Neuro-Cognitive Reward Modeling for Human-Centered Autonomous Vehicle Control

Supplementary Material

1. Ablation for EEG feature prediction model

We conducted ablation with (i) a temporal model (CNN+GRU), (ii) reduced temporal context (single-frame input), and (iii) removing temporal average pooling. Our baseline achieves the best speed-accuracy trade-off (See Tab. 1). It attains the highest accuracy (82) while preserving real-time throughput (204 FPS). The GRU provides comparable accuracy (82) but reduces fps (137 FPS), while using a single frame degrades performance (80), indicating that temporal context is beneficial. Removing average pooling further reduces accuracy (79), possibly due to overfitting.

	F_1	F_2	F_3	F_4	F_5	Mean	FPS
Ours	80	85	77	86	81	82	204
GRU	80	85	79	85	80	82	137
single frame	78	82	75	81	83	80	250
no avg pool	80	81	72	84	77	79	176

Table 1. Five-fold result and fps of EEG feature prediction model.

2. Human Data Collection Platform for the RLHF baseline

We implemented a canonical RLHF (human preference-based) baseline following [1]. We collected 2-second video clips of RL agents exhibiting diverse driving behaviors, then built a custom interface to present side-by-side clip pairs sampled from this dataset, as demonstrated in supplementary Fig. 1. Participants completed 2000 pairwise queries, each selecting "prefer left," "prefer right," or "cannot tell." Data collection required 10 hours of manual labeling. The responses were stored in csv format and used to train a human preference reward model via the Bradley-Terry loss. This model was then integrated into RL training under identical conditions to our ERP-based predictor. Across all evaluation metrics, our ERP-based reward model outperformed the human preference baseline, as summarized in the table 3.

3. Discussion

3.1. Human cognitive reward model

In this study, we present a novel framework that integrates human cognitive reward feedback into the RL paradigm, demonstrating its effectiveness in emergency braking scenarios. Building in the current advance of human reward model, including RLHF framework to predict human preference from human behavior [1], and a framework that predict eye-tracking features from text for enhancing reward modeling [4], we propose a cognitive reward prediction model capable of estimating EEG-derived features, a more direct and natural representation of human cognition, from visual scene images. Importantly, we apply this model to a novel domain: autonomous driving. To the best of our knowledge, this is the first work that leverages EEG-based cognitive feedback for reward modeling in an autonomous driving context. Currently, this model achieves an accuracy of 82%, which may be attributed to the low signal-to-noise ratio inherent in contemporary EEG devices. With advancements in EEG hardware, it is possible that higher accuracies will be attainable in the future.

3.2. Neuroscience Findings and ERP Interpretation

We present the averaged event-related potentials (ERPs) across all channels in supplementary Fig. 2. The maximum ERP variation occurs at approximately 392 ms, aligning with the typical latency range of the P3 component (300–500 ms) observed in cognitive processing tasks [5]. Notably, the highest amplitudes are recorded over frontal-central electrodes, consistent with prior findings that identify these regions as exhibiting maximal P3 activity [5]. For analyses, we focus on the Cz electrode, situated over the vertex of the scalp, which is commonly associated with motor-related cortical activity and is frequently utilized in studies investigating motor preparation and execution [3].

ERP plots in supplementary Fig. 3 visualize the brain's time-locked response to the front vehicle's braking event across many repeated trials. Each horizontal line corresponds to one EEG trial, aligned to the moment when the leading car begins to brake. Warmer colors reflect higher positive voltages, while cooler colors reflect more negative voltages. By stacking these trials vertically and sorting them by the participant's behavioral reaction time, the ERP image reveals a systematic pattern: a positive deflection emerging around 300–500 ms post-event, consistent with the classical P3 component associated with evaluating unexpected or safety-critical stimuli. Notably, the Pearson correlation between ERP peak latency and reaction time is statistically significant ($p = 0.0438$), indicating that later neural responses are associated with slower braking behavior.

