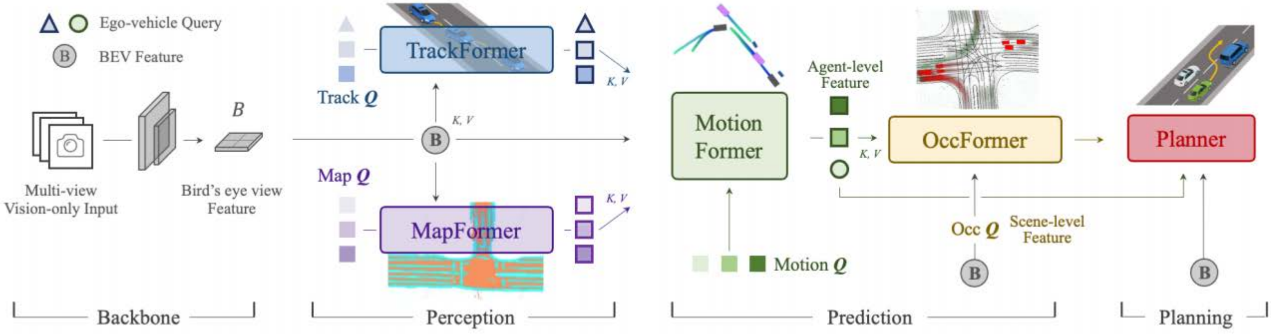
清華大學
Tsinghua University

Embodied Cognition Augmented End2End Autonomous Driving

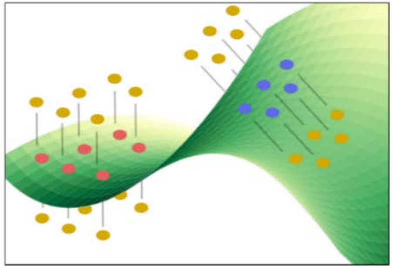
Ling Niu, Xiaoji Zheng, Han Wang, Chen Zheng,
Ziyuan Yang, Bokui Chen*, Jiangtao Gong*

