VLP: Vision Language Planning for Autonomous Driving

Supplementary Material

1. Experiment Setup

Model Components. UniAD is composed of a BEV extractor followed by five transformer decoder-based P3 modules for 3D object tracking, mapping, motion forecasting, occupancy prediction, and planning, respectively. VAD is composed of a BEV extractor followed by four transformer decoder-based modules for 3D object detection, scene segmentation, motion forecasting, and planning, respectively.

Model Hyperparameters. We provide the detailed training configurations for UniAD-based and VAD-based models in the Tab. 1.

2. Ablation Study

Prompt Information. We present ablation studies on prompt information for UniAD and VAD, accompanied by downstream task results in Tab. 5 and 6. Notably, training with completed ground truth information embedded in the prompts consistently yields superior performance across all downstream tasks for both UniAD and VAD. In comparison to the baseline (1st line in each table), models equipped with VLP, regardless of the ground truth information included, consistently outperform the model without VLP. These observations underscore the efficacy of the proposed VLP in enhancing model performance.

Different LMs. We explored several pretrained LMs for integration with our VLP. As shown in Tab. 2, each integration led to consistent improvements in decision-making over two baselines, proving the effectiveness of VLP design (leverages LM locally and globally). Also, it underscores our contribution of integrating language understanding into visual models for autonomous driving, irrespective of the specific choice of the LM. Our choice was driven by a balance between computational efficiency and effectiveness. Integration with LLAMA will be our future work.

ALP and SLP. As shown in Tab. 3, both SLP and ALP consistently improve all tasks. SLP focuses more on the final planning as it works for the ego-car query while ALP works for the shared BEV map. When ALP and SLP operate concurrently, the model puts more efforts on optimizing the final planning. While there is a relatively smaller margin of improvement in scene segmentation and occupancy prediction, the VLP still significantly surpasses the baseline (+VLP: +5.9%). Trajectory planning is the most important step which guarantees the safety and efficiency of ADS. Improvement on it is the main target of VLP.

3. Long-tail Generalization for 3D Object Detection

Tab. 4 presents the generalization ability of each VLP component on long-tail cases for 3D Object Detection. The results highlight the efficacy of each component in mitigating the long-tail detection problem. Particularly, the inclusion of SLP leads to a noticeable improvement over the baseline, and the combined utilization of both SLP and ALP further enhances the generalization ability. The long-tail classes, including construction vehicles, buses, motorcycles, bicycles, and trailers, constitute approximately 6% of the nuScenes dataset.

4. Visualization

We present several qualitative comparisons with the baseline in Fig. 1-7, using green arrows to highlight areas where our model outperforms the baseline. The visual comparison illustrates that our Vision Language Planning (VLP) framework help navigate the self-driving car in a more efficient and safer way.

5. Why human-like?

Humans interpret the visual scene as contextual cognitive semantics instead of plain digits, enabling them to navigate in unseen environments and recognize rare objects. We imbue the visual system with a more human-like and intuitively expected feature space via inspiring the reasoning and decision-making processes with designed prompts and pretrained LMs. During inference, our system retains the same enriched feature space and robust capabilities, which is evidenced by the marked improvements in the new-city generalization and long-tail tracking/detection.

References

[1] Bo Jiang, Shaoyu Chen, Qing Xu, Bencheng Liao, Jiajie Chen, Helong Zhou, Qian Zhang, Wenyu Liu, Chang Huang, and Xinggang Wang. Vad: Vectorized scene representation for efficient autonomous driving. *arXiv preprint arXiv:2303.12077*, 2023. 2

Configs	UniAD-based	VAD-based
point cloud range	$[-51.2, -51.2, -5.0, 51.2, 51.2, 3.0]$	$[-15.0, -30.0, -2.0, 15.0, 30.0, 2.0]$
transformer decoder dimension	256	256
BEV size	200×200	100×100
queue length	3	3
motion predict steps	12	6
motion predict modes	6	6
planning steps	6	6
backbone	RN101	RN50
optimizer	AdamW	AdamW
learning rate	2e-4	2e-4
weight decay	0.01	0.01
epoch	20	60
batch size	8	8

