming, and difficult to scale across the vast design space of modern vehicles [6]. In contrast, models of driver behavior, visual attention allocation, and multitasking open up the possibility of rapid, low-cost evaluation early in the design process [26]. Accordingly, such computational models can serve as an effective complement to user studies by enabling iterative prototyping, narrowing down design alternatives, and informing more targeted and efficient empirical investigations [30].

Second, predictive models of driver and gaze behavior could be a key enabler for the next generation of intelligent vehicle systems that collaborate seamlessly with their users. Rather than treating driver behavior as static, systems that incorporate an accurate model of the human driver can leverage behavioral predictions to anticipate driver needs, adapt to individual driving styles [39], and time interactions in a context-sensitive manner [35]. For example, notifications and warnings could be scheduled based on the driver's current attentional state, or assistance systems could proactively compensate for anticipated lapses in attention [40, 41]. Such capabilities are central to designing vehicles that are not only safer, but also more intuitive, less intrusive, and ultimately more aligned with human capabilities and limitations.

Data-driven approaches, such as models predicting the visual demand of in-vehicle interactions or driver distraction from vehicle data, have shown promise for evaluating automotive user interfaces and predicting driver behavior and multitasking [5, 15]. However, their utility can be limited by reduced interpretability and dependence on large, task-specific datasets, which may constrain generalizability across novel design conditions [26].

In contrast, computational rationality (CR) [9, 31] offers a principled alternative by modeling human behavior as the optimal behavior under environmental, task, and cognitive constraints. In practice, this entails specifying the task goals, available actions, uncertainty, and human-like bounds on perception, motor control or memory, and deriving a policy that predicts behavior under these constraints. This first-principles approach provides both explanatory power and predictive capability, making it particularly attractive for modeling human interaction behavior [31]. While CR has led to successful models of visual attention and decision-making in HCI [10, 27, 36, 42], its application to driving remains limited. Notably, existing approaches are based on abstract representations of the driving environment [11, 12] and treat vehicle control and eye movement behavior in isolation, failing to capture how drivers actively coordinate perception and action in a closed-loop, dynamic environment under uncertainty.

In this work, we address this gap by introducing a CR model that jointly accounts for visual attention and vehicle control in a highway driving task. We propose a dual-agent formulation in which a driving agent and a vision agent are coupled: the driving agent selects driving actions under partial observability based on a belief state, while the vision agent actively updates this belief through gaze allocations over a dynamically changing driving scene. By integrating foveated vision and biologically grounded models of eye movements, our approach captures key constraints of human visual processing. We evaluated the model in a controlled lane-switching task with eye tracking, comparing both driving performance and gaze behavior between agents and human participants. Through this, we demonstrate how computational rationality can serve as a unifying framework for modeling the interplay between perception and action in driving. Specifically, we make the following contributions:

- We implemented the Daimler lane change task [29] in an endless highway simulation in Unity that supports configurable path lengths, road signs, target speeds, and target lanes. The environment provides task-relevant measures,

such as the distance to the next target speed and target lane. It integrates the computational-rational models described later, and is publicly available on GitHub[1].
- We introduce a computational-rational dual-agent driving model in which a vision agent actively controls gaze over a dynamically changing first-person pixel stream with foveated and peripheral processing and EMMA-based eye-movement timing. Based on these observations, the vision agent updates the belief state of a separate driving agent responsible for steering, braking, and acceleration.
- We conducted a lab-based user study to compare the simulated behavior with human behavior in the same driving task, using measures for both driving performance and gaze behavior derived from eye-tracking data.
- We provide a human-centered evaluation of the proposed CR model by comparing not only task performance but also visual attention behavior, including gaze allocation, information uptake, and road-monitoring behavior.

2 Related Work

In contrast to traditional approaches such as EPIC [16] or ACT-R [2], CR [31] is an approach that does not rely on hand-crafted production rules to model human behavior, but utilizes the recent advancements in reinforcement learning (RL) to learn a human-like behavioral policy by modeling perceptual and cognitive *bounds*, such as peripheral and foveated vision, or limited memory. Therefore, the tasks are formulated as sequential decision making problems under constraints using a partially observable Markov decision process (POMDP). Several CR models have been proposed in the literature, notably on vision tasks and driving tasks that also involve vision.

2.1 Computational Models of Vision

Shi et al. [36] model human eye movements to guide information acquisition for static charts based on the theory of CR. The authors employ a two-level hierarchical control architecture. At the top-level, LLMs are used to comprehend the collected visual information in light of task relevance and to select a goal for the lower-level controllers. These lower-level controllers subsequently execute eye movements based on a sampling policy learned via RL. The authors find that the model successfully predicts human-like scan paths for task-driven chart reading scenarios. Another CR model that employs vision is *CR-Typist* by Shi et al. [37], which models touchscreen typing, including proof-reading, error correction, and accurate keystrokes. The architecture comprises three components: a vision module, a finger module, and working memory. To account for human-like vision, the vision module simulates peripheral and foveated vision, rather than processing the entire screen state. Concurrently, the finger module moves the finger given human constraints and performs keystrokes. The working memory maintains a belief state of the text typed thus far, integrating information from the vision and finger module. Based on the belief about the working memory state, an external supervisor sets goals for the vision and finger modules to correctly type the target phrase. Shi et al. [37] find that *CR-Typist* not only replicates human typing and proof-reading behavior but also generalizes to new keyboard layouts.

CR-Eyes is a CR model by Lorenz et al. [27] that simulates human visual sampling and gameplay behavior in Atari games. The model works on raw pixel inputs and integrates the EMMA model [33] and limited memory capacity as human bounds. *CR-Eyes* treats eye movements as goal-directed actions to perform better in the Atari games rather than as isolated saliency predictions. While the model achieves human-like game scores and saliency-maps, the resulting scanpaths differ from human players.

2.2 Computational Models of Driving and Vision

Lorenz et al. [26] performed a systematic literature review of computational models for in-vehicle user interface design, focusing on works that develop or evaluate a computational model that quantifies engagement with a secondary task. Target values for prediction include vision (glance behavior), driving (lateral and longitudinal control, crash risk), non-driving related tasks, and mental aspects (distraction, workload) [26].

Jokinen and Kujala [13] model adaptation of a driver's multitasking behavior to a driving assistance system, and investigate predictions made by the model. The primary task in this model is lane keeping at a fixed driving speed. Lateral road positions are part of the state space, and are modeled using discrete bins. The driver's belief regarding the lateral position is modeled as a belief distribution over the possible lateral positions. A lane-keeping assistance system is used that ensures the car stays in its lane. A secondary task, in-car visual search, is also modeled: a supervisor chooses between making observations for the driving task or the visual search task. To test the model, they turn off lane assistance during testing, for several in-car UI sizes [13]. The model that their work was based on (multitasking driving and visual search) [12] was tested against human data, in a within-subjects experiment (N=12) where participants interacted with the system at two different driving speeds and performed visual search tasks while wearing an eye tracker. The difference between our work and that of [12, 13] is that in those works, there is no visual processing of the road, but rather a probability of sampling the true lane position when attention is on the road (where the supervisor gives attention either to the lane-keeping task vs. the visual search task, like an on/off switch). In contrast, in the present work, there is visual processing of the view of the road.

Kujala and Salvucci [20] proposed an ACT-R-based cognitive model [2] to simulate visual attention allocation during multitasking while driving. The authors framed the scenario as an interleaving problem, where the primary driving task is interleaved with a visual search task on an in-car display. Handcrafted production rules were evaluated in a user study (N=12) in a mixed reality driving simulator. The primary driving task consisted of lane keeping on a straight three lane highway with a maximum speed of $80 km/h$. The developed model was able to predict human gaze metrics, especially mean glance duration and the total number of in-car glances.