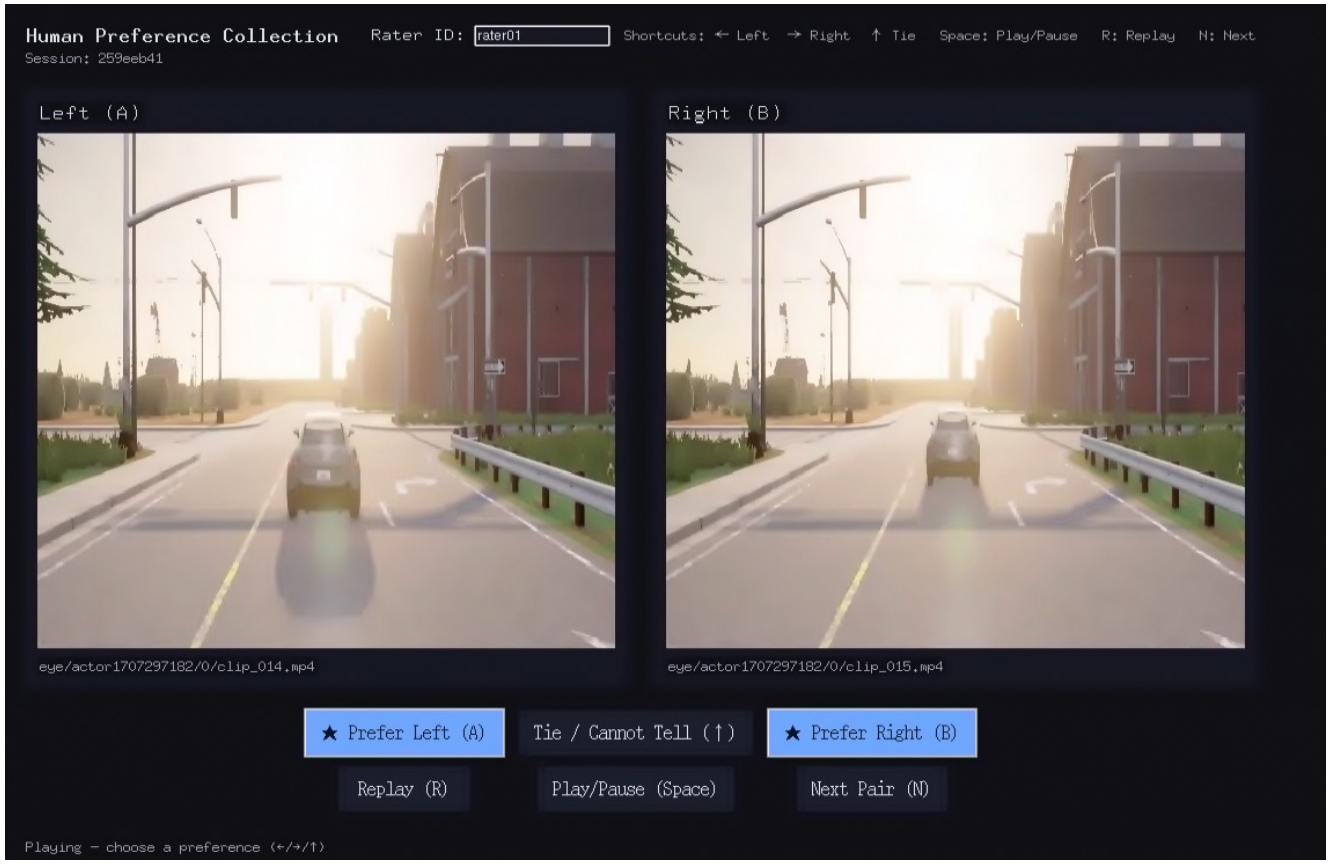

Figure 1. Customized data collection platform for reproducing the RLHF baseline. The users are able to choose Prefer Left/ Prefer Right/ Tie after watching a 2-second video clip.