Table 1. **Training configurations for UniAD-based and VAD-based models.**

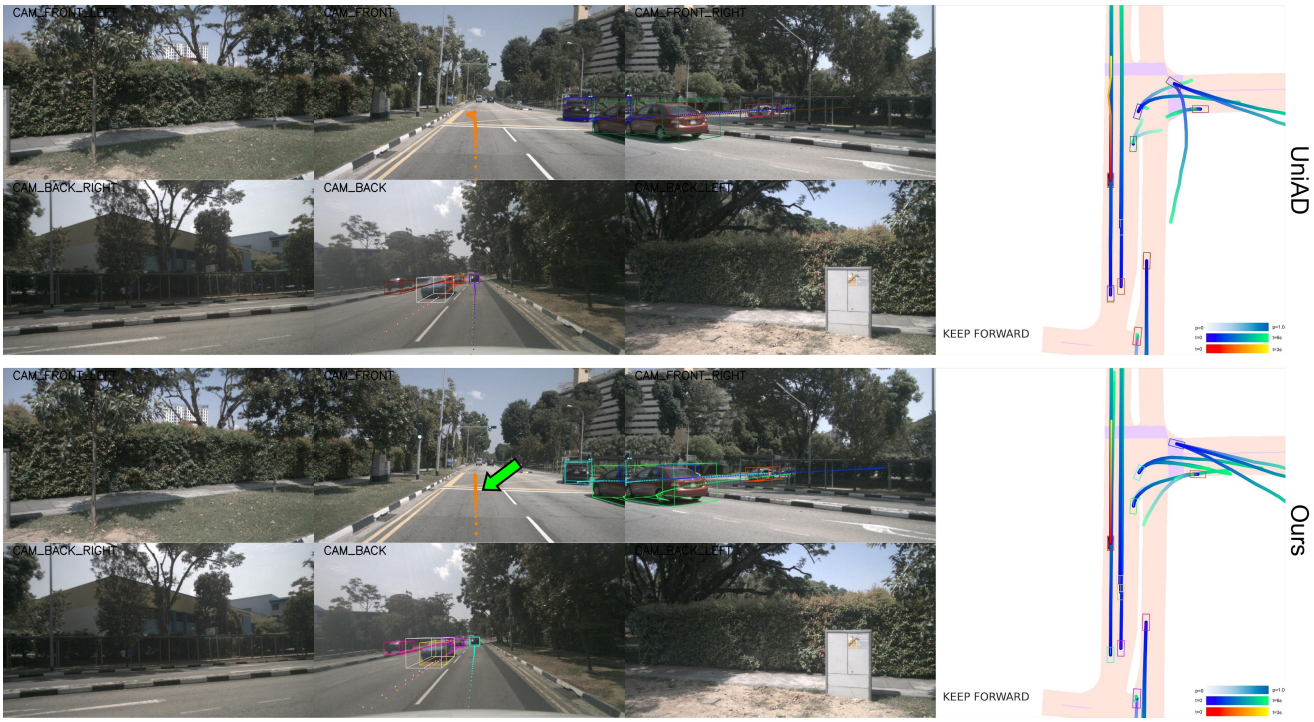

Figure 1. **Qualitative comparison between UniAD and Ours.** Green arrow is used to highlight areas where our VLP outperforms the baseline. The results indicate that our VLP enables the self-driving car to navigate more efficiently and safely.