Lee [21] integrated a saliency map with Distract-R, an ACT-R variant for driving, to account for both bottom-up visual features

[1] https://github.com/reapphil/AITentiveMultitasking

and top-down task influences when predicting drivers' glance behavior toward in-vehicle displays.

To summarize, existing CR vision models address human-like sampling in isolated or static tasks, and are not integrated with and continuously acting in fast-paced, dynamically changing environments.

3 A Computational-Rational Model for Highway Driving

Since prior driving models that incorporate vision rely on abstract or discrete attention mechanisms and do not model continuous visual processing of the driving environment, this work introduces a dual-agent framework with foveated vision and active gaze control, enabling fully closed-loop integration of perception and driving behavior. The proposed CR approach consists of a dual-agent setting in which a vision agent updates the belief state of a driving agent, which in turn controls the car's steering, acceleration, and braking.

3.1 The Task Environment: Adapting Speed and Lane Positions

The environment is implemented in Unity and consists of a straight three-lane road with several street signs indicating either the target speed or target lane. The target speed is indicated using speed-limit signs placed on both sides of the track, and the target lane via overhead signs with arrows pointing to the corresponding lane (as defined in the lane change task [29]). The number of street signs, their spacing, and the total length of the track can be configured. To enable realistic vehicle physics, we use the *Realistic Car Controller* asset from the Unity Asset Store. Figure 1 shows the in-car camera view, which is accessible to both the vision agent and the participants of the user study. The maximum possible speed was set to 150 kilometers per hour.

3.2 Partially Observable Markov Decision Processes

In RL, an agent in state s selects an action a, which leads to a transition to a new state s' with probability p and yields a reward r. The agent learns a policy π (with $\sum_{a\in A}\pi(a \mid s) = 1$) that maximizes the expected return. In case the agent has access to the true environment state, the RL-problem can be described as a fully observable MDP (S, A, P, R, γ), where the discount factor $\gamma \in [0, 1]$ determines how strongly future rewards are weighted. *Throughout, uppercase symbols (e.g., S, A, P, R, Ω) denote the corresponding sets of elements—states, actions, transition probabilities, rewards, and observations, respectively (analogously for the POMDP tuple)—while γ is a scalar discount factor.*

Value-based RL estimates the state-value function using the Bellman expectation equation:

$$v_\pi(s) = \sum_{a\in A}\pi(a \mid s)\sum_{s'\in S}\sum_{r\in R}p(s', r \mid s, a)\left[r + \gamma\, v_\pi(s')\right] \quad \forall s \in S. \tag{1}$$

Computing $v_\pi(s)$ requires recursive evaluation over successor states and rewards. Since the state space considered in this work is continuous, function approximation methods (e.g., neural networks) are used to estimate the value function and derive actions within this space.

In contrast to the fully observable MDP, the driving agent introduced in this work cannot fully observe the environment. Only the portion currently attended by the vision agent is observable and updated, while the remaining parts remain uncertain, which may negatively affect action selection. In our driving task, the concrete state spaces are instantiated in Section 3.3: the driving state primarily represents observed target speed, target lane, and a belief over the vehicle's projected lateral position, while the vision state represents the foveated camera input and belief-related information. Such situations, in which the true state is not known to the agents, are naturally modeled as a POMDP $(S, A, P, R, \gamma, H, \Omega, O)$, where H denotes the horizon, Ω the observation set, and O the observation function. $O(s', o, a)$ specifies the probability of observing o in state s' after taking action a (computed here using a fixed parameter o_p, see Formula 3).

Because the true state is not directly accessible, decisions are made based on a belief state $b(s)$, representing a probability distribution over S. The belief state is updated according to:

$$b'(s') \propto O(s', o \mid s, a)\sum_{s\in S}p(s' \mid a, s)\, b(s), \tag{2}$$

where the proportionality constant ensures that $\sum_{s'} b'(s') = 1$. Consequently, policies in the POMDP are defined over belief states b rather than the true states s.

3.3 Driving and Vision Agents

The CR environment consists of a dual-agent setup, in which the driving agent is responsible for the vehicle's acceleration, braking, and steering, while the vision agent updates the driving agent's belief state. The CR environment incorporates the following human-inspired perceptual constraints and internal-state assumptions:

- The internal representation of the car position is updated based on the point the vision agent is fixating.
- Internal representation of the last observed target lane and target speed.
- Eye movement over a pixel stream, following the EMMA model, with:
  - Foveated vision allowing perception of all objects inside the fovea.
  - Peripheral vision, where the area outside the fovea is provided as blurred, and objects in this area cannot be directly perceived.

3.3.1 Driving Agent. The driving agent κ_d, where d denotes driving, can be modeled as a POMDP, where the belief state $b(s_d)$ represents a probability distribution over the driving state space S_d encoding the probability associated with the lateral position of the car x on the lane, where $x \in [-13, 13]$, implying a total street width of 26 meters. To represent this belief state, the road width is discretized into 40 bins. Each bin corresponds to a possible future/projected lateral position and is assigned a probability indicating the likelihood that the car is located at that position. Here, x does not denote the vehicle's current physical position on the road. Instead, it denotes the projected position the vehicle would

reach along the lane if it continued driving at its current velocity until reaching the point currently fixated by the vision agent. Thus, the x-axis is defined relative to the vision agent's current gaze target and the vehicle's present speed. For instance, if the vision agent is observing a point 10 meters ahead, is currently at position $x = 5$, and has a velocity of $(1, 5)$, where the first component denotes lateral motion and the second longitudinal motion, then the vehicle would reach the observed point after $2s$, during which the lateral position would change by $2m$; the projected lateral position x is therefore 7. Using this projected position instead of the instantaneous position reduces jittering, because actions are evaluated based on their future effect: strong corrective movements in opposite directions may cancel out in the current position, but produce large deviations in the projected position (potentially even leaving the lane), thereby discouraging such oscillatory behavior. We assume that this formulation reflects human driving behavior, as drivers typically look ahead and act based on where they expect the car to be, rather than only reacting to its current position.

The transition probabilities p (Eq. 2) are based on a normal distribution of the car's velocity with a fixed standard deviation σ_P similar to that in our previous work [25]. The mean is sampled from a second normal distribution with a standard deviation σ_{mean}, centered on the estimated velocity. If the vision agent directs its gaze toward the road, the estimated velocity equals the true velocity. Otherwise, it is iteratively updated using samples from the second distribution, causing uncertainty to increase initially before stabilizing as σ_{mean} is halved at each update step.

The observation probability O (Eq. 2) is defined using a fixed parameter o_p:

$$O_d(s', o \mid s, a) = \begin{cases} o_p, & \text{if } s' = s_{\text{true}} \text{ and gaze is on the road,} \\ \dfrac{1 - o_p}{\#S - 1}, & \text{if } s' \neq s_{\text{true}} \text{ and gaze is on the road,} \\ 1, & \text{if gaze is not on the road.} \end{cases} \tag{3}$$

Here, $0 \leq o_p \leq 1$. If the vision agent has its gaze on the road and o_p is sufficiently large, the belief quickly converges to the true state. Otherwise, no observation is made and all states are equally likely. The belief state is updated at fixed intervals u, computed in parallel using the *Unity C# Job System* and the *Burst Compiler* [8].

S_d additionally includes the last observed target lane and target speed, the car's speed and engine Revolutions Per Minute (RPM) (as integers), the distance to the next speed and target lane sign, the entropy of the car position distribution, and the estimated velocity. The estimated velocity equals the true velocity when the vision agent focuses on the road; otherwise, it is sampled from the normal distribution used to update the car's belief position.

