A. Implementation Details

Model Details The feature dimension of all decoders in our framework is set as 256. We use $K_e = 1$, $K_{bev} = 3$, $K_{opy} = 3$, $C^K = 64$, $C^V = 256$ for the feature dimensions mentioned in Sec 3. The feature of the 5th stage in Resnet was used as the feature map f_i in the 2D backbone. We use Fully Connected Layer and Batch Normalization [30] to construct a simplified version of PointNet [48] to encode the information of raw LiDAR points in the 3D backbone.

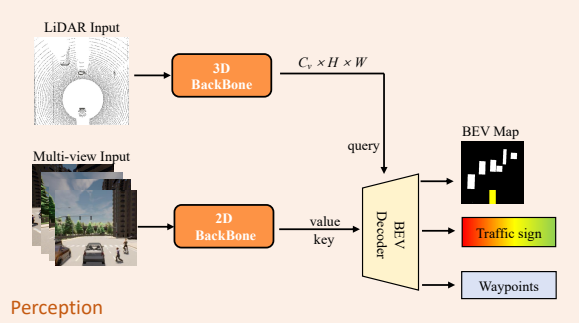

Figure 6. Overview of our pipeline for pretraining the perception module in the first training stage.

Training We train our models using the AdamW optimizer [44] and a cosine learning rate scheduler [43]. In the first training stage, the initial learning rate is set to $5e^{-4} \times \frac{BatchSize}{512}$ for the transformer encoder and the 3D backbone, and $2e^{-4} \times \frac{BatchSize}{512}$ for the 2D backbones. The weight decay is 0.07. We train the models for 35 epochs with the first 5 epochs for warm-up [27]. We used random scaling from 0.9 to 1.1 and color jittering to augment the collected RGB images. The overview of the first-stage framework can be found in Figure 6. In the second training stage, we freeze the perception module. The training schedule of the other two modules are similar to that in the first stage.

Sensors The RGB images are collected and cropped from one front-facing camera, two side-facing cameras, and one back-facing camera with a resolution of 800×600. Each camera has a 100° horizontal field of view (FOV), and the side cameras are angled at 60°. For the front image, we scale the shorter side of the front camera input to 256 and crop its center patch of 224×224. For the focusing-view image, we directly crop the center of the front camera input to get a 128×128 patch. For the other images, the shorter side of the camera input is scaled to 160 and a center patch of 128×128 is taken.

Other hyper-parameter values Some other hyper-parameter values used in ReasonNet are listed in Table 8.

B. Benchmark details

We evaluate our method on the CARLA public leaderboard [53], Town05 benchmark [47], and our proposed DOS benchmark. Adversarial events[3] are included in the first two benchmarks, and occlusion events are included in the last benchmark. In these benchmarks, the ego vehicle is required to complete a given route without collision or traffic rules violation.

CARLA Leaderboard The CARLA Autonomous Driving Leaderboard [53] is to evaluate the driving proficiency of autonomous agents in realistic traffic situations with a variety of weather conditions. The CARLA leaderboard provides a set of 76 routes for training and verifying agents and contains a secret set of 100 routes to evaluate the driving performance of the submitted agents.

Town05 benchmark In this benchmark, we use Town05 for evaluation and other towns for training. Following [47], the benchmark includes two evaluation settings: 1) Town05 Short: 10 short routes of 100-500m, each comprising 3 intersections, 2) Town05 Long: 10 long routes of 1000-2000m, each comprising 10 intersections. Town05 is a complex town with multi-lane roads, single-lane roads, bridges, highways and exits. The core challenge of the benchmark is how to handle dynamic dense agents and adversarial events.

CARLA 42 routes benchmark The CARLA 42 routes benchmark was proposed in NEAT [14], including six towns covering a variety of areas such as US-style intersections, EU-style intersections, freeways, roundabouts, stop signs, urban scenes and residential districts. The traffic density of each town is set to be comparable to busy traffic setting. We take the same configuration open-sourced by [47] when we evaluated the methods.

C. More Experimental results

In this section we report additional experimental results, including the CARLA leaderboard and two other benchmarks.

