

A Cognitive Data Collection

In this study, we recruited 27 participants. Due to incomplete or corrupted data from seven of them, only data from the remaining 20 participants (10 novice and 10 expert drivers) were analyzed. All experiments employed the Neuroscan 64-channel Quik-Cap¹ system and Headbox for EEG acquisition, with cables connecting the amplifier to the acquisition unit, the stimulator, and the headbox; stimulation and synchronization were performed using a YOGA stimuli host and E-Prime software. Data were recorded in real time via CURRY 9, and usage of experimental and analysis software was controlled by hardware dongles.

The experiment consisted of three phases: setup, recording, and termination. During setup, participants donned the electrode cap and applied conductive paste, the vehicle idled, and the in-car industrial PC and Docker environment were started; Dreamview was then launched to synchronize system time. During recording, the external operator configured the COM port in the tester software, CURRY 9 injected timestamps, and E-Prime triggered cross-modal synchronization via spacebar presses and audio cues (marker 161 for beeps, 192 for other events). At the end of the session, the vehicle parked and the parking brake was engaged, the operator stopped recording with Ctrl+C, verified completeness of the .record files, and saved the raw .cnt files.

Subsequent data processing involved electrode localization, removal of non-EEG electrodes (EKG, EMG, Trigger, etc.), re-referencing to M1/M2, band-pass filtering at 0.1–50 Hz, bad-electrodes rejection, and artifact correction via ICA; events were then imported and data were epoched according to driving conditions. Using the processed data, we conducted statistical comparisons of fatigue levels, workload (theta/alpha plasticity), and task engagement between expert and novice drivers across driving scenarios, finding that experts exhibited more stable fatigue management and greater EEG adaptability in complex road sections.

Table 1: EEG System Components and Accessories

Type	Hardware Materials	Type	Hardware Materials
Headbox	64-channel Quik-Cap, Headbox unit	Cables	Cable 1 (Amplifier → Acquisition), Cable 2 (Amplifier → Stimuli, white), Cable 3 (Amplifier → Headbox)
Master Unit 1	Amplifier (wide-band main unit), Amplifier power cable	Software Licenses	Experimental software dongle, Analysis software dongle, E-Prime dongle
Master Unit 2	Stimulus generator unit	Subject Consumables	EEG paste (bulk), paste syringe, abrasive gel, cotton swabs, adhesive tape, shampoo, hair dryer, disposable absorbent wipes
Master Unit 3	Acquisition unit, Power cable	Other Equipment	Power strip, Expansion dock, Speaker set, Portable power bank, Serial-port interface box, Eye-tracking EEG sync cable

B Hardware and Software

B.1 Equipment

Hardware

- DJI Action 3 action cameras ×6
- DJI Action battery charging case ×2
- Insta360 X3 panoramic camera
- Insta360 charging case
- SD memory cards ×6

¹<https://compumedicsneuroscan.com/products/caps/quik-cap/>



Figure 1: Example Snapshots of the Video Modalities: top left – Baseline (forward road view); top right – Driver1 (driver’s face view); bottom left – Driver2 (driver’s posture view); bottom right – Driver3 (driver’s feet view).

- Mounting brackets (multiple)
- High-power Bluetooth speaker (for audio synchronization)

Software

- E-Prime (for EEG synchronization)
- Jianying (video editing)

B.2 Modalities

RGB Action Cameras Six in-vehicle viewpoints captured by DJI Action 3 cameras:

- **Baseline:** forward road view (dashcam)
- **Driver1:** driver’s face
- **Driver2:** driver’s feet
- **Driver3:** driver’s posture
- **Passenger1:** front passenger posture
- **Passenger2:** rear passenger posture

Recording starts when synchronization is initiated and ends when the vehicle is safely parked.

360° Panoramic HDR 360° exterior view recorded by Insta360 X3, capturing surrounding road environment and vehicle pose. Recording interval matches the RGB cameras. All in-vehicle RGB action cameras (DJI Action 3) record at 1080 P resolution and 30 fps with a wide-angle lens (155°) and subsequent distortion correction; the panoramic camera (Insta360 X3) captures 360° surround video in 5.7 K HDR at 30 fps to ensure precise timestamp synchronization with the RGB cameras.

B.3 Collection Procedure

1. Preparation and Inspection

- 1) Verify battery levels and free storage on all cameras; prepare spares.
- 2) Clean windshield and remove obstacles to minimize glare.