The action $a \in A_d$ is represented as a two-dimensional continuous vector, where the first component corresponds to steering input and the second to longitudinal control, indicating acceleration when positive and braking when negative.

The driving reward r_d is defined as

$$r_d = r_{\text{lane}} \cdot r_{\text{speed}}, \tag{4}$$

with $r_{\text{lane}} \in [0, 1]$ and $r_{\text{speed}} \in [0, 1]$. Consequently, $r_d \in [0, 1]$. Due to the multiplicative form, both components jointly influence the reward, preventing optimization of only one term. The speed reward r_{speed} is defined as one minus the normalized absolute distance between the current speed and the target speed, scaled by the maximum possible deviation. The lane reward r_{lane} is given by

$$r_{\text{lane}}(d, \sigma_d) = \exp\left(-\frac{d^2}{2\sigma_d^2}\right), \tag{5}$$

where we set $\sigma_d = 2$, and d denotes the distance between the car's projected position and the target lane position. This Gaussian formulation yields a nonlinear reward with respect to the lane deviation.

3.3.2 Vision Agent. The vision agent κ_v, where v denotes vision, updates the driving agent's belief state by directing its gaze over a $160 \times 90 \times 3$ pixel input representing the driver's camera view including color information. It is modeled as a MDP and pursues four objectives:

(1) Update the target speed by detecting the next speed sign,
(2) Update the target lane by detecting the next lane sign,
(3) Update the car's belief position by observing the road,
(4) Determine the projection distance, defined as how far the agent looks into the distance.

s_v therefore includes the described pixel stream, where only the foveal region is sharp and the periphery is blurred using a shader. The blur increases with the distance from the gaze position and is computed via a box blur, with the final image obtained by interpolating between sharp and blurred regions. Furthermore, s_v includes the belief car position and the entropy of its distribution.

The action $a \in A_v$ is represented as a two-dimensional continuous vector, where x and y denote the target screen coordinates with $x \in [0, 1]$ and $y \in [0, 1]$, defined relative to the screen.

The gaze movement time is computed using the EMMA model. The encoding time of a visual object is given by

$$T_{\text{enc}} = K \cdot [-\log(f)] \cdot k^{k\epsilon}, \tag{6}$$

where K and k are constants, f is the object frequency, and ϵ is the eccentricity.

A saccade takes

$$T_s = t_{\text{prep}} + t_{\text{exec}} + D \cdot t_{\text{sacc}}, \tag{7}$$

where t_{prep}, t_{exec}, and t_{sacc} are constants, and D is the saccade distance.

If $T_{\text{enc}} < t_{\text{prep}}$, encoding occurs without eye movement; otherwise, a saccade reduces eccentricity before encoding completes.

Based on the position of the gaze, all Unity GameObjects that are within the fovea are returned and can be interpreted as seen. So for instance, in case the agent looks at a speed-limit sign the corresponding target speed is updated according to the speed of the speed-limit sign. The peripheral vision including color information allows the agent to detect patterns associated with a certain street sign, i.e., the speed-limit sign can be identified by the red circle around the number indicating the speed limit.

The reward r_v is defined as follows:

$$r_v = \begin{cases} r_d \cdot v & \text{if } r_d \text{ is positive,} \\ r_d \cdot (1 - v) & \text{otherwise} \end{cases} \tag{8}$$

where v represents the vision performance. v approximates how much the vision agent contributes to the reward r_d observed by the driving agent and is defined as the sum of three components:

$$v = \frac{1 - \text{entropy}_{01}}{2} + 0.25 \cdot \mathbb{I}_{\text{lane}} + 0.25 \cdot \mathbb{I}_{\text{speed}}, \tag{9}$$

where $\text{entropy}_{01} \in [0, 1]$ denotes the normalized entropy of the car position belief, and $\mathbb{I}_{\text{lane}}, \mathbb{I}_{\text{speed}} \in \{0, 1\}$ are indicator variables that equal 1 if the current or next target lane or speed has been observed, respectively, and 0 otherwise. Thus, $v \in [0, 1]$ increases with more certain car position estimates and successful observations of relevant traffic information.

4 Training

The CR agents were trained with Proximal Policy Optimization (PPO) [34] using Unity ML-Agents. Training was conducted in two stages. First, we performed hyperparameter optimization (HPO) using Ray Tune [23], evaluating 282 configurations for up to 2 million training steps per configuration. The search used Optuna for sampling [1] and ASHA for early stopping [22]. We optimized the shared reward of the coupled agent setup, as task success depends on both agents: the vision agent has to sample task-relevant information, while the driving agent has to use the resulting belief state for lane and speed control. The complete search space, training scripts, and configuration files are available in our GitHub repository.

After HPO, we selected the best-performing policy configuration and used it for the final training run. The final model was trained with parallel Unity environments for 40 million vision-agent steps and 200 million driving-agent steps. During the training run, cumulative reward increased in early training and then stabilized, indicating convergence of the selected policy. The internal CR-agent metrics showed a consistent pattern: belief entropy decreased over training, while position awareness and vision performance increased and subsequently stabilized. These training curves indicate that the reported behavior reflects a stable learned policy selected through HPO.

5 User Study

For evaluation, we conducted a lab study in which participants drove through a predefined driving sequence specified by the number of road signs, their intervals, and the distances between the values of the speed limits. We used the collected data to evaluate how closely the behavior of the driving and vision agents matches that of human participants. We investigated the following aspects:

- **Gaze behavior.** We compared human and agent gaze behavior by constructing phase-specific spatial gaze distributions from normalized gaze coordinates and visualizing them as heatmaps over the driving scene. To quantify similarity between human and agent attention patterns, we compared the resulting gaze distributions using Pearson correlation, Jensen–Shannon divergence and Earth Mover's Distance. By computing heatmaps separately for distance-based phases relative to the next relevant sign, the analysis captures when sign-directed gaze begins to emerge instead of only averaging over the entire approach.
- **Driving performance.** In addition to gaze behavior, we compared overall task performance, including adherence to target speed and lane, to relate differences in visual attention to driving outcomes.

5.1 Participants

Overall, $N = 20$ participants completed the experiment, including 7 female and 13 male participants, aged between 23 and 51 years ($M = 35.30$, $SD = 8.87$ years). For four participants, no eye-tracking data were available due to technical issues. Because the analyses relied on synchronized eye-tracking and driving-performance data, these participants could not be included in the analyses, resulting in a final analytic sample of $n = 16$. Participants were mainly recruited from a university setting, with a small number recruited through the researchers' local networks. Participation was open to adults aged 18 years and older. Inclusion criteria required possession of a valid driver's license and normal or corrected-to-normal vision. All participant data were pseudo-anonymized using randomly generated alphanumeric codes.

5.2 Procedure

For the driving simulator, a Logitech PRO RACING WHEEL 11 and Logitech PRO RACING PEDALS were used. The steering wheel was positioned 100 cm in front of a 4K 65-inch TV (Hisense 65E77NQ). The participants could adjust the driving seat to suit their proportions, with the center of the seat located between 130 and 155 cm from the screen. The simulated vehicle used an automatic transmission. At the beginning of the session, participants received information about the study and provided informed consent by signing a form. After this, participants were equipped with Pupil Labs Neon Glasses, and the eye tracker was calibrated individually for each participant.

