DrivePI: Spatial-aware 4D MLLM for Unified Autonomous Driving Understanding, Perception, Prediction and Planning

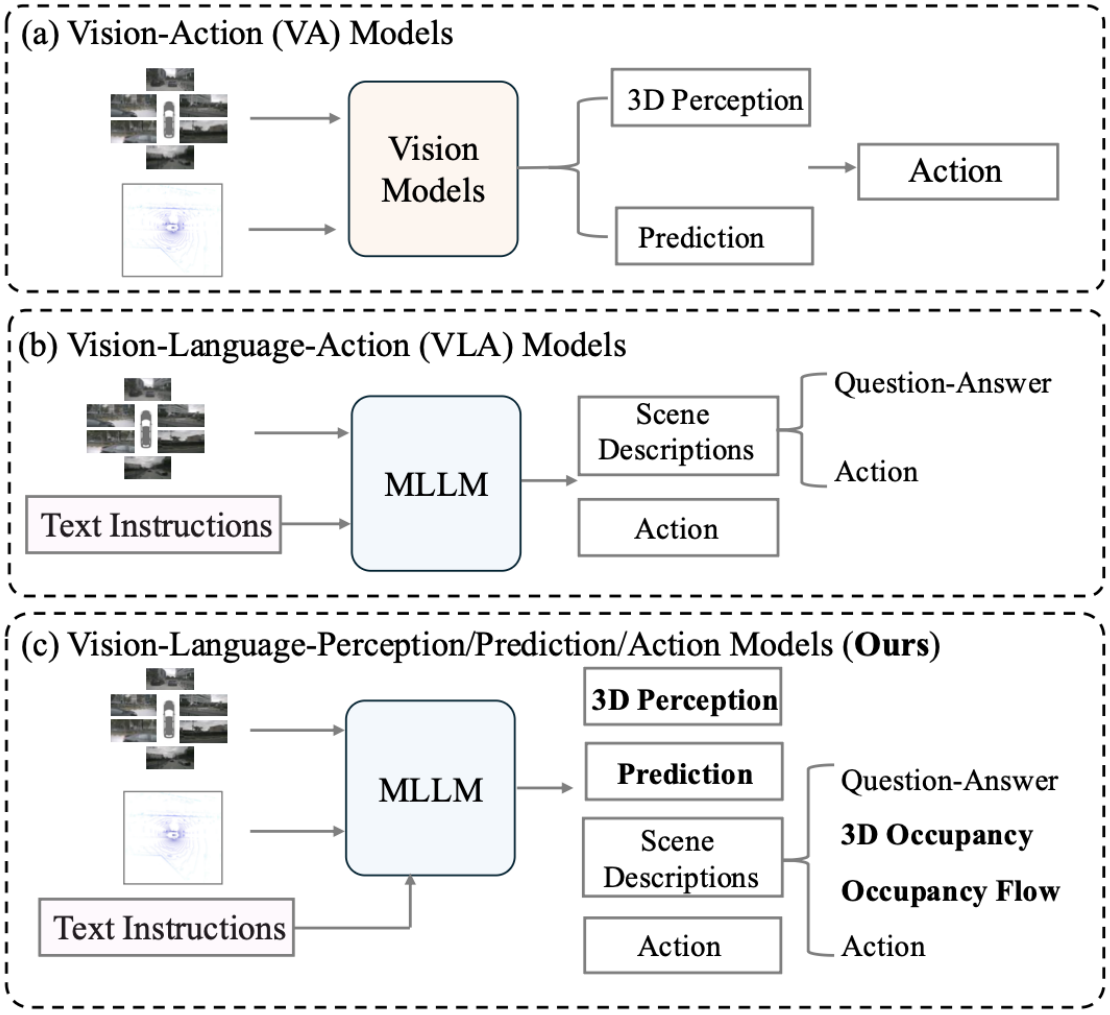
Zhe Liu[1], Runhui Huang[1], Rui Yang[1], Siming Yan[2],
Zining Wang[2], Lu Hou[2], Di Lin[3], Xiang Bai[4], Hengshuang Zhao[1,✉]