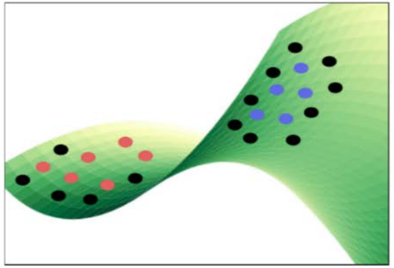
01 Motivation

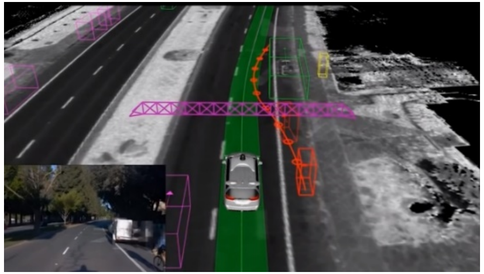
Background



- 1) Enhanced safety
- 2) Improved framework completeness



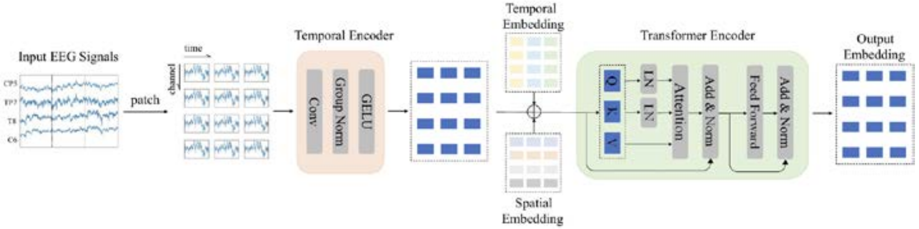


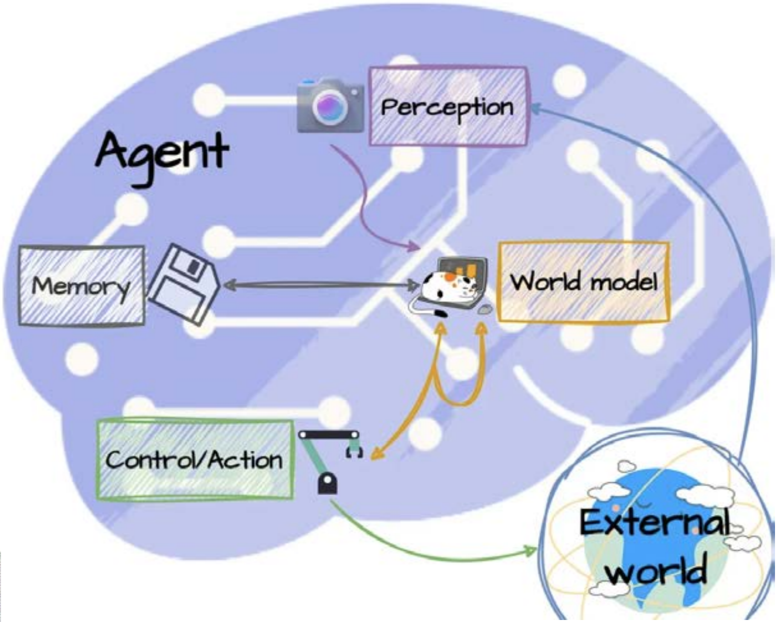


limited capacity on OOD data

Motivation

Human driver can drive safely even in unseen scenarios.





Contributions

1

The E^3AD paradigm is the first to incorporate human driving cognition to enhance end-to-endautonomous driving models.

2

Achieving substantial improvements in driving performance with small computational cost, while reaching the level ofstate-of-the-art method.

3

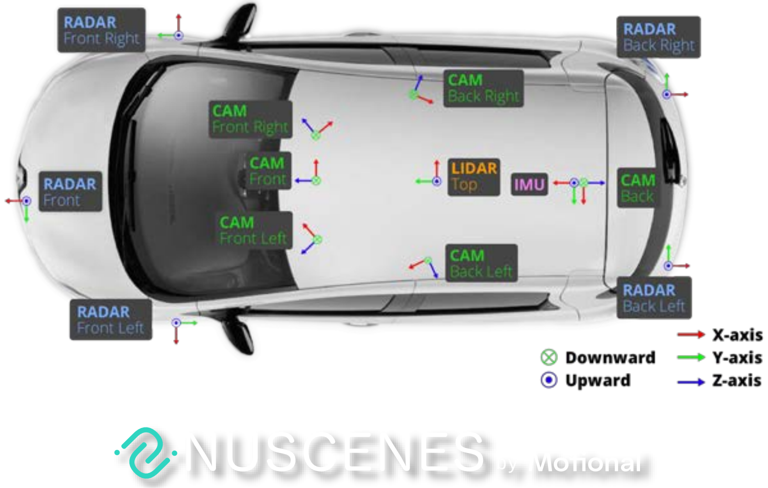
Make novel contributions to the field of embodiedhuman intelligence augmentation in AI algorithm