- 3) Mount and secure all six Action 3 cameras at predetermined positions; tighten brackets.
- 4) Check each camera’s field of view against the setup diagram.
- 5) For the test drive, power on only the Baseline camera.
- 6) Configure the Insta360 X3 to HDR mode and 360° dual-lens capture; secure on roof mount and level the camera.

2. Power-On

- 1) During the practice drive, the external operator starts the Baseline camera.
- 2) Before the formal trial, the in-vehicle operator gives the start command; the external operator powers on all cameras.

3. Recording Start

- 1) In-vehicle operator issues a voice command to initiate recording on all RGB cameras simultaneously.
- 2) External operator confirms LED indicators and starts the Insta360 X3.

4. Recording Stop

- 1) Upon safe stop of the vehicle, the in-vehicle operator issues a voice command to stop all RGB cameras.
- 2) External operator stops the Insta360 X3 recording on command.

C Comparison and Analysis

First and foremost, it should be clarified that, in line with many recent studies[5, 4], we view L2 error solely as an indicator of model convergence, whereas the ultimate goal of autonomous driving is safe, collision-free operation. In this study, we doubled the coefficient of the L2 loss term in the source code of VAD-Base[2] and found that, although the L2 error decreased by 0.15 m (20.8%, respectively), the collision rate increased by 0.07% (31.8%, respectively), as shown in Table. 2. Additionally, Ego-MLP[5] performs planning using only the vehicle’s historical state information without any visual input, the L2 error achieved is comparable to that of VAD, yet both the collision rate and closed-loop evaluation results are terrible. These findings motivate us to adopt collision rate as the primary performance metric for open-loop evaluation during model design and hyperparameter selection, while relegating L2 error to a secondary role as an indicator of model convergence, and coefficient settings are kept consistent with the baseline model to ensure a fair comparison. Additionally, due to the low computational efficiency of UniAD[1], we primarily select VAD as the baseline model in the large-scale experiments discussed in the following sections.

As shown in Table. 3, we compare the number of blocks, the dropout rate, and the number of cross-attention heads in the "Interact with the Ego Query" framework. Our observations indicate that, even with only one layer, the model achieves a stable improvement over the baseline. When the number of cross-attention heads is reduced by half, the model converges more easily and exhibits lower L2 error, while the collision rate remains unaffected. Additionally, dropout is a key parameter; excessively low dropout rates can lead to overfitting, resulting in degraded driving performance on the test set.

The comparison results of hyperparameters for the "Interact with Planning Features" framework are presented in Table.4. We observe that employing just one layer of our designed Decoder block can already lead to notable performance improvements. Nevertheless, appropriately increasing the number of layers can further enhance the overall performance of the model. It is also crucial to select a suitable dropout value, as excessively large or small dropout rates may lead to underfitting or overfitting, respectively, thereby degrading performance. Additionally, moderately reducing the number of attention heads can facilitate model convergence and improve driving performance. However, setting the number of attention heads too low may compromise the model’s learning capacity.

For the "Attach to the Spatio-temporal Features" framework, we compare the impact of the number of vision feature queries n_s , which are selected by human visual attention. We observe that as n_s increases, the model’s L2 error gradually decreases. However, the collision rate does not decrease accordingly and even increases when n_s is excessively high, as shown in Table. 5. This phenomenon can be attributed to the fact that this framework only captures visual features attended by the human brain, without learning planning-related driving cognition. These visual features primarily help the

model to fit the expert driving trajectories, but their contribution to enhancing the model’s driving performance is limited. Furthermore, when the visual features become overly redundant, they may even impair the model’s performance.

Table. 6 presents a comparison of detailed metrics between our model and baseline models under closed-loop evaluation. It can be observed that our model substantially reduces the rate of vehicle collisions and red light violations by 11.82% (38.8%, respectively) and 1.37% (37.6%, respectively). The rate of stop infraction is also significantly reduced by 0.46% (16.8%, respectively). In addition, occurrences of collisions with layouts and lane departures have also been mitigated. These results demonstrate that the E^3AD paradigm achieves significant improvements in driving performance, particularly in complex driving tasks such as interacting with other vehicles and understanding traffic signals.

Table 2: Relationship Between L2 Error and Collision Rate.