Participants then completed a short familiarization phase during which they received verbal instructions about the task. Following the definition of the lane change task, they were asked to follow the speed and lane information shown by the road signs and to adapt their driving only after passing the corresponding sign. During the task, participants drove on a straight, three-lane highway and continuously monitored the scene for relevant road signs to adjust their driving behavior. Driving data and event markers were synchronized through iMotions. Eye-tracking data exported from Pupil Labs were then time-aligned with the iMotions recordings, and the synchronized data streams were merged into a common dataset for subsequent analysis.

5.3 Measurements

Driving data were sampled at 50 Hz and included distance to target position, speed deviation, and lateral velocity as an indicator of swerving. Distance to the target position was defined as absolute lateral deviation between the vehicle center and a piecewise linear target lane-change path. This path remained in the previous lane until one second of travel after the overhead sign, transitioned linearly to the target lane over the following 3.6 seconds of travel, and then continued in the target lane. Speed deviation was defined as the difference between the current speed and the target speed.

Eye movements were mapped to the screen and represented as normalized two-dimensional gaze coordinates. From these, we constructed phase-specific gaze heatmaps for each participant and for the CR agent (Figure 2). Phases were defined by distance to the next relevant sign. For **speed signs**, *near*, *mid*, and *far* corresponded to distances of $< 30m$, $30–120m$, and $> 120m$, respectively. For **lane signs**, the corresponding ranges were $< 60m$, $60–200m$, and $> 200m$. Agent–human similarity was computed by comparing each participant heatmap to the agent heatmap, and a human–human baseline was obtained by averaging comparisons between each participant and all others within the same phase. Similarity was quantified using three complementary measures, to capture different aspects of spatial gaze similarity. Pearson correlation captures similarity in the overall spatial pattern of gaze allocation across the heatmap, but is less informative about whether two distributions differ in their overall probability mass or are spatially shifted relative to one another. Therefore, we additionally computed Jensen–Shannon divergence to quantify similarity between normalized gaze distributions and Earth Mover's Distance to quantify the spatial displacement between distributions, i.e., how far gaze mass would have to be shifted to transform one heatmap into the other.

To assess driving adaptation after passing a speed or target lane sign, we analyzed driving performance trajectories over a 5 s window following each sign, with the event of passing the sign by the vehicle being defined as the moment of minimum absolute distance to the next speed or lane sign. For speed signs, driving performance was measured as absolute target-speed deviation; for lane signs, as the absolute lateral deviation between the vehicle center and the closest point on the piecewise linear target lane-change path. CR-agent events were further separated by whether the upcoming sign had been observed before passage, yielding three conditions per task: humans, observed-agent events, and not-observed-agent events. Trajectories were summarized using post-passage Area Under Curve (AUC), final error, correlation with the human mean trajectory, and Root Mean Square Error (RMSE).

As an additional driving-performance analysis, we computed aggregate lane- and speed-control measures for observed and not-observed samples. For lane control, we first computed the absolute lateral deviation between the current vehicle position and the target lane-change path,

$$d_i = \left| x_i^{\text{target}} - x_i^{\text{current}} \right|.$$

To relate this measure to established lane-position measures [18, 40], we report Mean Lateral Position (MLP) and Standard Deviation of Lateral Position (SDLP) as

$$MLP_c = \frac{1}{N_c} \sum_{i \in c} d_i$$

and

$$SDLP_c = \sqrt{\frac{1}{N_c} \sum_{i \in c} (d_i - d_{\text{avg},c})^2},$$

where $d_{\text{avg},c} = MLP_c$. In our case, d_i was not defined as the distance to a lane boundary, but as the absolute lateral deviation between the vehicle center and the target lane-change path.

For speed control, we computed analogous measures based on the absolute deviation between current and target speed,

$$s_i = \left| v_i^{\text{target}} - v_i^{\text{current}} \right|.$$

We refer to these measures as Mean Absolute Speed Deviation (MASD) and Standard Deviation of Absolute Speed Deviation (SDASD):

$$MASD_c = \frac{1}{N_c} \sum_{i \in c} s_i$$

and

$$SDASD_c = \sqrt{\frac{1}{N_c} \sum_{i \in c} (s_i - s_{\text{avg},c})^2},$$

where $s_{\text{avg},c} = MASD_c$.

We additionally recorded internal CR-agent metrics aggregated over episodes: **belief entropy**, **visibility**, **position awareness**, and **vision performance**, characterizing uncertainty, observation of car position, belief accuracy, and overall perceptual effectiveness v (see Equation 9).

6 Results

6.1 Gaze Distributions

Table 1 summarizes the mean similarity scores and their standard deviations for the phase-specific spatial gaze distributions of humans and the CR agent in the speed-sign and lane-sign conditions. For each phase, agent-human similarity was compared against a human-human baseline, defined as the average similarity between a given participant and all other participants in the same phase. For the correlation-based gaze analyses, Pearson correlation coefficients were Fisher-z transformed prior to inferential testing to better satisfy the assumptions of parametric tests. Normality of the transformed values and paired difference scores was assessed using Shapiro–Wilk tests. As no significant deviations from normality were detected, we used one-sample t-tests to test whether agent-human correlations were significantly greater than zero and paired-samples t-tests to compare agent-human correlations against the human-human baseline.

Across both sign types, Pearson correlations between agent and human heatmaps were consistently positive but weak, ranging from .126 to .141 for speed signs and from .119 to .166 for lane signs, and remained below the corresponding human-human baseline in all phases (speed: .218–.290; lane: .232–.252). One-sample t-tests on Fisher-z transformed correlations showed that all agent-human correlations were significantly greater than zero for both lane signs (far: $t(15) = 5.98$, $p < .001$; mid: $t(15) = 6.97$, $p < .001$; near: $t(15) = 6.50$, $p < .001$) and speed signs (far: $t(15) = 6.77$, $p < .001$; mid: $t(15) = 4.98$, $p < .001$; near: $t(15) = 5.52$, $p < .001$). At the same time, paired-samples t-tests on Fisher-z transformed correlations confirmed that agent-human similarity was significantly lower than the human-human baseline in every phase for both lane signs (far: $t(15) = -4.80$, $p < .001$; mid: $t(15) = -5.32$, $p < .001$; near: $t(15) = -4.80$, $p < .001$) and speed signs (far: $t(15) = -4.56$, $p < .001$; mid: $t(15) = -4.96$, $p < .001$; near: $t(15) = -3.60$, $p = .003$).

Jensen–Shannon divergence was consistently higher for agent-human than for human-human comparisons, with agent-human values ranging from 0.709 to 0.717 for speed signs and from 0.694