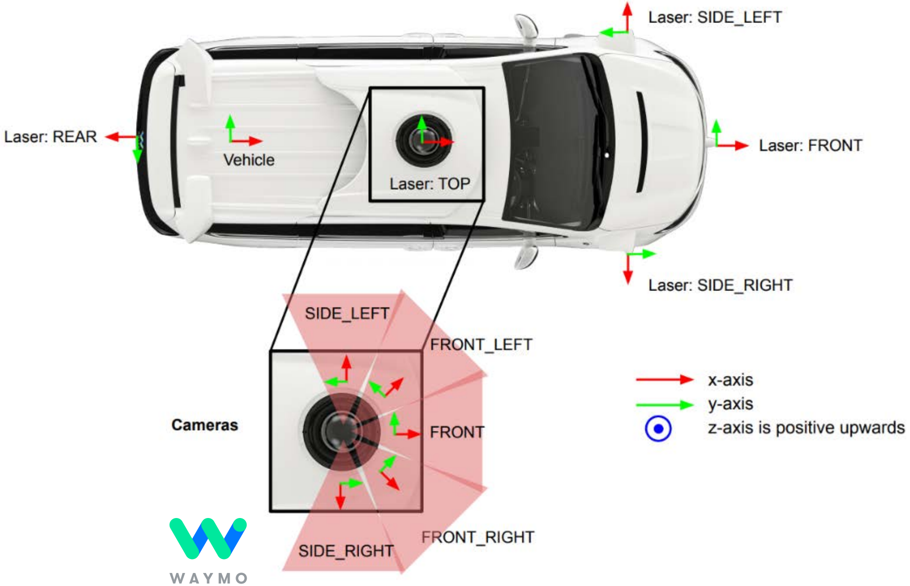
清華大学

Tsinghua University

02 Cognitive Dataset Collection

Cognitive Data Collection





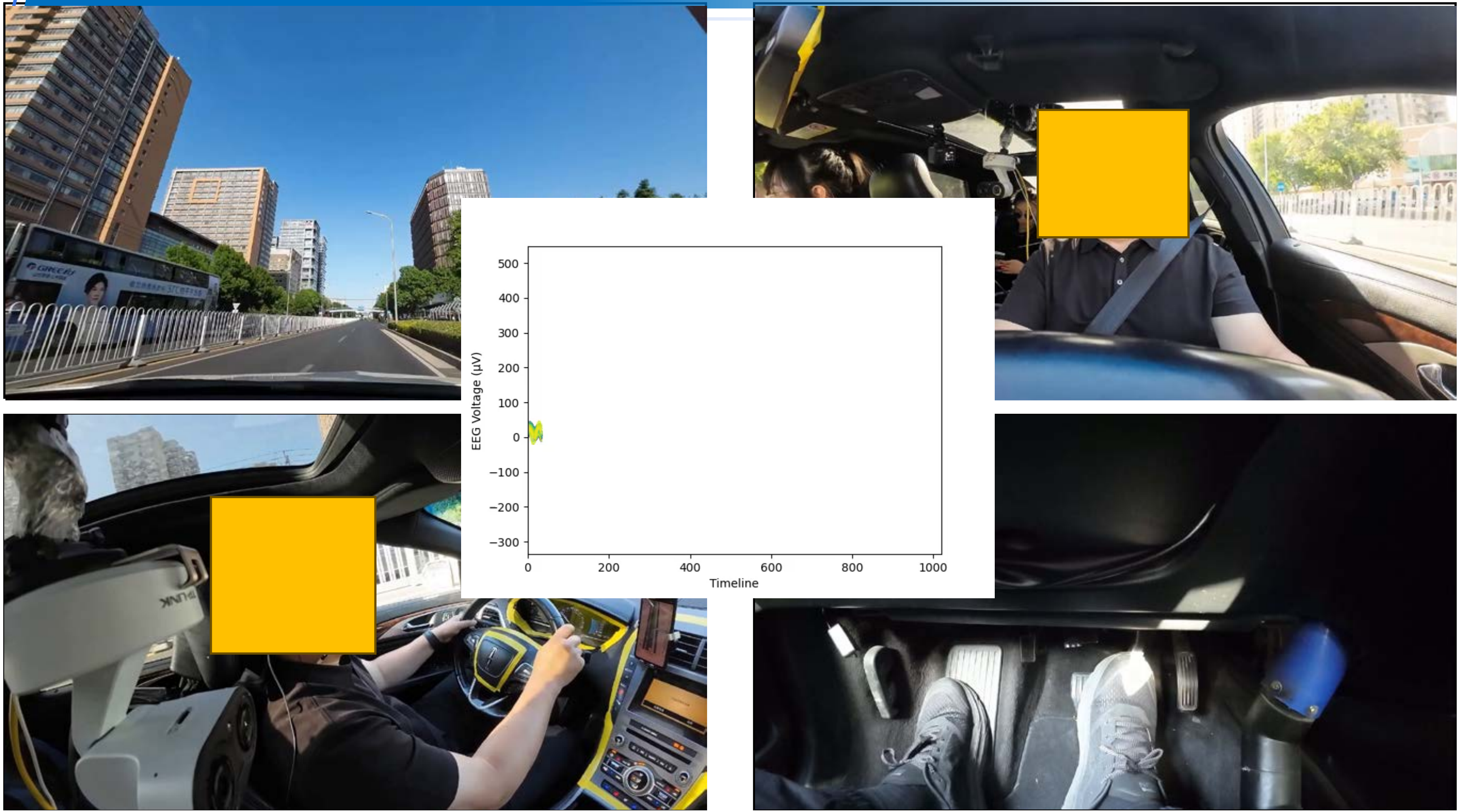
https://www.nuscenes.org/

https://waymo.com/open/

Cognitive Data Collection

Participants: 20 drivers (10 expert, 10 novice)
Driving Conditions: 14 varied conditions
Total Driving Time: 400 minutes (20 minutes per driver)
Model Training Data Duration: ~100 minutes
~50 minutes from expert drivers
~50 minutes from novice drivers