[1]The University of Hong Kong, [2]Yinwang Intelligent Technology Co. Ltd.,
[3]Tianjin University, [4]Huazhong University of Science and Technology

Zhe Liu
2026/5/18

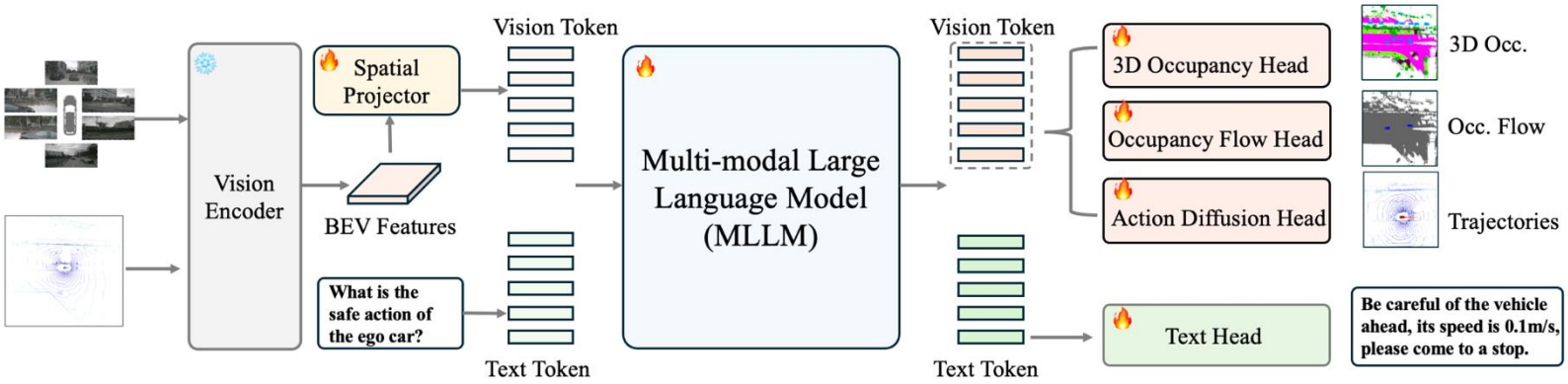
Motivation

Challenges: Interpretability and safety guarantees



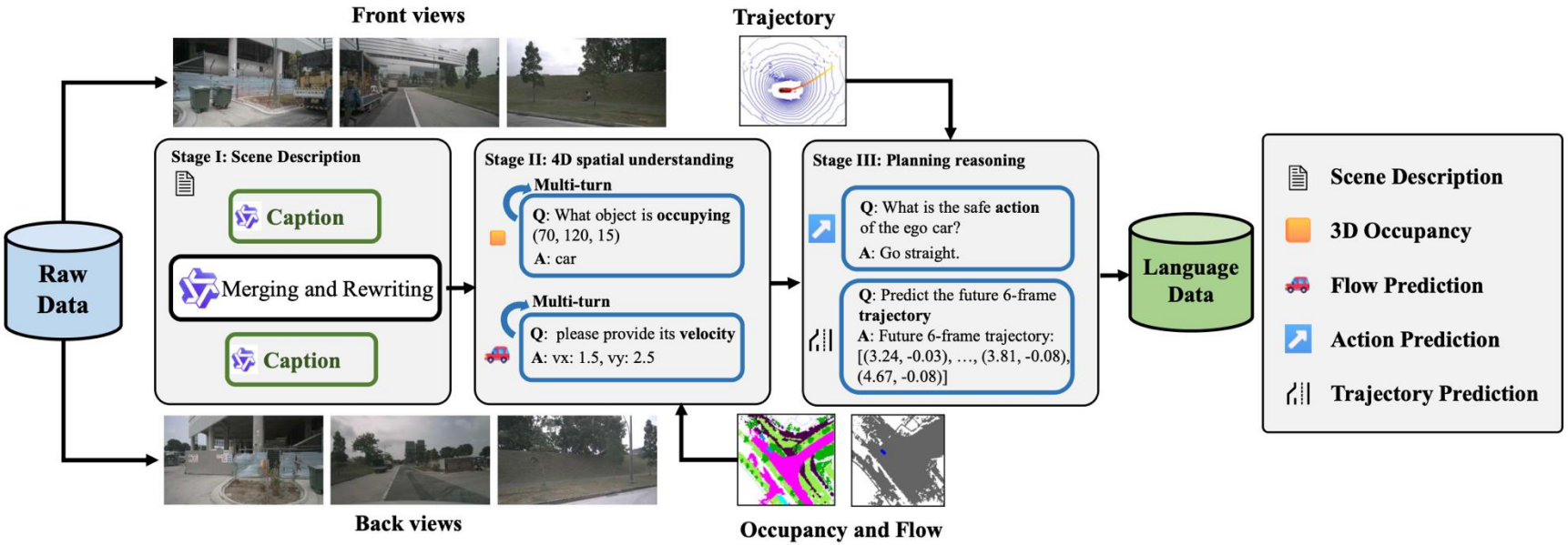
- (a) presents the pipeline of mainstream vision-based models (or vision-action models) for end-to-end autonomous driving.
- (b) illustrates mainstream Vision-Language-Action (VLA) frameworks.
- (c) DrivePI combines coarse-grained linguistic understanding with fine-grained 3D perception and prediction, inheriting advantages from both vision-based and VLA approaches.