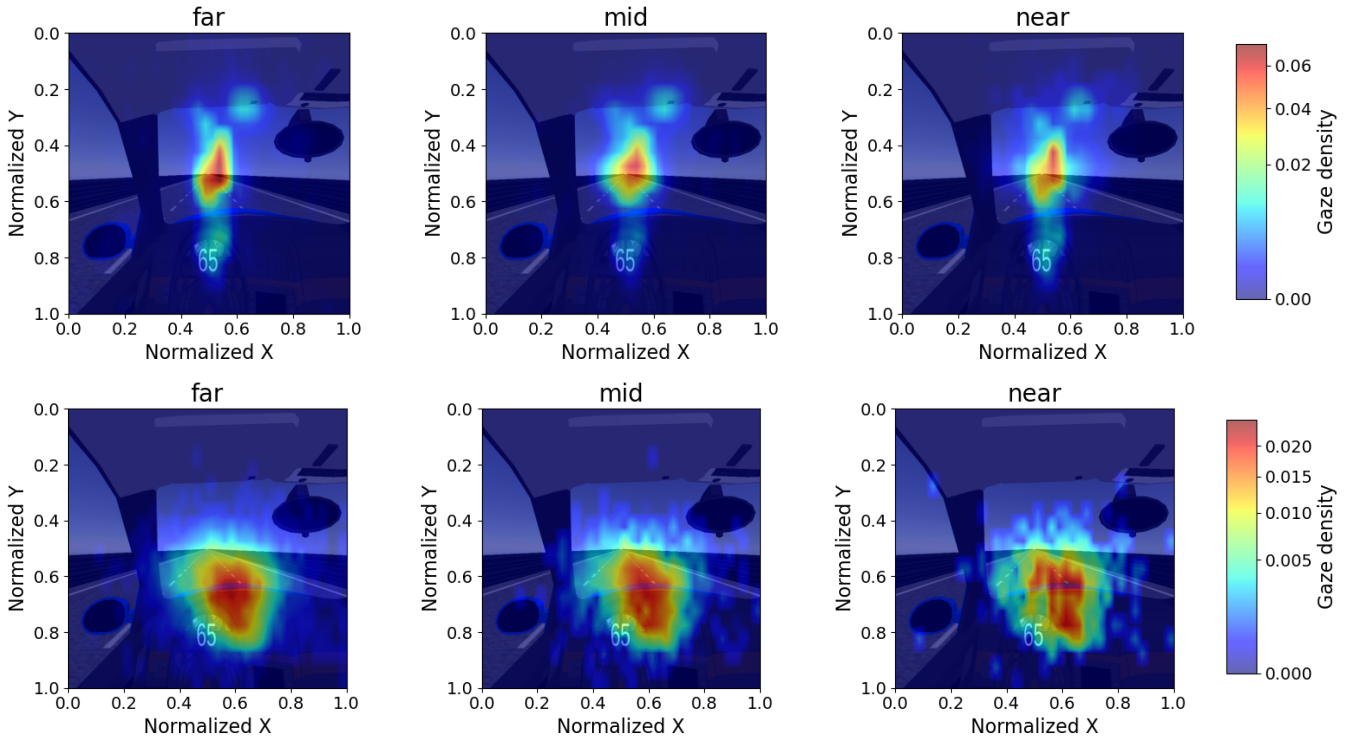


Figure 2: Spatial distribution of gaze during speed-sign approach for human participants (top) and the CR agent (bottom), separated into the phases *far*, *mid*, and *near*. Heatmaps are overlaid on the egocentric driving scene and visualize normalized gaze density. Across all phases, human gaze shows a more clearly bimodal distribution between road monitoring and glances toward the speed display, whereas the CR agent exhibits a broader, unimodal allocation of attention around the road center and sign-relevant region.

to 0.710 for lane signs, compared to 0.585–0.631 and 0.601–0.615 for the corresponding human-human baselines.

The clearest discrepancy was observed for Earth Mover's Distance, which was consistently higher for agent-human than for human-human comparisons across all phases and both sign types (speed: 3.126–3.233 vs. 1.756–1.912; lane: 3.072–3.179 vs. 1.772–1.828). This pattern indicates that, although the CR agent learned to attend to relevant regions of the driving scene, its gaze distributions only partly align with those of human participants.

6.2 Driving Performance

For the driving performance analysis, we compared driving trajectories of humans and the CR agent after passing speed and lane signs (Figure 3). For speed signs, the CR agent produced 801 sign passages in total, of which 354 (44.2%) were classified as observed and 447 as not observed. For lane signs, the CR agent likewise produced 801 sign passages, of which 535 (66.8%) were observed and 266 were not observed.

Table 2 summarizes the resulting similarity and error measures. For **speed signs**, the observed CR-agent condition was closer to the human trajectory than the not-observed condition, with a substantially lower Root Mean Square Error (RMSE) (3.224 vs. 10.048), a smaller post-passage Area Under the Curve (AUC) (73,534.870 vs. 107,345.936), and a lower remaining error at 5 s (12.557 vs. 21.098). However, both agent conditions remained above the human reference trajectory, with an error of 7.656 at 5 s. Correlations with the human trajectory were high for both agent conditions (r = .996 for observed, r = .993 for not observed).

For **lane signs**, the observed CR-agent condition again showed smaller deviations from the human trajectory than the not-observed condition. The observed condition yielded a lower RMSE (0.671 vs. 1.015), a lower error at 5 s (1.966 vs. 3.151), and a smaller AUC (12,391.751 vs. 14,201.874). In addition, its correlation with the human trajectory was strongly positive (r = .953), whereas the not-observed condition showed a negative correlation (r = −.414). Overall, these results indicate that prior observation of the relevant sign was associated with smaller deviations from human driving behavior in both the speed and lane conditions.

As a complementary driving-performance analysis, we computed lane- and speed-control measures at the participant level for humans and at the episode level for the CR agent (see Table 3). Because human participants and agent episodes were not meaningfully paired, we compared the human and agent distributions using independent-samples Mann–Whitney U tests. For lane control, human MLP was 1.958 (SD = 0.479), compared to 1.410 (SD = 0.358) for observed-agent episodes and 4.295 (SD = 1.065)

Table 1: Similarity of phase-specific spatial gaze distributions for speed-sign and lane-sign events. Pearson correlation is higher for more similar heatmaps, whereas Jensen-Shannon divergence and Earth Mover's Distance are lower for more similar heatmaps.

Sign	Phase	Pearson correlation		Jensen-Shannon divergence		Earth Mover's Distance	
		Agent–Human	Human–Human	Agent–Human	Human–Human	Agent–Human	Human–Human
Speed	far	.126 (.07)	.218 (.10)	.709 (.09)	.631 (.12)	3.126 (.75)	1.842 (.79)
Speed	mid	.138 (.11)	.290 (.14)	.717 (.11)	.585 (.15)	3.233 (.81)	1.756 (.80)
Speed	near	.141 (.10)	.233 (.10)	.710 (.12)	.626 (.12)	3.140 (.78)	1.912 (.74)
Lane	far	.119 (.08)	.232 (.11)	.710 (.10)	.615 (.13)	3.179 (.76)	1.828 (.79)
Lane	mid	.135 (.08)	.252 (.12)	.703 (.09)	.609 (.13)	3.108 (.78)	1.773 (.76)
Lane	near	.166 (.10)	.252 (.12)	.694 (.11)	.601 (.13)	3.072 (.79)	1.803 (.77)

for not-observed-agent episodes. Both agent conditions differed significantly from the human distribution (observed: $U = 212.000$, $Z = 3.166$, $p = .001$; not observed: $U = 5.000$, $Z = -4.636$, $p < .001$). Human SDLP was 1.219 ($SD = 0.388$), compared to 1.623 ($SD = 0.453$) in the observed-agent condition and 2.525 ($SD = 0.461$) in the not-observed condition. Again, both differences were significant (observed: $U = 60.000$, $Z = -2.563$, $p = .010$; not observed: $U = 6.000$, $Z = -4.598$, $p < .001$). For speed control, human MASD was 8.563 ($SD = 3.668$), compared to 12.881 ($SD = 3.360$) for observed-agent episodes and 23.950 ($SD = 7.127$) for not-observed-agent episodes, with both differences being significant (observed: $U = 40.000$, $Z = -3.317$, $p = .001$; not observed: $U = 5.000$, $Z = -4.636$, $p < .001$). Human SDASD was 10.657 ($SD = 2.996$), compared to 11.085 ($SD = 3.255$) in the observed-agent condition and 15.130 ($SD = 2.889$) in the not-observed condition. Here, the observed-agent condition did not differ significantly from humans ($U = 119.000$, $Z = -0.339$, $p = .752$), whereas the not-observed condition did ($U = 25.000$, $Z = -3.882$, $p < .001$).