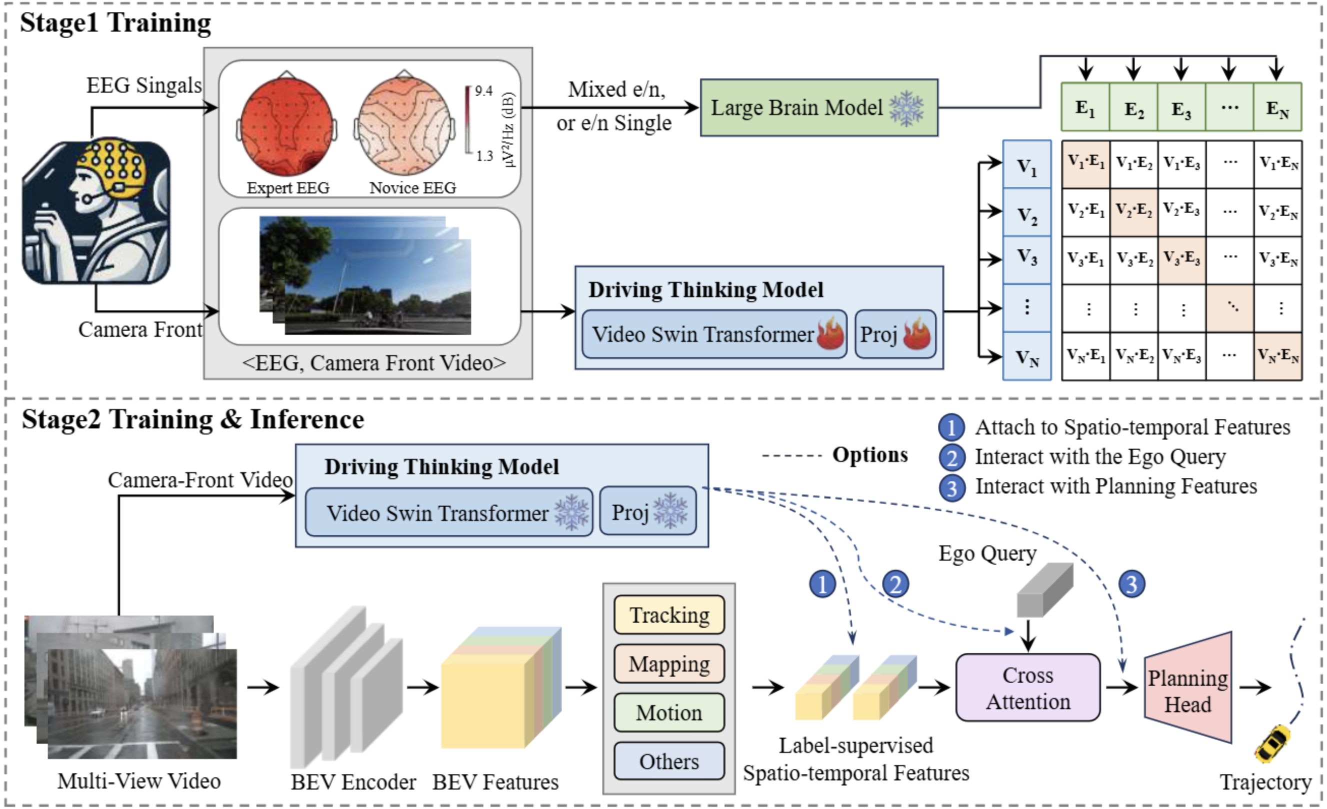
Cognitive Data Collection



Tsinghua University

03 METHOD

METHOD


Stage1 Training
EEG Singals
Expert EEG
Novice EEG
9.4
1.3
μV²/Hz (dB)
Camera Front
<EEG, Camera Front Video>
Mixed e/n,
or e/n Single
Large Brain Model
Driving Thinking Model
Video Swin Transformer
Proj
E1 E2 E3 ··· EN
V1 V2 V3 ⋮ VN
V1·E1 V1·E2 V1·E3 ··· V1·EN
V2·E1 V2·E2 V2·E3 ··· V2·EN
V3·E1 V3·E2 V3·E3 ··· V3·EN
VN·E1 VN·E2 VN·E3 ··· VN·EN
Stage2 Training & Inference
Camera-Front Video
Driving Thinking Model
Video Swin Transformer
Proj
Options
1 Attach to Spatio-temporal Features
2 Interact with the Ego Query
3 Interact with Planning Features
Ego Query
Multi-View Video
BEV Encoder
BEV Features
Tracking
Mapping
Motion
Others
Label-supervised
Spatio-temporal Features
Cross
Attention
Planning
Head
Trajectory

04 Results and Discussion

Open-Loop Evaluation

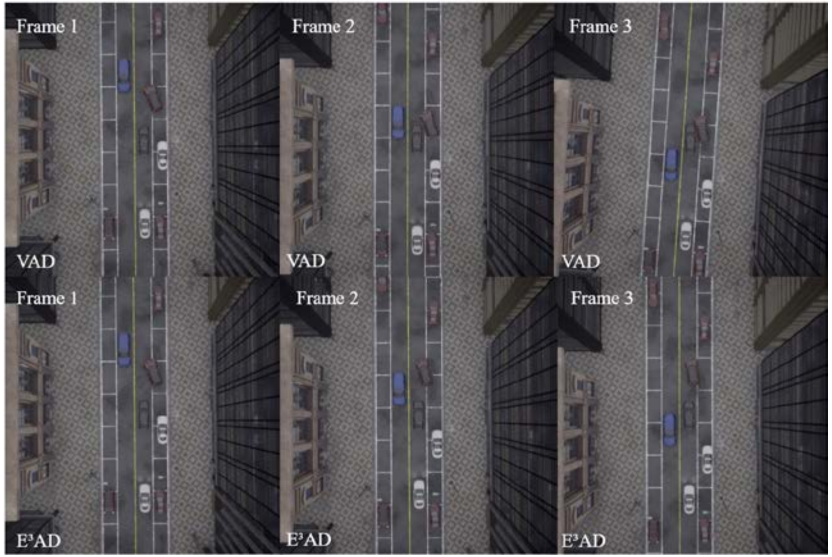
Method	ST-P3 Metrics								UniAD Metrics								FPS
	L2(m) ↓				Collision(%) ↓				L2(m) ↓				Collision(%) ↓				
	1s	2s	3s	Avg.	1s	2s	3s	Avg.	1s	2s	3s	Avg.	1s	2s	3s	Avg.	
ST-P3[9]	1.33	2.11	2.90	2.11	0.23	0.62	1.27	0.71	-	-	-	-	-	-	-	-	1.6
VAD-Base[12]	0.41	0.70	1.05	0.72	0.07	0.17	0.41	0.22	-	-	-	-	-	-	-	-	4.5
VAD-Tiny[12]	0.46	0.76	1.12	0.78	0.21	0.35	0.58	0.38	-	-	-	-	-	-	-	-	16.8
UniAD[10]	-	-	-	-	-	-	-	-	0.48	0.96	1.65	1.03	0.05	0.17	0.71	0.31	1.8