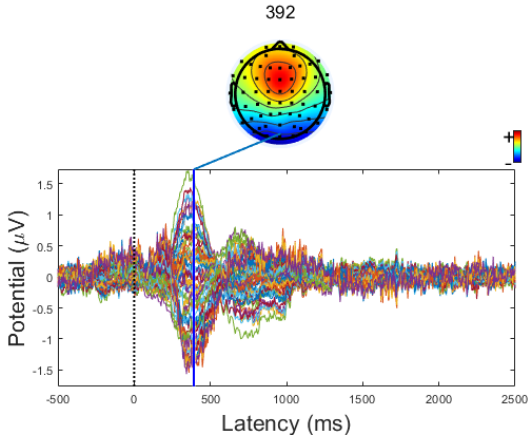

Figure 2. The relative time courses of the averaged ERP at all channels. The scalp map shows the topographic distribution of the average potential at 392 ms, which is the latency of maximum ERP data variance.

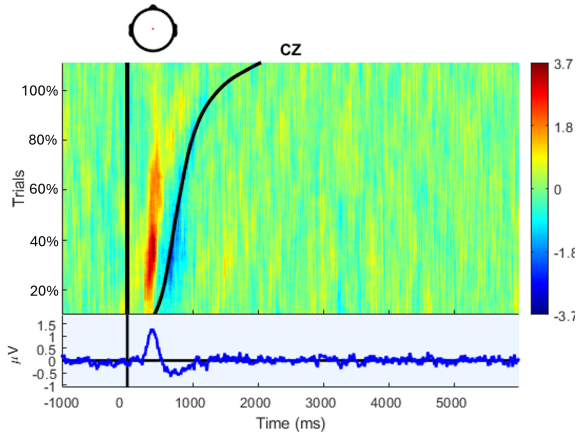

Figure 3. The ERP image sorted by the reaction time. Every horizontal line represents EEG amplitude in a single experimental trial.

3.3. Intersubject variation

EEG signals are inherently subject to intersubject variability, arising from differences in individual physiolog-

ical states and environmental conditions. This variability poses challenges for the generalization of EEG-based models across different individuals. To address this, we employed a five-fold cross-validation strategy, wherein the EEG dataset was randomly partitioned into five folds. The cross-validation results indicated consistent classification accuracy across the different folds, with only minor fluctuations observed. Our visualization result in Fig. 4 shows a low standard error, as revealed by the shaded area, and also shows the consistency among subjects. Further research with more participants and integrating techniques to identify outliers could potentially mitigate this problem.

4. Training details and reward setting

All reinforcement learning models were trained for 400 epochs. Experiments and training were conducted using an NVIDIA L4 GPU, accumulating more than 25 GPU days in total. This extended runtime primarily reflects our rigorous evaluation protocol, which includes training each model across five distinct random seeds and assessing performance in unseen environments. Notably, our proposed method requires under 5 GPU hours to train. A summary of the hyperparameters employed during training is provided in supplementary Table 2.

To define the time-gap reward in our reinforcement learning setup, we adopt the formulation introduced by [2], which encourages safe following behavior based on the temporal spacing between vehicles. The reward r_{gap} is determined by the time gap T_{gap} between the ego and lead vehicles as follows:

$$r_{gap} = \begin{cases} T_{gap}, & \text{if } T_{gap} \in [1, 2] \\ \max(-1/T_{gap}, -10), & \text{if } T_{gap} < 1 \\ \max(-T_{gap}, -10), & \text{if } T_{gap} > 2 \end{cases} \quad (1)$$

The time gap is calculated using $T_{\text{gap}} = \frac{Dis}{V_{\text{ego}}}$, where Dis is the distance to the lead vehicle and V_{ego} is the speed of the ego vehicle. This reward mechanism promotes maintaining a safe and efficient following distance, discouraging both tailgating and excessive spacing, thereby supporting smooth and collision-free driving behavior.

5. ERP-Based Validation of the EEG Feature Prediction Model

To verify the cognitive relevance of our EEG feature prediction model, we conducted an ERP-based analysis comparing trials predicted to induce high ERP responses versus those predicted to induce low ERP responses. We first grouped test trials based on the model's binary predictions with high ERP and low ERP, and then computed the average ERP waveform for each group across all subjects and trials. This analysis was performed using held-out EEG recordings across all five folds of the cross-validation that were not used during training.

As shown in supplementary Fig. 4, trials predicted as high ERP consistently exhibited a prominent positive deflection in the 300–500 ms range, characteristic of the P300 component. In contrast, trials predicted as low ERP showed markedly attenuated or absent ERP components in the same temporal window. This analysis confirms that our model's predictions align with neurophysiological markers of cognitive processing, supporting its use in RLHF as a proxy for online neural signals. By doing so, we can scale RLHF training without the need for real-time EEG recordings, preserving both performance and interpretability.