DrivePI



- A vision encoder to extract features from images and LiDAR data, obtaining latent BEV features
- Then convert BEV features into vision tokens by a spatial projector.
- Feed both vision tokens and text tokens into the MLLM to generate output tokens.
- MLLM produces responses through four specialized heads: a text head for scene understanding in an auto-regressive manner, a 3D occupancy head for accurate spatial perception, an occupancy flow head for pixel-level motion prediction, and an action diffusion head for trajectory planning.

Data Engine



1） Generate captions of front and back views, respectively

2） Use InternVL3-78B to combine these captions to merge and polish generated scene descriptions

3） Generate various instruction QA pairs based on occupancy and flow ground truth by multi-turn conversations to improve the 4D spatial understanding ability

4) Generate planning QA pairs to allow MLLM to predict the future actions of ego-vehicle

Experiments: Perception, Prediction, Planning

Table 1. 3D occupancy and occupancy flow performance on the OpenOcc validation set.

Method	VLM-based	OccScore↑	RayIoU (3D Occ.)↑	mAVE (Occ. Flow)↓	RayIoU_{1m}	RayIoU_{2m}	RayIoU_{4m}
OccNeRF [66]		28.5	31.7	–	16.6	29.3	49.2
RenderOcc [40]		33.0	36.7	–	20.3	32.7	49.9
LetOccFlow [35]		36.4	40.5	–	25.5	39.7	56.3
OccNet [49]		35.7	39.7	–	29.3	39.7	50.0
BEVDetOcc-SF [15]		33.0	36.7	1.420	31.6	37.3	41.1
FB-Occ [26]		39.2	39.0	0.591	32.7	39.9	44.4
F-Occ [68]		41.0	39.9	0.491	33.9	40.7	45.2
CascadeFlow [29]		40.9	39.6	0.470	33.5	40.3	45.0
ALOcc-Flow-3D [6]		43.0	41.9	0.556	35.6	42.8	47.4
DrivePI (Ours)	✓	**49.3**	**49.3**	**0.509**	**45.0**	**50.0**	**52.9**

Table 2. Planning performance on the nuScenes validation set. Note that our unified model DrivePI does not incorporate ego status during training by default to avoid potential shortcut learning.