6.3 CR-Agent Behavior

The CR-agent shows moderate perceptual performance with a belief entropy of 0.24, indicating a relatively confident but still uncertain estimate of the car's position. The agent can observe the car 39% of the time, reflecting limited but regular gaze allocation to the road. It correctly identifies its position in 69% of cases, while achieving a vision contribution of 0.73, suggesting effective integration of perception into driving decisions. Overall, the agent demonstrates coordinated but still imperfect perception-action coupling under partial observability.

7 Discussion

Our proposed CR model can partly capture key aspects of the perception-action loop in driving. Because gaze allocation and driving performance reflect different aspects of behavior, we interpret the model's alignment with human data separately for each measure, using the human-human baseline as a reference for inter-individual variability. Under this criterion, the agent did not fully reproduce human gaze behavior: although agent-human correlations were significantly greater than zero and the heatmaps show attention to task-relevant regions, similarity remained consistently

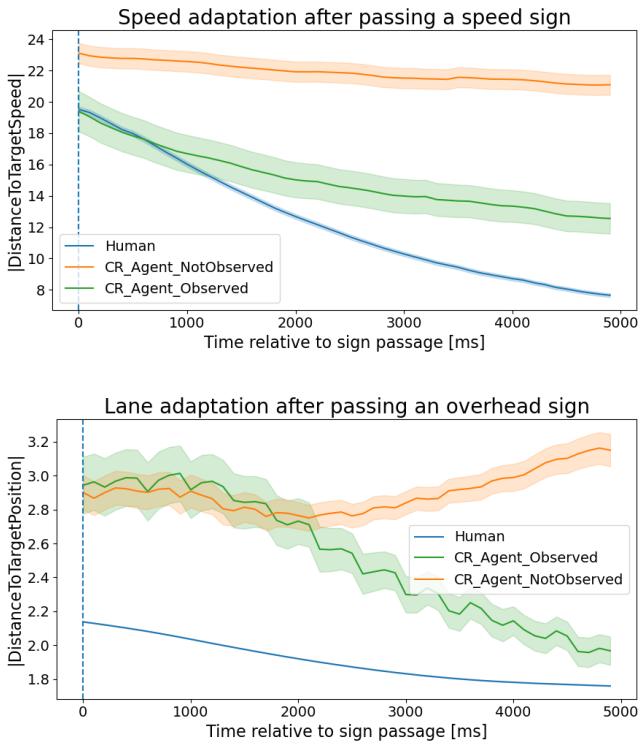


Figure 3: Event-aligned driving adaptation after passing a relevant road sign for humans, CR-agent events in which the upcoming sign was observed, and CR-agent events in which it was not observed. Curves show mean absolute task error relative to sign passage, with speed adaptation shown on the top and lane adaptation shown on the bottom; shaded regions indicate variability across events.

below the human-human baseline. Pearson correlation, Jensen–Shannon divergence, and Earth Mover's Distance led to the same overall conclusion and suggest that the main difference was a spatial shift in attention rather than completely different gaze behavior.

Table 2: Performance similarity for event-aligned driving trajectories in the speed-sign and lane-sign conditions. Lower values indicate better agreement with human performance for post-pass AUC, error at 5 s, and RMSE, whereas higher correlation indicates greater trajectory similarity.

Task	Condition	Post-pass AUC	Error at 5 s	Correlation	RMSE
Speed	Human	60346.402	7.656	1.000	0.000
Speed	Observed	73534.870	12.557	.996	3.224
Speed	Not observed	107345.936	21.098	.993	10.048
Lane	Human	9331.891	1.759	1.000	0.000
Lane	Observed	12391.751	1.966	.953	0.671
Lane	Not observed	14201.874	3.151	-.414	1.015

Table 3: Lane- and speed-control deviations for humans and the CR agent as a function of prior sign observation.

Measure	Human M (SD)	Agent observed M (SD)	Agent not observed M (SD)
MLP	1.958 (0.479)	1.410 (.358)	4.295 (1.065)
SDLP	1.219 (0.388)	1.623 (.453)	2.525 (.461)
MASD	8.563 (3.668)	12.881 (3.360)	23.950 (7.127)
SDASD	10.657 (2.996)	11.085 (3.255)	15.130 (2.889)

Qualitative inspection of the gaze heatmaps supports this interpretation: human attention is bimodally distributed between the horizon and the speedometer, whereas the agent's attention is unimodal and more broadly distributed.

In contrast, driving performance was substantially closer to human behavior when the relevant visual information was successfully scanned. This was reflected primarily in lower RMSE, smaller post-passage AUC, and lower final error for observed compared to not-observed agent events. Trajectory correlations point in the same direction, but should be interpreted with caution because the constrained task structure likely contributes to similarity in the overall curve shape. The additional lane- and speed-control measures further support this interpretation: prior sign observation reduced agent deviations in MLP, SDLP, MASD, and SDASD, and the observed-agent condition was closer to the human distribution than the not-observed condition across all measures. Notably, the observed-agent condition did not differ significantly from humans for SDASD, suggesting that successful visual sampling brought the variability of speed deviations close to the human range for this measure. However, observed-agent behavior still differed significantly from humans for the remaining measures, indicating that successful visual sampling improved driving performance but did not make the agent fully human-like.

Overall, these findings suggest that the model captures the functional role of visual sampling in driving: when task-relevant signs were observed, driving behavior became substantially more human-like, even though gaze allocation and several control measures still differed from humans.

7.1 Modeling Vision, Driving Behavior, or Both?

Our results suggest that perception-action coupling is an important element for modeling driving behavior with CR. Driving performance and alignment with human data varied depending on whether the agent successfully scanned visual information, demonstrating the importance of perception-action coupling. Accordingly, we argue that our approach complements saliency-based models that decouple visual behavior from action [3, 14] and gaze estimation approaches that focus solely on prediction accuracy [43]. Our approach further extends prior CR models in the driving domain that rely on abstract representations of the driving environments and simplified kinematic driving models [13, 20]. Because our approach is implemented in a Unity simulation environment, it allows us to render a visually rich pixel-based driving scene that serves as an input to the model while integrating a high-fidelity kinematic driving model. This allows the model to be easily applied to various other driving environments.

However, our results also show that although the model exhibits human-like driving behavior, its sampling behavior still differs from that of human participants. At present, however, we interpret this pattern primarily as a limitation of the current model. For example, unlike the human participants, the model does not sample the speedometer. This is most likely because the current model can infer the exact driving speed by visually sampling the road, which is not true for humans, whose perception-based speed inference is noisy [38]. They therefore, and in contrast to the model, also need to sample the speedometer to get a more accurate account of the current speed.

Given these considerations, researchers and designers should align model choice with their goals. Traditional saliency models may be the better choice when sufficient training data is available, highly accurate saliency predictions are the priority, and the black-box nature of data-driven models is not a hindrance. CR models such as the one presented here, in turn, may be the better choice when interpretable behavior linking perception and action is key — for example, to answer questions like *"Did driving performance degrade because the driver sampled inefficiently?"* This may be particularly

relevant for training or validating safety systems based on other drivers' reactions.

7.2 Potential Applications

Our work demonstrates that it is possible to model and train agents to learn to perform a driving task involving visual processing of the scene. The agents' driving behavior is responsive to observed road signs in terms of both speed and lane adaptation. This means this model could potentially be extended to other 3D driving environments, learning driving and visual attention allocation from first principles. Our comparably simple lane change tasks did not contain other vehicles and ambiguities, and future work should investigate the dual-agent model in different and more complex environments.