Pretrained LM	Planning avg.L2 ↓	Planning avg.Col ↓	Pretrained LM	Planning avg.L2 ↓	Planning avg.Col ↓
UniAD			VAD		
-	1.03	0.31	-	0.82	0.93
+VLP			+VLP		
GPT2	0.75 (+27.2%)	0.16 (+48.4%)	GPT2	0.61 (+25.6%)	0.29 (+68.8%)
CLIP-RN50x64-LM	0.74 (+28.2%)	0.16 (+48.4%)	CLIP-RN50x64-LM	0.62 (+24.4%)	0.30 (+67.7%)
CLIP-RN101-LM	0.76 (+26.2%)	0.21 (+32.3%)	CLIP-RN101-LM	0.57 (+30.5%)	0.35 (+62.4%)
CLIP-ViT-L/14-336px-LM	0.73 (+29.1%)	0.24 (+22.6%)	CLIP-ViT-L/14-336px-LM	0.64 (+22.0%)	0.45 (+51.6%)

Table 2. Open-loop planning results with various LMs.

VLP	Vectorized Scene Seg. mIoU↑	Occupancy Pred. IoU-n.↑	IoU-f.↑	VPQ-n.↑	VPQ-f.↑	Plan Val. avg.L2 ↓	avg.Col ↓	Plan Singapore avg.L2 ↓	avg.Col ↓
-	39.8	63.4	40.2	54.7	33.5	0.78	0.38	0.78	0.39
+SLP	47.2	64.2	40.7	55.8	34.5	0.50	0.23	0.66	0.25
+ALP	47.6	64.5	41.0	56.1	34.7	0.52	0.26	0.68	0.28
+VLP	45.7	64.1	40.2	55.9	34.1	0.55	0.15	0.63	0.20

Table 3. Ablation study on each component of VLP.

Model	SLP	ALP	mAP↑	NDS↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
VAD [1]			17.6	29.9	0.79	0.33	0.95	0.65	**0.17**
VLP-VAD	✓		**19.1**	**31.2**	**0.76**	0.33	1.00	**0.50**	0.25
VLP-VAD	✓	✓	**20.4**	**33.2**	**0.75**	0.33	**0.83**	**0.60**	0.19

Table 4. **3D Object detection in challenging long-tail scenarios.**

Included Ground Truth				Tracking			Mapping		Motion Forecasting			Occupancy Prediction				Planning	
Label	Bbox	Trajectory	Command	AMOTA↑	AMOTP↓	IDS↓	IoU-lane↑	IoU-road↑	minADE↓	minFDE↓	MR↓	IoU-n.↑	IoU-f.↑	VPQ-n.↑	VPQ-f.↑	avg.L2↓	avg.Col.↓
✗	✗	✗	✗	35.2	1.353	720	29.5	66.7	0.83	1.24	0.187	59.5	38.5	49.9	29.4	0.93	0.38
✗	✓	✓	✓	35.5	1.331	670	30.8	66.8	0.78	1.23	0.180	59.8	39.1	51.2	30.7	0.87	0.31
✓	✗	✓	✓	35.9	1.331	650	31.2	67.5	0.75	1.19	0.162	61.9	38.5	53.1	31.4	0.82	0.32
✓	✓	✗	✓	35.4	1.340	710	31.1	67.2	0.81	1.20	0.179	62.1	38.7	53.8	31.5	0.89	0.36
✓	✓	✓	✗	36.1	1.329	610	31.6	68.9	0.76	1.12	0.168	61.7	38.9	54.1	31.9	0.80	0.29
✓	✓	✓	✓	36.2	1.320	620	31.7	69.1	0.72	1.11	0.156	62.4	39.3	54.2	32.9	0.78	0.24

Table 5. **Detailed ablations on the effectiveness of prompt information with UniAD-based models.** Our ablation studies on prompt information for UniAD reveal a significant performance boost when training with completed ground truth information embedded in the prompts.