Method	VLM-based	Ego Status	*L2 (m)*↓				*Col. (%)*↓			
			1s	2s	3s	avg.	1s	2s	3s	avg.
ST-P3 []			1.33	2.11	2.90	2.11	0.23	0.62	1.27	0.71
FF [13]			0.55	1.20	2.54	1.43	0.06	0.17	1.07	0.43
EO [21]			0.67	1.36	2.78	1.60	0.04	0.09	0.88	0.33
UniAD [14]			0.48	0.96	1.65	1.03	0.05	0.17	0.71	0.31
VAD [20]			0.41	0.70	1.05	0.72	0.07	0.17	0.41	0.22
VAD [20]		✓	0.17	0.34	0.60	0.37	0.07	0.10	0.24	0.14
OmniDrive [53]	✓	✓	0.14	0.29	0.55	0.33	0.00	0.13	0.78	0.30
ORION [9]	✓	✓	0.17	0.31	0.55	0.34	0.05	0.25	0.80	0.37
OpenDriveVLA-7B [71]	✓	✓	0.20	0.58	1.21	0.66	0.00	0.22	0.55	0.25
DrivePI (Ours)	✓		**0.24**	**0.46**	**0.78**	**0.49**	**0.38**	**0.27**	**0.48**	**0.38**
DrivePI (Ours)	✓	✓	**0.19**	**0.36**	**0.64**	**0.40**	**0.00**	**0.05**	**0.28**	**0.11**

Experiments: Understanding

Ext., Cnt., Obj., Sts., Cmp. and Acc. are short for exist, count, object, status, comparison, and the overall accuracy.

Method	Ext.↑	Cnt.↑	Obj.↑	Sts.↑	Cmp.↑	Acc.↑
LLaMA-AdapV2 [11]	19.3	2.7	7.6	10.8	1.6	9.6
LLaVA1.5 [34]	45.8	7.7	7.8	9.0	52.1	26.2
LiDAR-LLM [63]	74.5	15.0	37.8	45.9	57.8	48.6
BEVDet+BUTD [41]	83.7	20.9	48.8	52.0	67.7	57.0
OpenDriveVLA-0.5B [71]	83.9	22.0	50.2	57.0	68.4	58.4
OpenDriveVLA-3B [71]	84.0	22.3	50.3	56.9	68.5	58.5
OpenDriveVLA-7B [71]	84.2	**22.7**	49.6	54.5	**68.8**	58.2
DrivePI (Ours)	**85.3**	22.4	**57.5**	**59.1**	68.3	**60.7**

Text Understanding performance on the nuScenes-QA validation set.

Ablation Study: Text and vision head

Table 5. Ablation study for text head and vision head in DrivePI.

#	Text Head	Vision Head	*3D Occ.* RayIoU↑	*Occ. Flow* mAVE↓	*Planning* L2↓	*Planning* Col.↓	*QA* Acc.↑
I	✓	–	–	–	–	–	**61.2**
II	–	✓	47.5	0.69	1.02	0.39	–
III	✓	✓	**49.3**	**0.51**	**0.50**	**0.38**	60.7

Table 7. Ablation study for the balancing weights in DrivePI.

#	*Occ. Weight*	*Flow Weight*	*3D Occ.* RayIoU↑	*Occ. Flow* mAVE↓	*Planning* L2↓	*Planning* Col.↓	*QA* Acc.↑
I	0.2	0.2	48.1	0.57	0.46	0.19	**61.1**
II	0.5	0.5	**49.3**	0.54	0.49	0.40	60.9
III	1.0	1.0	**49.3**	**0.51**	**0.50**	**0.38**	60.7

Ablation Study: Text data scaling

Table 6. The ablation study of DrivePI exploring data scaling. The columns of *Occ. Status*, *Occ. Class*, *Occ. Flow*, *Action Status* denote the occupancy status (*i.e.*, "yes" or "no"), occupancy category, the occupancy flow, the action commands (*i.e.*, "straight", "right", "left", and "stop"), respectively.

Model Size	Training Data size					*Occ. Status*	*Occ. Class*	*Occ. Flow*	*Action Status*	*Planning*		*QA*
	Caption	OCC.	Flow	Action	QA	Acc.↑	Acc.↑	mAVE↓	Acc.↑	L2↓	Col.↓	Acc.↑
0.5B	84k	420k	140k	24k	377k	86.0	50.4	0.91	**83.8**	**0.79**	0.63	60.7
3B	84k	28k	–	–	–	73.0	14.3	–	–	–	–	
	84k	56k	–	–	–	74.2	22.4	–	–	–	–	–
	84k	280k	–	–	–	86.8	54.7	–	–	–	–	–
	84k	560k	–	–	–	87.0	59.2	–	–	–	–	–
	84k	420k	140k	24k	377k	**88.6**	**59.8**	**0.69**	83.3	0.86	**0.62**	**63.0**

1） 84K captions + 28K occupancy QA pairs： only 73% accuracy in occupied status prediction and 14.3% accuracy in occupancy category prediction.