C.1. CARLA leaderboard

Table 5 shows the detailed comparison between our method and the baselines on the CARLA public Leaderboard [53]. Our method also leads the vehicle collision and offroad infraction numbers among all the methods.

C.2. Town05 and CARLA 42 routes

Table 6 and Table 7 additionally compare the driving score, road completion, and infraction score of the presented approach to prior state-of-the-art on the CARLA

[3] Adversarial events include unexpected agents rushing into the road from occluded regions, vehicles running red traffic lights, etc. Please refer to https://leaderboard.carla.org/scenarios/ for detailed descriptions.

Rank	Method	Driving Score	Route Completion	Infraction Score	Vehicle Collisions	Pedestrian Collisions	Layout Collisions	Red light Violations	Offroad Infractions	Blocked Infractions
1	ReasonNet (Ours)	**79.95**	89.89	**0.89**	**0.13**	0.02	0.01	0.08	**0.04**	0.33
2	InterFuser [51]	76.18	88.23	0.84	0.37	0.04	0.14	0.22	0.13	0.43
3	TCP [63]	75.14	85.63	0.87	0.32	**0.00**	**0.00**	0.09	**0.04**	0.54
4	LAV [10]	61.85	**94.46**	0.64	0.70	0.04	0.02	0.17	0.25	**0.10**
5	TransFuser [15]	61.18	86.69	0.04	0.71	0.81	0.01	**0.05**	0.23	0.43
6	Latent TransFuser [15]	45.20	66.31	0.72	1.11	0.02	0.02	**0.05**	0.16	1.82
7	GRIAD [8]	36.79	61.85	0.60	2.77	**0.00**	0.41	0.48	1.39	0.84
8	TransFuser+ [1]	34.58	69.84	0.56	0.70	0.04	0.03	0.75	0.18	2.41
9	Rails [9]	31.37	57.65	0.56	1.35	0.61	1.02	0.79	0.96	0.47
10	IARL [54]	24.98	46.97	0.52	2.33	**0.00**	2.47	0.55	1.82	0.94
11	NEAT [14]	21.83	41.71	0.65	0.74	0.04	0.62	0.70	2.68	5.22

Table 5. Comparison of our method and the state-of-the-art on the public CARLA leaderboard [53] (accessed Nov 2022). Methods are ranked by the driving score as the main metric. Driving Score, Route Completion, Infraction Score are higher the better, and the other metrics are lower the better. We outperform all other methods by a wide margin. We also lead the vehicle collision, offroad infraction numbers among all the methods.

	Town05 Short		**Town05 Long**	
Method	Driving Score ↑	Road Completion ↑	Driving Score ↑	Road Completion ↑
CILRS [18]	7.47±2.51	13.40±1.09	3.68±2.16	7.19±2.95
LBC [11]	30.97±4.17	55.01±5.14	7.05±2.13	32.09±7.40
TransFuser [47]	54.52±4.29	78.41±3.75	33.15±4.04	56.36±7.14
NEAT [14]	58.70±4.11	77.32±4.91	37.72±3.55	62.13±4.66
Roach [66]	65.26±3.63	88.24±5.16	43.64±3.95	80.37±5.68
WOR [9]	64.79±5.53	87.47±4.68	44.80±3.69	82.41±5.01
InterFuser [51]	94.95±1.91	95.19±2.57	68.31±1.86	94.97±2.87
ReasonNet (Ours)	**95.71**±1.88	**96.23**±3.17	**73.22**±1.91	**95.88**±2.31

Table 6. Comparison of our ReasonNet with six state-of-the-art methods in Town05 benchmark. Our method outperformed other strong methods in all metrics and scenarios.

