LMDrive: Closed-Loop End-to-End Driving with Large Language Models

Supplementary Material

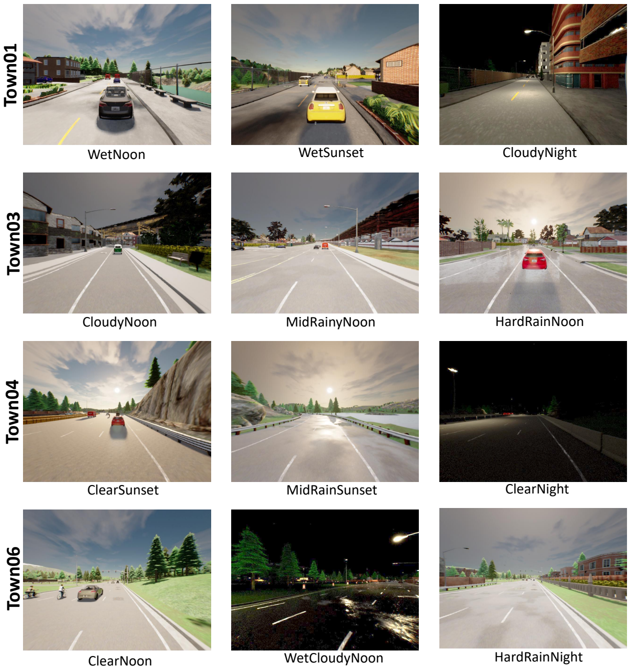

Figure 7. The visualization of some weather and daylight conditions used in the LangAuto Benchmark.

A. Implementation Details

Model details For the vision encoder, we use $K_{enc} = 1$ encoder layers and $K_{dec} = 3$ decoder layers. The feature of the 5-th stage in the ResNet is employed as the extracted feature map, then we apply an MLP layer to convert its dimension to 768, which is the feature dimension of the following Q-Former. Following LAV [3], we build a simplified version of PointNet [34] with several MLP layers and Batch Normalization layers to encode LiDAR point cloud data. For the Q-Former, we utilize the model architecture and pre-trained weights from the BLIP-2 [24]. In our work, the visual tokens fed into the Q-Former include 400 BEV tokens, 5 future waypoint tokens and 1 traffic light token.

Sensor configuration We use one front-facing camera, two side-facing cameras, and one back-facing camera to collect RGB images. Each camera has a resolution of 800 × 600 and a 100° horizontal field of view (FOV). The two side cameras are angled at 60°. For the font 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 which can capture distance traffic light status. For the other images, the shorter side of the camera input is scaled to 160 and then takes a 128 × 129 center patch. For the LiDAR sensor, the rotation frequency is 10Hz and the upper/lower field-of-view is 10/-30. The number of channels is 64.

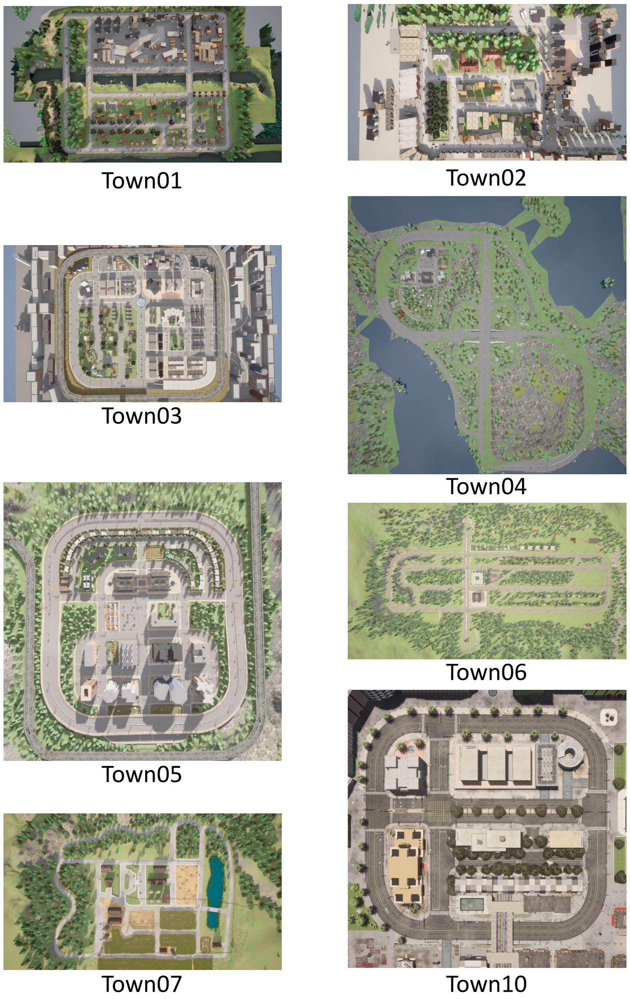

Figure 8. The visualization of eight town maps used in the LangAuto Benchmark.