Included Ground Truth				3D Object Detection		Vectorized Scene Segmentation				Motion Forecasting				Planning	
Label	Bbox	Trajectory	Command	mAP↑	NDS↑	Boundary↑	Divider↑	Crossing↑	mIoU↑	minADE↓	minFDE↓	MR↓	EPA↑	avg.L2↓	avg.Col.↓
✗	✗	✗	✗	22.6	33.3	43.3	43.6	34.0	40.3	0.86	1.20	0.143	0.526	0.98	0.80
✗	✓	✓	✓	24.9	35.7	45.2	42.5	32.9	40.9	0.76	1.06	0.127	0.522	0.64	0.23
✓	✗	✓	✓	25.7	37.6	46.5	44.4	30.4	40.4	0.78	1.05	0.119	0.549	0.56	0.30
✓	✓	✗	✓	27.0	38.7	44.9	40.0	34.8	40.0	0.86	1.19	0.141	0.532	0.59	0.36
✓	✓	✓	✗	26.6	38.0	44.0	40.3	31.4	38.9	0.79	1.06	0.121	0.551	0.59	0.26
✓	✓	✓	✓	27.3	39.0	47.1	45.3	36.0	42.8	0.77	1.05	0.117	0.551	0.52	0.17

Table 6. **Detailed ablations on the effectiveness of prompt information with VAD-based models.** Our ablation studies on prompt information for VAD shows a considerable performance improvement when training with completed ground truth information embedded in the prompts.

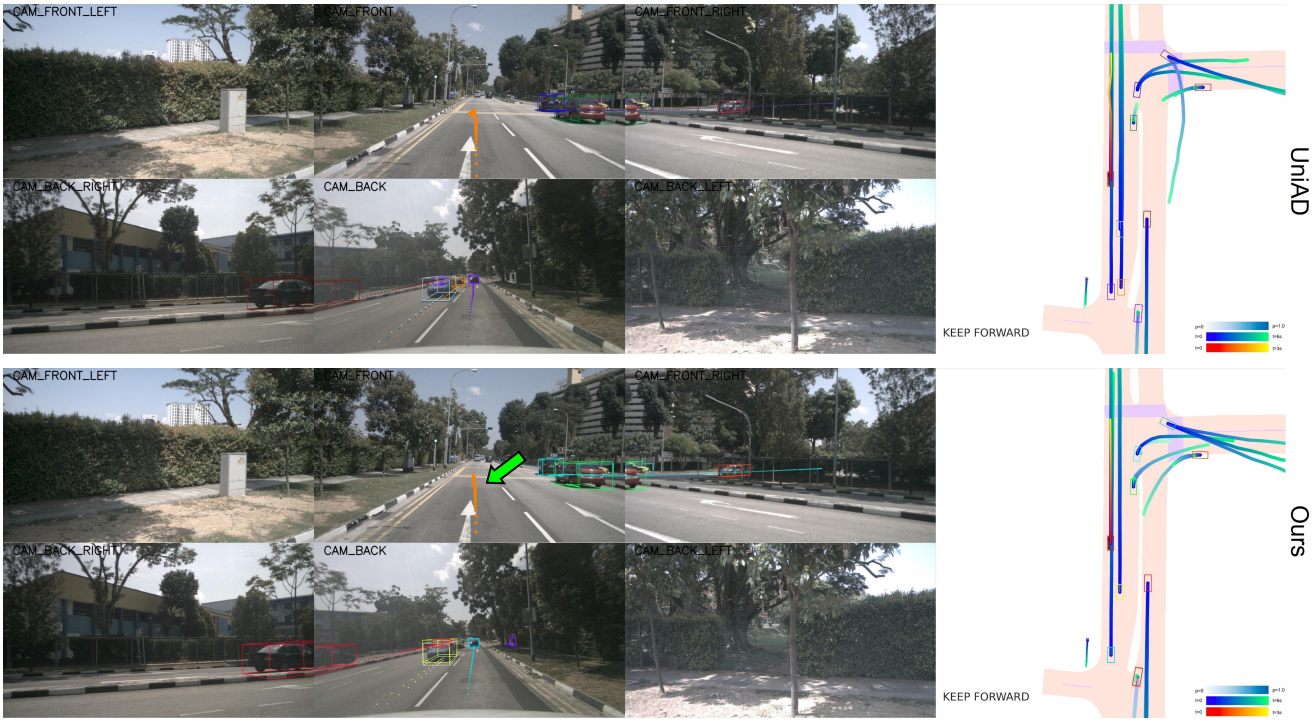

Figure 2. **Qualitative comparison between UniAD and Ours.** Green arrow highlights areas where our VLP outperforms the baseline.