GenAD[31]	0.28	0.49	0.78	0.52	0.08	0.14	0.34	0.19	0.36	0.83	1.55	0.91	0.06	0.23	1.00	0.43	6.7
BEV-Planner[18]	0.28	0.42	0.68	0.46	0.04	0.37	1.07	0.49	-	-	-	-	-	-	-	-	-
LAW[16]	0.26	0.57	1.01	0.61	0.14	0.21	0.54	0.30	-	-	-	-	-	-	-	-	19.5
VAD-Tiny[12]	0.46	0.76	1.12	0.78	0.21	0.35	0.58	0.38	-	-	-	-	-	-	-	-	9.8†
E³AD(VAD-Tiny)	**0.40**	**0.66**	**0.98**	**0.68**	**0.18**	**0.33**	**0.55**	**0.35**	-	-	-	-	-	-	-	-	**8.8†**
VAD-Base[12]	0.41	0.70	1.05	0.72	0.07	0.17	0.41	0.22	-	-	-	-	-	-	-	-	3.8†
E³AD(VAD-Base)	**0.35**	**0.62**	**0.96**	**0.64**	**0.06**	**0.13**	**0.36**	**0.18***	-	-	-	-	-	-	-	-	**3.7†**
UniAD[10]	-	-	-	-	-	-	-	-	0.48	0.96	1.65	1.03	0.05	0.17	0.71	0.31	1.4†
E³AD(UniAD)	-	-	-	-	-	-	-	-	**0.48**	**0.96**	**1.64**	**1.03**	**0.07**	**0.10**	**0.52**	**0.23***	**1.4†**
GenAD(official checkpoint)[31]	0.25	0.46	0.76	0.49	0.11	0.21	0.45	0.26	0.33	0.81	1.58	0.91	0.06	0.43	1.19	0.56	6.7†
GenAD(Reproduce)	0.25	0.46	0.76	0.49	0.14	0.26	0.50	0.30	0.33	0.79	1.58	0.90	0.17	0.49	1.15	0.60	6.7†
E³AD(GenAD)	**0.24**	**0.44**	**0.74**	**0.47**	**0.10**	**0.21**	**0.42**	**0.24**	**0.32**	**0.78**	**1.52**	**0.87**	**0.15**	**0.35**	**1.09**	**0.53**	**6.6†**
LAW[16]	0.26	0.57	1.01	0.61	0.14	0.21	0.54	0.30	-	-	-	-	-	-	-	-	16.5†
E³AD(LAW)	**0.28**	**0.57**	**0.98**	**0.61**	**0.11**	**0.13**	**0.42**	**0.22**	-	-	-	-	-	-	-	-	**16.1†**