Training details We first introduce the details of the vision encoder pretraining. We adopt the AdamW optimizer [31] and a cosine learning rate scheduler [30] for the training. The initial learning rate set for the transformer encoder and 3D backbone is $\frac{BatchSize}{512} \times 5e^{-4}$. And $\frac{BatchSize}{512} \times 2e^{-4}$ is the learning rate for the 2D backbone because the backbone is initialized with the ImageNet pre-train weights. We train the models for 35 epochs with the first 5 epochs for warm-up [17]. We used random scaling from 0.9 to 1.1 and color

Sample rate	DS ↑	RC ↑	IS ↑
1	49.5±1.5	58.5±2.7	0.83±0.03
2	**50.6±1.7**	**60.0±3.4**	**0.84±0.04**
4	46.0±2.1	59.5±2.7	0.79±0.03

Table 6. Ablation study on the different choices of the sample rate. The experiment is conducted on the LangAuto-Short benchmark with the backbone of LLaVA-v1.5.

jittering to augment the collected RGB images.

Then we introduce the training details in the instruction-finetuning stage. We also adopt the cosine learning rate scheduler and the initial learning rate is $1e^{-4}$ for a batchsize 32. We train the models for 15 epochs with the first 2000 iteration steps for warm-up. The weight decay is set to 0.07. The maximum historic horizon T_{max} is set to 40, and we will truncate the data clip to keep the recent 40 frames if its number of frames exceeds 40.

B. Additional Experiments

In this section, we delve further into our method by conducting additional ablation studies. First, we assess our methodology using a variety of sample rates in Table 6. The term ”sample rate” represents the fixed interval at which training frames are sampled. When the sample rate is set at 1, the horizon becomes narrow, which prevents the application of temporal augmentation, and consequently leads to a decrease in performance. Conversely, setting the sample rate at 4 can create an excessively large gap between two consecutive frames, which might obtain a poor driving score. A sample rate of 2 achieves a good trade-off.

Second, we conduct ablation studies on the usage rate of notice instructions as shown in Table 7. In our method, we randomly removed 75% notice instructions in the data clips to avoid overfitting. It’s worth noting that we use the LangAuto-Short here, rather than LangAuto-Notice, where the AV can not receive any notice instruction. We first remove all notice data, and get a worse driving score and infraction score. This suggests that incorporating notice instructions during training may improve the AV’s ability to both attend to and understand adverse events, thus reducing collisions. However, when we tried to include all notice data, we found it did not enhance performance. Utilizing 25% of the instructions achieves a good trade-off.

C. Efficiency Analysis

Benefiting from the Q-Former, only 4 tokens need to be processed in the LLM. Hence, the LLM doesn’t bring massive computational costs. We also use FlashAttention to speed up the LLM inference. Figure 9 shows our model (multi-frame input) has slightly larger latency than InterFuser (one-frame input), and outperforms ReasonNet (multi-frame input).

Notice Data %	DS ↑	RC ↑	IS ↑
0	45.2±2.8	**67.1±2.5**	0.68±0.03
25	**50.6±1.7**	60.0±3.4	**0.84±0.04**
100	49.1±1.9	58.2±2.4	0.83±0.04

Table 7. Ablation study on the usage rate of notice instructions. The experiment is conducted on the LangAuto-Short benchmark with the backbone of LLaVA-v1.5.

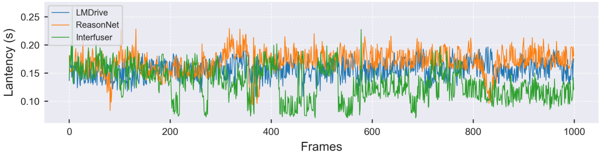

Figure 9. Latency comparison with SOTA methods. (InterFuser only takes one frame as input, ReasonNet and ours take multiple frames as input).

D. Benchmark Details

In Figure 7, we show the 12 different environmental conditions (16 conditions are used in our benchmark in total). In Figure 8, we show all 8 town maps we used in this work. In Table 8, we list the basic statistical information of the LangAuto benchmarks across various tracks (LangAuto, LangAuto-Short, LangAuto-Tiny).