Pretrained LM	Tracking			Mapping		Motion Forecasting			Occupancy Prediction				Planning	
	AMOTA↑	AMOTP↓	IDS↓	IoU-lane↑	IoU-road↑	minADE↓	minFDE↓	MR↓	IoU-n.↑	IoU-f.↑	VPQ-n.↑	VPQ-f.↑	avg.L2↓	avg.Col.↓
-	35.2	1.353	720	29.5	66.7	0.83	1.24	0.187	59.5	38.5	49.9	29.4	0.93	0.38
CLIP-RN50x64-LM	32.9	1.384	780	29.2	65.4	0.70	1.01	0.140	59.7	38.0	50.3	29.1	0.81	0.23
CLIP-RN101-LM	36.2	1.320	620	31.7	69.1	0.72	1.11	0.156	62.4	39.3	54.2	32.9	0.78	0.24
CLIP-ViT-L/14-336px-LM	32.6	1.358	923	29.8	67.1	0.70	1.01	0.138	59.0	37.6	49.1	28.4	0.57	0.87
GPT-2	34.1	1.379	784	29.8	67.0	0.72	1.03	0.148	60.6	38.3	50.3	28.9	0.93	0.22

Table 7. Detailed ablations investigating the impact of various language models on UniAD-based models.

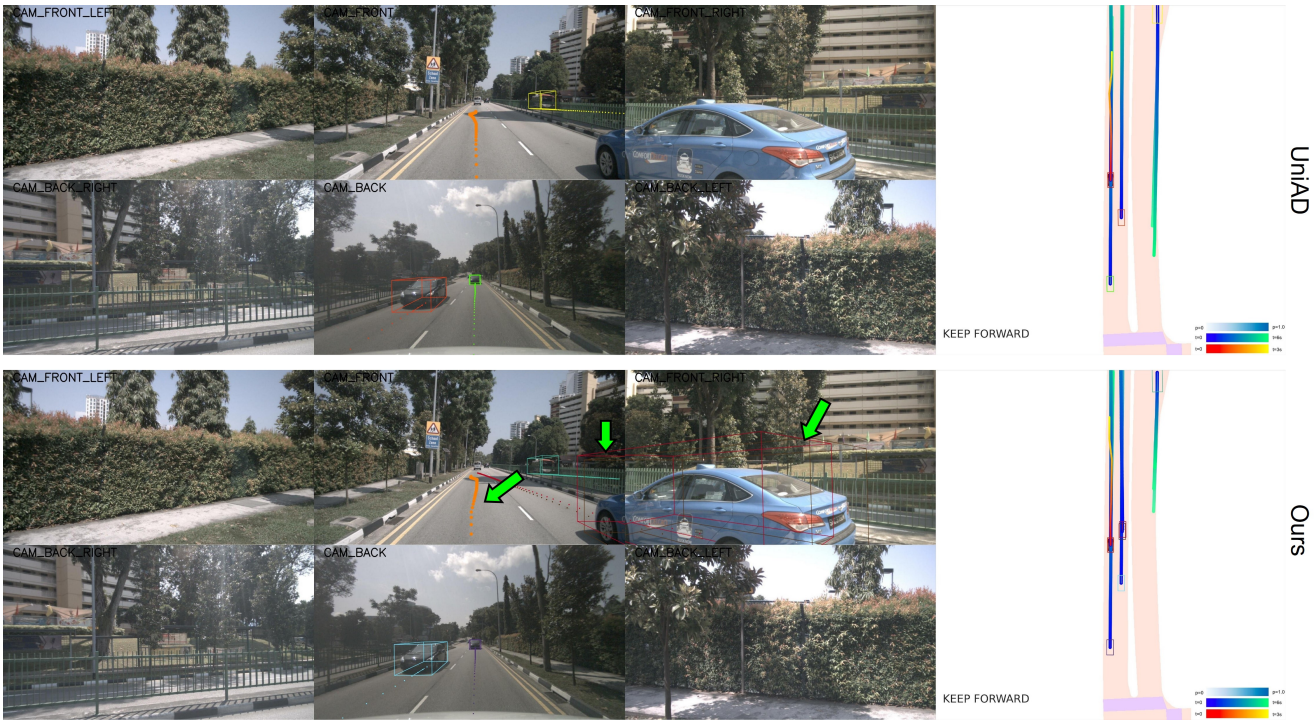

Figure 3. **Qualitative comparison between UniAD and Ours.** Green arrow highlights areas where our VLP outperforms the baseline.

Figure 4. **Qualitative comparison between UniAD and Ours.** Green arrow highlights areas where our VLP outperforms the baseline.

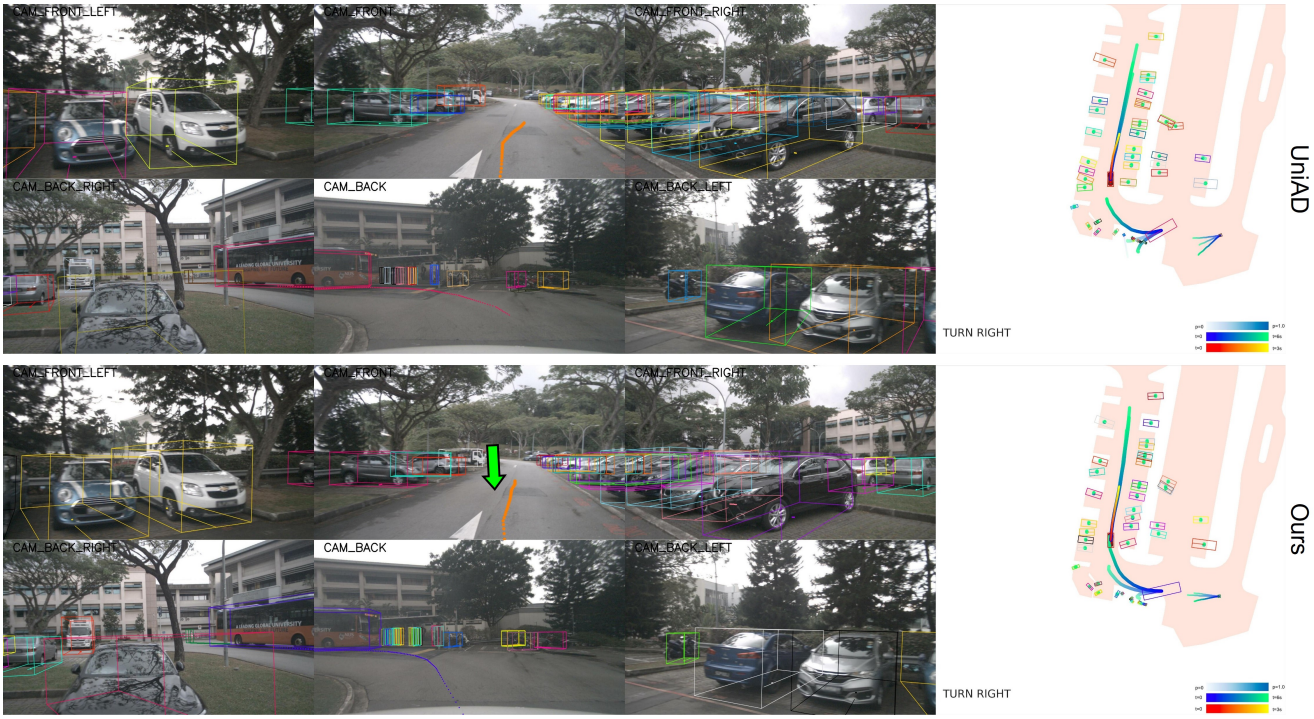

Figure 5. **Qualitative comparison between UniAD and Ours.** Green arrow highlights areas where our VLP outperforms the baseline.