We see various application areas for our model. First, it could benefit driver monitoring and driver assistance systems by giving them a model of human gaze and driving behavior. Eye-tracking data have already been used to dynamically assess situation awareness (SA), for instance, during control takeovers [44]. The dual-agent model could similarly be used to predict where an attentive driver should focus for safe driving. Deviations between predicted and actual gaze may then indicate reduced SA, and an intelligent driver monitoring system may warn the driver. Second, it could support driver training. If the model is parametrized to simulate expert driving behavior, it could inform novice drivers where and how to sample the scene, building on evidence that exposure to experts' eye movements shifts novices' scanning behavior toward expert-like patterns [28]. Third, it could help to train and validate automated driving and safety systems in simulated multi-agent environments, which remains an open challenge [19]. Future models could enable complex accident analysis using counterfactual predictions (e.g., "what if the driver looked at the sign earlier") and model the diverging dynamics of novice and experienced drivers. Fourth, it could support design evaluation of in-vehicle HMIs [6], including assessing how distracting a new infotainment design might be. We see our model as a foundation for simulation approaches that predict how long a driver needs to look away, and in which driving situations such glances lead to severely degraded driving performance [17]. It could likewise inform adaptive systems that better address drivers' cognitive limitations and emerging multitasking demands [24, 25, 41].

7.3 Limitations

The driving agent operates on a hand-crafted task representation rather than directly on raw pixel input, as we model the internal belief state of the driving task using a discretized state space along the x-axis. An open question is whether a pure pixel-based stream could serve as a sufficient state representation. This is likely not the case, since the internal state depends not only on visual perception but also on cognitive factors such as memory and other latent processes. Our work investigates only one type of scene in simulation, namely a straight highway with three lanes without other traffic. Future work should investigate different types of roads and investigate more realistic simulations and eventually real-world scenes.

8 Conclusion

We propose a dual-agent computational-rational model for a driving task that involves visual processing of a dynamic road scene. The scene is implemented as a highway simulation in Unity. The model combines both a vision agent with human-related perceptual constraints and a driving agent. The vision agent observes the true environment and updates the belief state that is used by the driving agent. This belief state contains the target lane, the target speed, and a probability distribution over the car's future position. Eye-tracking data from human study participants ($n = 16$ of $N = 20$) were compared with the behavior of the driving and vision agents. Although agent–human gaze correlations were weak, the dual-agent model produced gaze patterns covering task-relevant regions of the scene that partially aligned with human behavior. Driving performance showed that the dual-agent model was closer to human behavior in those cases where task-relevant visual information was successfully perceived. Our work shows that the dual-agent model is a promising way for agents to learn driving behavior from first principles, learning the driving task based on visual observations of dynamic highway scenes.

Acknowledgments

This project is supported by the Austrian Science Fund (FWF) under grant Nr. P35976-N (AITentive). We thank Anna Palatkina, Elpiniki Prinianaki, and Jonathan Kubik for conducting the driving study and for making the study data available for this work.

References

[1] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A Next-generation Hyperparameter Optimization Framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (Anchorage, AK, USA) *(KDD '19)*. Association for Computing Machinery, New York, NY, USA, 2623–2631. doi:10.1145/3292500.3330701

[2] John R Anderson. 2014. Production systems and the ACT-R theory. In *Rules of the mind*. Psychology Press, New York, NY and Hove, East Sussex, 1–15.

[3] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. 2018. What do different evaluation metrics tell us about saliency models? *IEEE transactions on pattern analysis and machine intelligence* 41, 3 (2018), 740–757.

[4] Thomas A. Dingus, Feng Guo, Suzie Lee, Jonathan F. Antin, Miguel Perez, Mindy Buchanan-King, and Jonathan Hankey. 2016. Driver Crash Risk Factors and Prevalence Evaluation Using Naturalistic Driving Data. *Proceedings of the National Academy of Sciences* 113, 10 (Feb. 2016), 2636–2641. doi:10.1073/pnas.1513271113

[5] Patrick Ebel, Christoph Lingenfelder, and Andreas Vogelsang. 2023. On the Forces of Driver Distraction: Explainable Predictions for the Visual Demand of in-Vehicle Touchscreen Interactions. *Accident Analysis & Prevention* 183 (April 2023), 106956. doi:10.1016/j.aap.2023.106956

[6] Patrick Ebel, Julia Orlovska, Sebastian Hünemeyer, Casper Wickman, Andreas Vogelsang, and Rikard Söderberg. 2021. Automotive UX Design and Data-Driven Development: Narrowing the Gap to Support Practitioners. *Transportation Research Interdisciplinary Perspectives* 11 (Sept. 2021), 100455. doi:10.1016/j.trip.2021.100455

[7] Mica R. Endsley. 1995. Toward a Theory of Situation Awareness in Dynamic Systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 37, 1 (March 1995), 32–64. doi:10.1518/001872095779049543

[8] C Ferreira and M Geig. 2018. Get Started with the Unity* Entity Component System (ECS), C# Job System, and Burst Compiler. *Intel Corporation,[Online]. Available: https://www.intel.com/content/www/us/en/developer/articles/guide/get-started-with-the-unity-entity-component-system-ecs-c-sharp-job-system-and-burst-compiler.html (visited on 2022-11-12)* 1 (2018), 1.

[9] Samuel J Gershman, Eric J Horvitz, and Joshua B Tenenbaum. 2015. Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science* 349, 6245 (2015), 273–278.

[10] Jussi Jokinen, Aditya Acharya, Mohammad Uzair, Xinhui Jiang, and Antti Oulasvirta. 2021. Touchscreen Typing As Optimal Supervisory Control. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 720, 14 pages. doi:10.1145/3411764.3445483

[11] Jussi Jokinen, Patrick Ebel, and Tuomo Kujala. 2025. Predicting Multitasking in Manual and Automated Driving with Optimal Supervisory Control. arXiv:2503.17993 [cs.HC] https://arxiv.org/abs/2503.17993

[12] Jussi PP Jokinen, Tuomo Kujala, and Antti Oulasvirta. 2021. Multitasking in driving as optimal adaptation under uncertainty. *Human factors* 63, 8 (2021), 1324–1341.

[13] Jussi P. P. Jokinen and Tuomo Kujala. 2021. Modelling Drivers' Adaptation to Assistance Systems. In *13th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (Leeds, United Kingdom) *(AutomotiveUI '21)*. Association for Computing Machinery, New York, NY, USA, 12–19. doi:10.1145/3409118.3475150

[14] Tilke Judd, Frédo Durand, and Antonio Torralba. 2012. A benchmark of computational models of saliency to predict human fixations. https://dspace.mit.edu/handle/1721.1/68590

[15] Dina Kanaan, Suzan Ayas, Birsen Donmez, Martina Risteska, and Joyita Chakraborty. 2019. Using Naturalistic Vehicle-Based Data to Predict Distraction and Environmental Demand. *International Journal of Mobile Human Computer Interaction* 11, 3 (July 2019), 59–70. doi:10.4018/ijmhci.2019070104

[16] Davis E. Kieras and Davis E. Meyer. 1997. An Overview of the EPIC Architecture for Cognition and Performance With Application to Human-Computer Interaction. *Human–Computer Interaction* 12, 4 (1997), 391–438. doi:10.1207/s15327051hci1204_4

[17] Sheila Klauer, Thomas Dingus, T Neale, J. Sudweeks, and D Ramsey. 2006. *The Impact of Driver Inattention on Near-Crash/Crash Risk: An Analysis Using the 100-Car Naturalistic Driving Study Data.* Technical Report. U.S. Department of Transportation, National Highway Traffic Safety Administration / Virginia Tech Transportation Institute, 3500 Transportation Research Plaza (0536) Blacksburg, Virginia 24061.