Method	L2(m)↓				Collision(%)↓			
	1s	2s	3s	Avg.	1s	2s	3s	Avg.
VAD-Base[2]	0.41	0.70	1.05	0.72	0.07	0.17	0.41	0.22
VAD-Base (Doubling L2 loss)	0.31	0.54	0.85	0.57	0.16	0.21	0.51	0.29
Ego-MLP[5]	0.46	0.76	1.12	0.78	0.21	0.35	0.58	0.38

Table 3: Comparison of Hyperparameters in the "Interact with the Ego Query" Framework.

Framework	Options			L2(m)↓				Collision(%)↓			
	Layers	Dropout	heads	1s	2s	3s	Avg.	1s	2s	3s	Avg.
VAD[2]	—	—	—	0.41	0.70	1.05	0.72	0.07	0.17	0.41	0.22
VAD (Reproduced)	—	—	—	0.40	0.70	1.04	0.71	0.10	0.16	0.44	0.23
Ours	1	0.1	8	0.37	0.66	1.04	0.69	0.06	0.14	0.41	0.20
	2	0.1	8	0.39	0.68	1.03	0.70	0.09	0.19	0.42	0.23
	4	0.1	8	0.36	0.62	0.95	0.64	0.08	0.19	0.41	0.23
	1	0.05	8	0.39	0.66	1.00	0.68	0.09	0.18	0.37	0.21
	1	0.1	4	0.33	0.58	0.92	0.61	0.07	0.14	0.39	0.20

Table 4: Comparison of Hyperparameters in the "Interact with the Planning Features" Framework.

Framework	Options			L2(m)↓				Collision(%)↓			
	Layers	Dropout	heads	1s	2s	3s	Avg.	1s	2s	3s	Avg.
VAD	—	—	—	0.41	0.70	1.05	0.72	0.07	0.17	0.41	0.22
VAD (Reproduced)	—	—	—	0.40	0.70	1.04	0.71	0.10	0.16	0.44	0.23
Ours	4	0.1	8	0.33	0.59	0.93	0.62	0.06	0.14	0.41	0.20
	1	0.1	8	0.39	0.70	1.07	0.72	0.03	0.14	0.44	0.20
	2	0.1	8	0.34	0.59	0.91	0.61	0.12	0.20	0.32	0.21
	6	0.1	8	0.38	0.64	0.96	0.66	0.09	0.18	0.39	0.22
	4	0.05	8	0.38	0.67	1.04	0.70	0.07	0.15	0.41	0.21
	4	0.15	8	0.34	0.61	0.96	0.63	0.06	0.17	0.40	0.21
	4	0.1	4	0.35	0.62	0.96	0.64	0.06	0.13	0.36	0.18
	4	0.1	2	0.33	0.59	0.93	0.62	0.06	0.14	0.41	0.20

D Visualization

Similar to other end-to-end methods[4, 1, 2, 3], we also provide visualization results, as shown in Fig. 2. It demonstrates a successful case from the closed-loop simulation experiments: when a vehicle parked ahead on the right suddenly starts moving, the baseline model adopts a more aggressive and risky avoidance maneuver, attempting to pass quickly and causing a severe collision. In contrast, our model learns to interact with the suddenly moving vehicle by slowing down and waiting until it has

Table 5: Comparison of Hyperparameters in the "Attach to the Spatio-temporal Features " Framework.

Framework	Options	L2(m)↓				Collision(%)↓			
	n_s	1s	2s	3s	Avg.	1s	2s	3s	Avg.
VAD	—	0.41	0.70	1.05	0.72	0.07	0.17	0.41	0.22
VAD (Reproduced)	—	0.40	0.70	1.04	0.71	0.10	0.16	0.44	0.23
Ours	4	0.38	0.67	1.04	0.70	0.09	0.17	0.400	0.22
	8	0.38	0.67	1.02	0.69	0.09	0.17	0.38	0.21
	16	0.34	0.60	0.95	0.63	0.12	0.19	0.49	0.27

Table 6: Comparison of E^3AD and Baseline Models on Closed-Loop Metrics.

Method	Closed-loop Metrics ↓					
	Layouts Collision(%)	Pedestrians Collision(%)	Vehicles Collision(%)	Running Red light(%)	Stop Infraction(%)	Off-road(%)
VAD-Base	15.46	0.91	30.46	3.64	2.73	23.64
Ours	15.00(↓ 3.1%)	0.91	18.64(↓ 38.8%)	2.27(↓ 37.6%)	2.27(↓ 16.8%)	22.73(↓ 3.8%)

126 completely departed before accelerating to proceed along the route, thus avoiding a collision. This
 127 case to some extent suggest that E^3AD exhibits behaviors that are closer to human driving style and
 128 are characterized by enhanced safety.