2） Scaling the occupancy QA pairs from 28K to 560K: improvement with 14% accuracy on predicting occupancy status and 44.9% accuracy on predicting occupancy category.

Some Findings

Table 8. The learned importance weights of all hidden states in the MLLM with Qwen-2.5 0.5B model, including the input embedding (indexed as 0). The ***Index*** and **Weight** column indicates the index and the learned importance weight of each hidden state.

Index	**Weight**	***Index***	**Weight**	***Index***	**Weight**
0	0.0328	*10*	0.0375	*20*	0.0463
1	0.0332	*11*	0.0381	*21*	0.0466
2	0.0337	*12*	0.0388	*22*	0.0472
3	0.0341	*13*	0.0397	*23*	0.0477
4	0.0346	*14*	0.0409	*24*	0.0468
5	0.0350	*15*	0.0422		
6	0.0355	*16*	0.0435		
7	0.0360	*17*	0.0449		
8	0.0365	*18*	0.0455		
9	0.0370	*19*	0.0458		

Table 9. The learned importance weights of all hidden states in the MLLM with Qwen-2.5 3B model, including the input embedding (indexed as 0). The ***Index*** and **Weight** column indicates the index and the learned importance weight of each hidden state.

Index	**Weight**	***Index***	**Weight**	***Index***	**Weight**
0	0.0254	*13*	0.0259	*26*	0.0277
1	0.0251	*14*	0.0260	*27*	0.0283
2	0.0251	*15*	0.0261	*28*	0.0287
3	0.0252	*16*	0.0262	*29*	0.0294
4	0.0253	*17*	0.0263	*30*	0.0297
5	0.0252	*18*	0.0264	*31*	0.0303
6	0.0253	*19*	0.0265	*32*	0.0304
7	0.0254	*20*	0.0266	*33*	0.0304
8	0.0255	*21*	0.0271	*34*	0.0302
9	0.0255	*22*	0.0265	*35*	0.0303
10	0.0256	*23*	0.0251	*36*	0.0331
11	0.0257	*24*	0.0258		
12	0.0258	*25*	0.0267		