Closed-Loop Evaluation

Table A.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%)

Method	Open-loop Metric	Closed-loop Metric	
	Avg.L2(m) ↓	DS ↑	SR(%) ↑
AD-MLP	3.64	18.05	0.00
UniAD-Tiny	0.80	40.73	13.18
UniAD-Base	0.73	45.81	16.36
VAD	0.91	42.35	15.00
E^3AD(UniAD-Base)	**0.69**	**50.07**	**20.12**
E^3AD(VAD)	**0.86**	**47.63**	**19.54**



Ablation Study and Comparison

Table 3: Ablation study on the impact of the Driving-Thinking model on driving models is conducted, using the $E^3AD(VAD-Base)$ as the experimental baseline.

Contrastive Learning		Freeze	L2(m) ↓				Collision(%) ↓			
Expert	Novice	Video Encoder	1s	2s	3s	Avg.	1s	2s	3s	Avg.
-	-	✓	0.39	0.68	1.02	0.70	0.13	0.21	0.40	0.25
✓	✓	-	0.40	0.64	1.04	0.69	0.15	0.18	0.42	0.25
✓	-	✓	0.33	0.59	0.92	**0.61**	0.05	0.18	0.40	**0.21**
-	✓	✓	0.38	0.65	0.99	0.67	0.07	0.18	0.38	0.21
✓	✓	✓	0.35	0.62	0.96	**0.64**	0.06	0.13	0.36	**0.18**

Table 4: Comparison of different Frameworks, using the $E^3AD(VAD-Base)$ as the baseline.

Framework	L2(m) ↓				Collision(%)↓			
	1s	2s	3s	Avg.	1s	2s	3s	Avg.
Attach to Spatio-temporal features	0.38	0.67	1.02	0.69	0.09	0.17	0.39	0.22
Interact with the Ego Query	0.37	0.66	1.04	0.69	0.06	0.14	0.41	0.20
Interact with the Planning Features	0.35	0.62	0.96	0.64	0.06	0.13	0.36	0.18