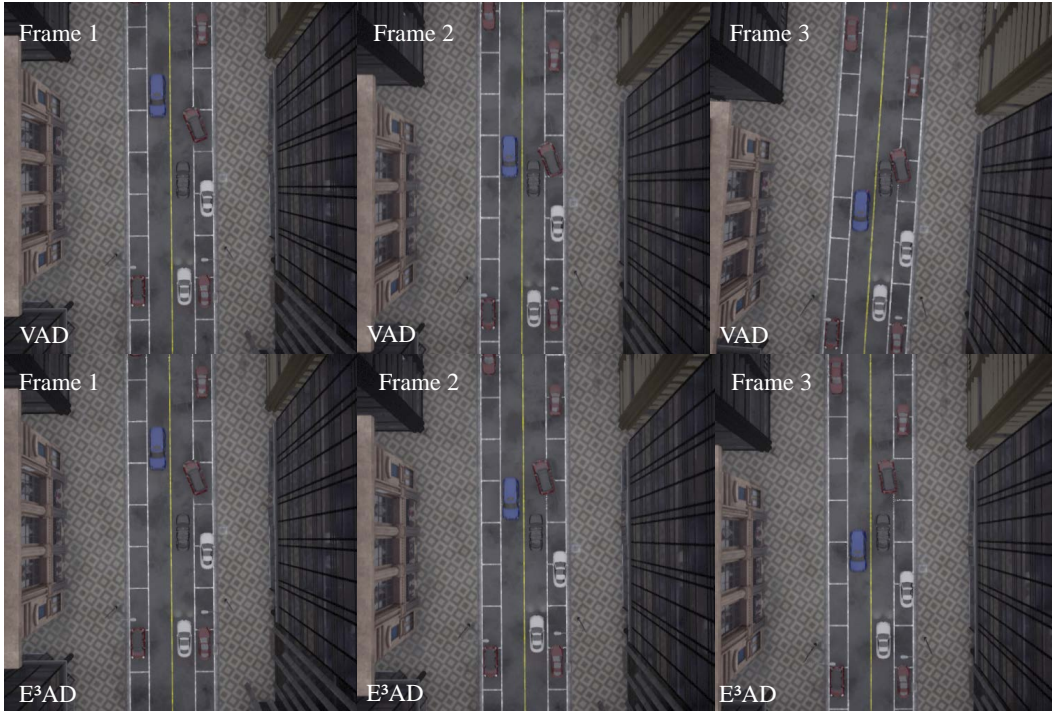


Figure 2: Visualization Comparison of E^3AD (VAD-Base) and the Baseline on Closed-loop Evaluation.

129 References

- 130 [1] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du,
 131 Tianwei Lin, Wenhai Wang, et al. Planning-oriented autonomous driving. In *Proceedings of the*
 132 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17853–17862, 2023.
- 133 [2] Bo Jiang, Shaoyu Chen, Qing Xu, Bencheng Liao, Jiajie Chen, Helong Zhou, Qian Zhang, Wenyu
 134 Liu, Chang Huang, and Xinggang Wang. Vad: Vectorized scene representation for efficient

- 135 autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer*
136 *Vision*, pages 8340–8350, 2023.
- 137 [3] Yingyan Li, Lue Fan, Jiawei He, Yuqi Wang, Yuntao Chen, Zhaoxiang Zhang, and Tieniu
138 Tan. Enhancing end-to-end autonomous driving with latent world model. *arXiv preprint*
139 *arXiv:2406.08481*, 2024.
- 140 [4] Zhiqi Li, Zhiding Yu, Shiyi Lan, Jiahao Li, Jan Kautz, Tong Lu, and Jose M Alvarez. Is ego status
141 all you need for open-loop end-to-end autonomous driving? In *Proceedings of the IEEE/CVF*
142 *Conference on Computer Vision and Pattern Recognition*, pages 14864–14873, 2024.
- 143 [5] Jiang-Tian Zhai, Ze Feng, Jinhao Du, Yongqiang Mao, Jiang-Jiang Liu, Zichang Tan, Yifu
144 Zhang, Xiaoqing Ye, and Jingdong Wang. Rethinking the open-loop evaluation of end-to-end
145 autonomous driving in nuscen. *arXiv preprint arXiv:2305.10430*, 2023.