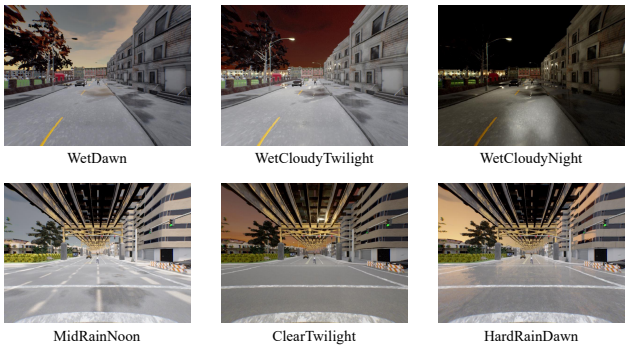

Figure 7. Different types of weather in our dataset.

Town05 benchmark [47] and CARLA 42 routes benchmark [14].

D. Data statistics

We describe the detailed statistics for each town and their corresponding maps in Table 9. In Figure 7, we show six types of weathers among our dataset. For the submission for the online leaderboard, the model is trained in all eight towns. For the ablation studies, we train the models on five towns (Town01, Town03, Town04, Town06 ,and Town07).

E. Videos

To investigate how different scenarios run in the CARLA simulator, we provide four videos for each scenario. The four videos are named **parked_cars.mp4**, **sudden_brake.mp4**, **red_light_infraction.mp4** and **left_turn.mp4**. For demonstration, we use the rule-based expert agent to control the ego car in these scenarios.

F. License of Assets

We use the open-source CARLA driving simulator [21]. CARLA is released under the MIT license. Its assets are under the CC-BY license. The pretrained ResNet model is under the MIT license. The source code for our work will be publicly available once accepted and they are under the CC-BY-NC 4.0 license.

Method	Driving Score ↑	Road Completion ↑	Infraction Score ↑
CILRS [18]	22.97±0.90	35.46±0.41	0.66±0.02
LBC [11]	29.07±0.67	61.35±2.26	0.57±0.02
AIM [47]	51.25±0.17	70.04±2.31	0.73±0.03
TransFuser [47]	53.40±4.54	72.18±4.17	0.74±0.04
NEAT [14]	65.17±1.75	79.17±3.25	0.82±0.01
Roach [66]	65.08±0.99	85.16±4.20	0.77±0.02
WOR [9]	67.64±1.26	90.16±3.81	0.75±0.02
InterFuser [51]	91.84±2.17	**97.12±1.95**	0.95±0.02
ReasonNet (Ours)	**93.25±2.91**	96.84±2.17	**0.96±0.02**

Table 7. Comparison of our ReasonNet with other methods in CARLA 42 routes benchmark. Our method outperformed other strong methods in driving score and infraction score.

Notation	**Description**	**Value**
	BEV Map and Controller	
a_{max}	Maximum acceleration	1.0 m/s
v_{max}	Maximum velocity	7.5 m/s^2
H, W	Size of the BEV map	50, 50
	Size of the BEV area	50 meter × 50 meter
H_b	The detection range for the backward of the ego vehicle	20
	Scale factor for bounding box size of pedestrians and bicycles	2
	Learning Process	
	Number of epochs	35
	Number of warm-up epochs	5
λ_{sign}	Weight for the traffic sign loss	0.2
λ_{w}	Weight for the waypoints loss	0.4
λ_{BEV}	Weight for the BEV map loss	0.4
λ_{opy}	Weight for the occupancy map loss	0.2
$\lambda_{consistency}$	Weight for the consistency loss	0.05
	Max norm for gradient clipping	10.0
	Weight decay	0.07
	Batch size	256

Table 8. The parameter used for ReasonNet.

Town Name	#Frames	Description
Town01	342846	A basic town layout consisting of "T junctions"
Town02	197240	Similar to Town01, but smaller
Town03	469115	The most complex town, with a 5-lane junction, a roundabout, unevenness, a tunnel, and more
Town04	429979	An infinite loop with a highway and a small town
Town05	297140	Squared-grid town with cross junctions and a bridge. It has multiple lanes per direction.
Town06	148495	Long highways with many highway entrances and exits. It also has a Michigan left
Town07	55299	A rural environment with narrow roads, barns and hardly any traffic lights
Town10	69039	A city environment with different environments such as an avenue or promenade

Table 9. Detailed statistics of the number of frames and a brief description of each town.