Figure 6. **Qualitative comparison between UniAD and Ours.** Green arrow highlights areas where our VLP outperforms the baseline.

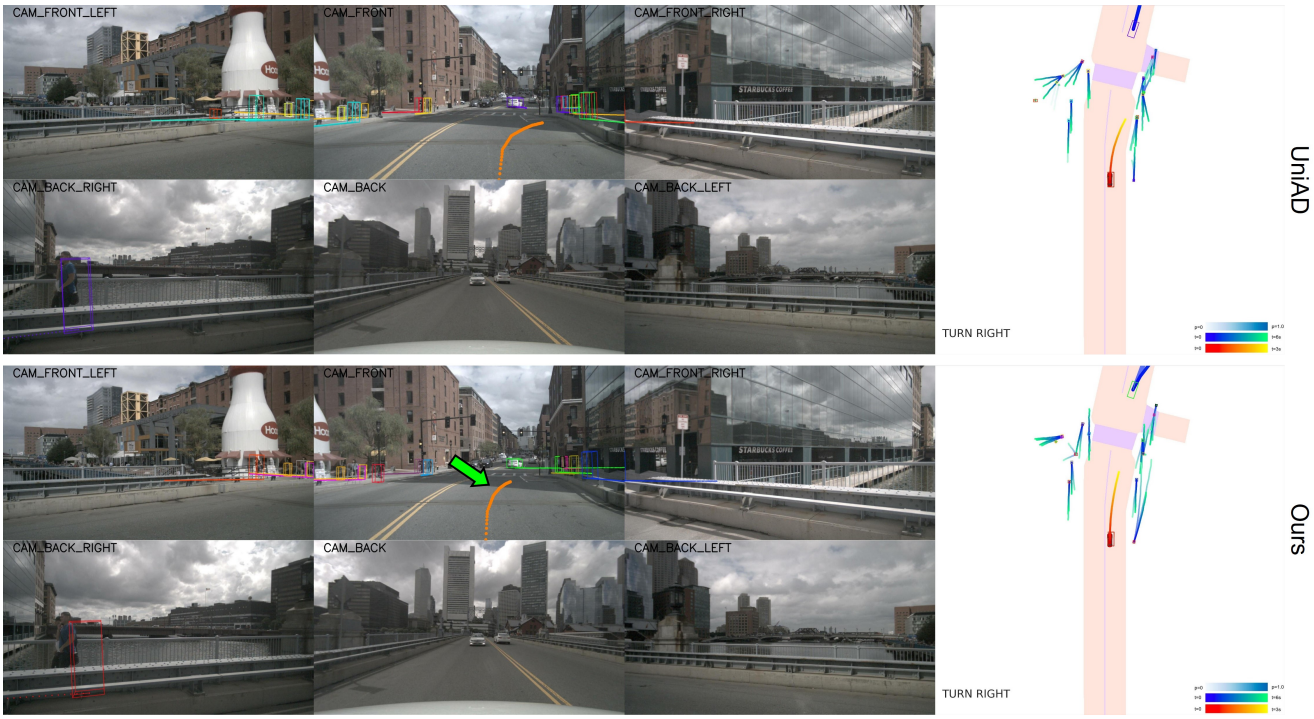

Figure 7. **Qualitative comparison between UniAD and Ours.** Green arrow highlights areas where our VLP outperforms the baseline.