Visualization

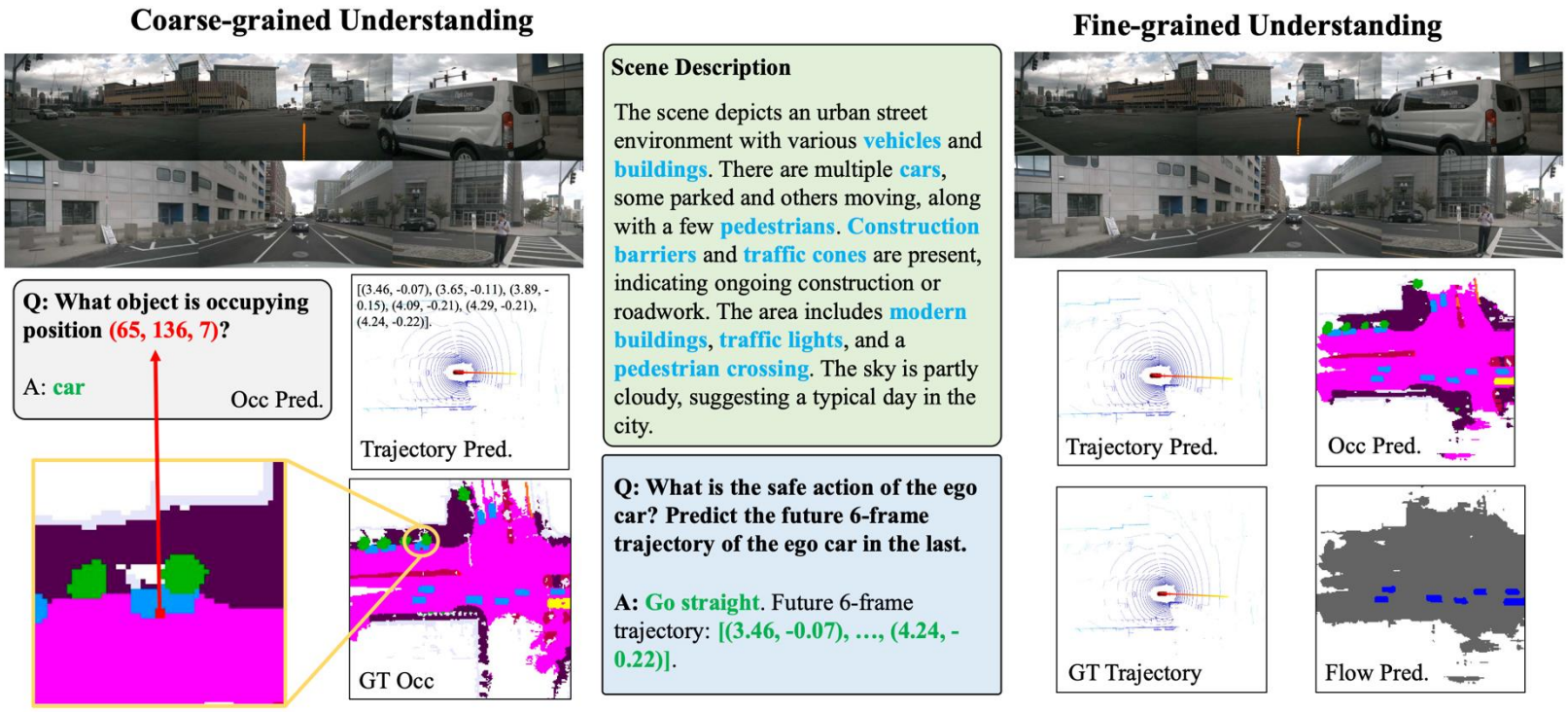


Figure 4. Visualization of coarse-grained scene descriptions and fine-grained 3D perception, prediction and planning of DrivePI.

Coarse-grained Perception, Prediction, and Planning

Task 1: Caption

The scene depicts a busy urban street with several vehicles, both parked and moving. There are multiple cars, some stopped at intersections and others in motion. Pedestrians are visible walking along the sidewalks. Construction cranes and buildings under construction are present in the background, indicating ongoing development. Traffic cones are placed on the road, likely for directing traffic or marking construction areas. The environment appears to be a bustling city area with a mix of modern architecture and active traffic.

Task 2: Occupancy and Flow prediction

Question: What can be seen at position $< OCC > (159, 78, 4) < /OCC >$? If there is an object, please provide its name and predict the velocity; otherwise, answer 'free'

Answer: {label: car}, {vx: 0.1, vy: -0.07}

Task 3: Action decision

Question: What is the safe action of the ego car?

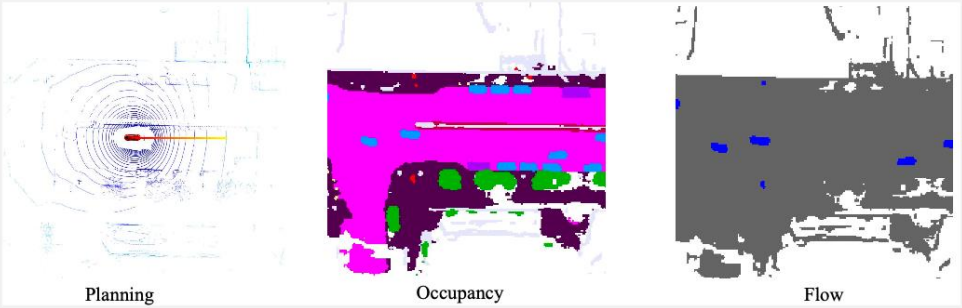
Answer: Go straight.

Task 4: Trajectory prediction

Question: Predict the future 6-frame trajectory of the ego car in the last.

Answer: Future 6-frame trajectory: [(3.42, -0.03), (3.52, -0.02), (3.59, -0.04), (3.72, -0.07), (3.87, -0.10), (3.88, -0.09)].

Fine-grained Perception, Prediction, and Planning



Thanks!