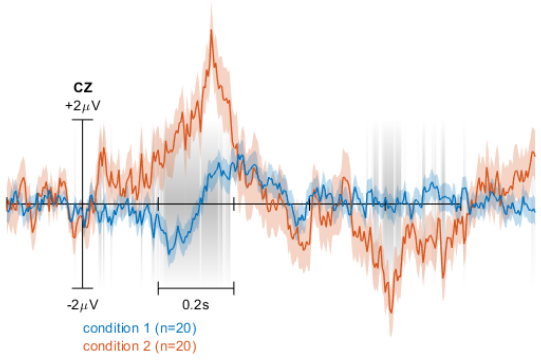

Figure 4. Average ERP waveform at the Cz channel for the EEG feature prediction model. Condition 1 corresponds to trials predicted with low ERP, and Condition 2 to trials predicted with high ERP. Shaded regions represent the standard error, and the gray area highlights time points with statistically significant differences between the two conditions.

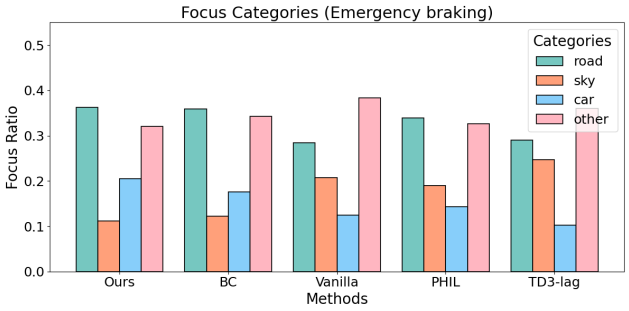

Figure 5. The ratio of focus categories of the policy network. Our model achieved the highest focus ratio on critical objects (e.g., the car).

Table 2. Hyperparameters of the RL in the current work.

Type	Description	Details
Buffer capacity	Capacity of the replay buffer	38400
Minibatch size	Capacity of minibatch	16
Actor learning rate	Initial learning rate of policy network	$5e^{-4}$
Critic learning rate	Initial learning rate of value network	$2e^{-4}$
Learning rate decay	Delay of learning rate per episode	0.995
Activation function	Activation function of the networks	Relu
Initial exploration	Initial exploration rate of noise in ϵ greedy	0.5
Final exploration	Final exploration rate of noise in ϵ greedy	0.05
Gamma (γ)	Discount factor of the Bellman equation	0.95
Soft updating factor	Parameter update frequency to target networks	$1e^{-3}$
Noise scale	Noise amplitude of action	0.2
Policy decay	Update frequency of critic over actor	1

6. Focus information

We systematically assess where the model focuses its attention by aligning the attention heatmaps with the corresponding segmentation images. For each category, we compute the proportion of pixels with attention values greater than 0.1 and present the results in supplementary Fig. 5. Our model allocates more attention to critical objects, particularly cars, than all other baselines, demonstrating that the cognitive reward model enhances the internal representation learned by the policy network in RL.

References

[1] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017. 1

[2] Resul Dagdanov, Halil Durmus, and Nazim Kemal Ure. Self-improving safety performance of reinforcement learning based driving with black-box verification algorithms. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5631–5637. IEEE, 2023. 3

[3] Yuxia Hu, Lipeng Zhang, Mingming Chen, Xiaoyuan Li, and Li Shi. How electroencephalogram reference influences the movement readiness potential? *Frontiers in Neuroscience*, 11:683, 2017. 1

[4] Angela Lopez-Cardona, Carlos Segura, Alexandros Karatzoglou, Sergi Abadal, and Ioannis Arapakis. Seeing eye to ai: Human alignment via gaze-based response rewards for large language models. *arXiv preprint arXiv:2410.01532*, 2024. 1

[5] John Polich. Updating p300: an integrative theory of p3a and p3b. *Clinical neurophysiology*, 118(10):2128–2148, 2007. 1