[18] Gwendolin Knappe, Andreas Keinath, Klaus Bengler, and Cristina Meinecke. 2007. Driving simulator as an evaluation tool–assessment of the influence of field of view and secondary tasks on lane keeping and steering performance. In *20th International Technical Conference on the Enhanced Safety of Vehicles (ESV) National Highway Traffic Safety Administration*. National Highway Traffic Safety Administration, Lyon, France, 1–11.

[19] Philip Koopman and Michael Wagner. 2017. Autonomous vehicle safety: An interdisciplinary challenge. *IEEE Intelligent Transportation Systems Magazine* 9, 1 (2017), 90–96.

[20] Tuomo Kujala and Dario D. Salvucci. 2015. Modeling visual sampling on in-car displays: The challenge of predicting safety-critical lapses of control. *International Journal of Human-Computer Studies* 79 (2015), 66–78. Integrating Knowledge of Multitasking and Interruptions Across Different Perspectives and Research Methods. doi:10.1016/j.ijhcs.2015.02.009

[21] Joonbum Lee. 2014. *Integrating the saliency map with distract-r to assess driver distraction of vehicle displays.* Ph. D. Dissertation. The University of Wisconsin-Madison.

[22] Liam Li, Kevin Jamieson, Afshin Rostamizadeh, Ekaterina Gonina, Moritz Hardt, Benjamin Recht, and Ameet Talwalkar. 2020. A System for Massively Parallel Hyperparameter Tuning. arXiv:1810.05934 [cs.LG] https://arxiv.org/abs/1810.05934

[23] Richard Liaw, Eric Liang, Robert Nishihara, Philipp Moritz, Joseph E. Gonzalez, and Ion Stoica. 2018. Tune: A Research Platform for Distributed Model Selection and Training. arXiv:1807.05118 [cs.LG] https://arxiv.org/abs/1807.05118

[24] Alexander Lingler, Helena Anna Frijns, Nelson J. S. Silva, Patrick Ebel, and Philipp Wintersberger. 2026. Enhancing Mobile Multitasking with Reinforcement Learning-Based Attention Management. *International Journal of Human–Computer Interaction* 0, 0 (2026), 1–28. doi:10.1080/10447318.2026.2688492

[25] Alexander Lingler, Dinara Talypova, Jussi PP Jokinen, Antti Oulasvirta, and Philipp Wintersberger. 2024. Supporting Task Switching with Reinforcement Learning. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–18.

[26] Martin Lorenz, Tiago Amorim, Debargha Dey, Mersedeh Sadeghi, and Patrick Ebel. 2024. Computational Models for In-Vehicle User Interface Design: A Systematic Literature Review. In *Proceedings of the 16th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (Stanford CA USA). ACM, New York, NY, USA, 204–215. doi:10.1145/3640792.3675735

[27] Martin Lorenz, Niko Konzack, Alexander Lingler, Philipp Wintersberger, and Patrick Ebel. 2026. CR-Eyes: A Computational Rational Model of Visual Sampling Behavior in Atari Games. arXiv:2603.26527 [cs]. doi:10.48550/arXiv.2603.26527

[28] Andrew Mackenzie and Julie Harris. 2015. Using Experts' Eye Movements to Influence Scanning Behaviour in Novice Drivers. *Journal of Vision* 15, 12 (September 2015), 367. doi:10.1167/15.12.367

[29] Stefan Mattes. 2003. The lane-change-task as a tool for driver distraction evaluation. *Quality of Work and Products in Enterprises of the Future* 57, 60 (2003), 7.

[30] Roderick Murray-Smith, Antti Oulasvirta, Andrew Howes, Jörg Müller, Aleksi Ikkala, Miroslav Bachinski, Arthur Fleig, Florian Fischer, and Markus Klar. 2022. What simulation can do for HCI research. *Interactions* 29, 6 (Nov. 2022), 48–53. doi:10.1145/3564038

[31] Antti Oulasvirta, Jussi P. P. Jokinen, and Andrew Howes. 2022. Computational Rationality as a Theory of Interaction. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 359, 14 pages. doi:10.1145/3491102.3517739

[32] Bryan Reimer. 2009. Impact of cognitive task complexity on drivers' visual tunneling. *Transportation Research Record* 2138, 1 (2009), 13–19.

[33] Dario D Salvucci. 2001. An integrated model of eye movements and visual encoding. *Cognitive Systems Research* 1, 4 (2001), 201–220.

[34] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal Policy Optimization Algorithms. arXiv:1707.06347 [cs.LG] https://arxiv.org/abs/1707.06347

[35] Rob Semmens, Nikolas Martelaro, Pushyami Kaveti, Simon Stent, and Wendy Ju. 2019. Is Now A Good Time? An Empirical Study of Vehicle-Driver Communication Timing. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3290605.3300867

[36] Danqing Shi, Yao Wang, Yunpeng Bai, Andreas Bulling, and Antti Oulasvirta. 2025. Chartist: Task-driven Eye Movement Control for Chart Reading. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–14. doi:10.1145/3706598.3713128

[37] Danqing Shi, Yujun Zhu, Jussi PP Jokinen, Aditya Acharya, Aini Putkonen, Shumin Zhai, and Antti Oulasvirta. 2024. CRTypist: Simulating touchscreen typing behavior via computational rationality. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–17.

[38] Alan A Stocker and Eero P Simoncelli. 2006. Noise characteristics and prior expectations in human visual speed perception. *Nature Neuroscience* 9, 4 (March 2006), 578–585. doi:10.1038/nn1669

[39] Yanbing Wang, Ziran Wang, Kyungtae Han, Prashant Tiwari, and Daniel B. Work. 2021. Personalized Adaptive Cruise Control via Gaussian Process Regression. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)* (Indianapolis, IN, USA). IEEE Press, Piscataway, NJ, USA, 1496–1502. doi:10.1109/ITSC48978.2021.9564498

[40] Philipp Wintersberger, Andreas Riener, Clemens Schartmüller, Anna-Katharina Frison, and Klemens Weigl. 2018. Let Me Finish before I Take Over: Towards Attention Aware Device Integration in Highly Automated Vehicles. In *Proceedings of the 10th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (Toronto, ON, Canada) *(AutomotiveUI '18)*. Association for Computing Machinery, New York, NY, USA, 53–65. doi:10.1145/3239060.3239085

[41] Philipp Wintersberger, Clemens Schartmüller, and Andreas Riener. 2019. Attentive user interfaces to improve multitasking and take-over performance in automated driving: the auto-net of things. *International Journal of Mobile Human Computer Interaction (IJMHCI)* 11, 3 (2019), 40–58.

[42] June-Seop Yoon, Hee-Seung Moon, Ben Boudaoud, Josef Spjut, Iuri Frosio, Byungjoo Lee, and Joohwan Kim. 2025. Modeling Visually-Guided Aim-and-Shoot Behavior in First-Person Shooters. *International Journal of Human-Computer Studies* 199 (2025), 103503. doi:10.1016/j.ijhcs.2025.103503

[43] Yifei Zhang, James Song, Siyi Gu, Tianxu Jiang, Bo Pan, Guangji Bai, and Liang Zhao. 2025. Saliency-bench: A comprehensive benchmark for evaluating visual explanations. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*. ACM, New York, NY, USA, 5924–5935.

[44] Feng Zhou, X Jessie Yang, and Joost CF De Winter. 2021. Using eye-tracking data to predict situation awareness in real time during takeover transitions in conditionally automated driving. *IEEE transactions on intelligent transportation systems* 23, 3 (2021), 2284–2295.