Benchmark Type	LangAuto	LangAuto-Short	LangAuto-Tiny
Avg. Driving Distance (m)	635.8	305.9	122.4
Avg. Navigation Instructions	20.3	10.8	5.1
Avg. Notice Instructions	5.8	3.3	1.7

Table 8. Comparative Analysis of Different Benchmark Tracks

E. Instruction Details

Our work considers 56 different types of navigation and notice instructions, and we use ChatGPT to generate eight different phrases for the same navigation instruction. Table 9 shows an example where we generate eight phrases for the navigation instruction ’Turn Right’. We present the full list of 56 different types of navigation and notice instructions in Table 10. Table 11 demonstrates how we generate the misleading instructions. The first column is the driving scenario in which the agent is located, and the second column represents the possible misleading instructions generated by our framework. For example, when the AV is on a single-lane road, the misleading instruction “*Change your route to the left-hand lane*” violates the traffic rules and raises safety concerns. In Table 12, we show some examples of the connected instructions.

Instruction	**Eight instructions of one kind of instruction**
Turn right	After [x] meters, execute a right turn. After [x] meters, take a right. Right in [x] meters. After [x] meters, hang a right. In [x] meters, prepare to turn right. Continue for [x] meters, then turn right. Go [x] meters, then just take a right. After [x] meters, a right turn is required.

Table 9. Eight different phrazes of one "turn" instruction generated by the ChatGPT API.

Type	One randomly chosen instruction of each instruction type
Follow	Transition to the left lane for travel.
	You might want to switch to the right lane.
	In about [x] meters, you'll want to switch to the left lane.
	In [x] meters, reposition to the right lane.
	Keep going on this road, you're doing great!
	Continue on the highway.
	Maintain your course along this route for precisely [x] meters.
	Cruise down the highway for about [x] meters.
	Continue in a straight line along your current path.
	Keep going straight until you reach the next junction, you're on the right track!
	Preserve your current trajectory for exactly [x] meters.
	Stay straight for [x] meters until the next intersection.
	Veering to the left, prepare to enter the highway.
	Execute a right maneuver, prepare for highway exit.
	In [x] meters, slide left and plan to hop off the highway.
	In [x] meters, proceed to the right, prepare for immediate highway departure.
Turn	Prepare to turn left up ahead.
	Proceed ahead and make a right turn.
	Continue for [x] meters, then turn left.
	After [x] meters, execute a right turn.
	You're going to be turning left at the next junction, alright?
	It is mandatory to take a right turn at the imminent intersection.
	Straight through the next crossroads.
	After navigating [x] meters, a left turn at the intersection is obligatory.
	After moving forward [x] meters, prepare to make a right turn at the intersection.
	[x] meters more, then straight on at the intersection, piece of cake.
	When you reach the next traffic signal, you will need to turn left.
	Please execute a right turn upon reaching the upcoming traffic signal.
	Please maintain your course straight at the next traffic signal.
	Just another [x] meters, then you'll be turning left at the light, okay?
	After [x] meters, take a right at the light.
	After traversing [x] meters, it's crucial to continue straight at the traffic signal.
	Next T-junction, turn left.
	At the forthcoming T-intersection, execute a right turn.
	Just keep on going straight through the next T-junction, sound good?
	In [x] meters, hang a left at the T.
	After [x] meters, take a right at the T, no biggie.
	After a distance of [x] meters, maintaining a straight course at the T-intersection is imperative.
	Find your way out at the first exit on the roundabout, please.
	Depart at the second exit on the roundabout.
	Shoot out on the third exit.
Others	Feel free to start driving.
	Please implement an immediate reduction in your speed.
	Hit the brakes, stop now.
	Drive freely.
	Please navigate towards the designated point, which is [x] meters in front of you and [y] meters to your left/right.
Notice	Please watch out for the pedestrians up ahead.
	Attention is required for the bicycle ahead.
	Watch for the car that's just stopped up front.
	Be mindful of the vehicle crossing on a red light to your left.
	Car ran red light ahead.
	Please be alert of the uneven road surface in the vicinity ahead.
	Watch for the tunnel coming up.
	Just a heads up, there's a red light ahead.
	Green light ahead.
	Be mindful of the yellow light ahead.

Table 10. Full list of the 56 different types of navigation instructions and notice instructions considered in our framework.

Driving Scenarios	**One randomly chosen instruction of each misleading type**
Single-lane Road	Change your route to the left-hand lane.
	Transition to the right lane for travel.
	Proceed ahead and make a left turn.
	A right turn is required up ahead.
	Hang a left at the next crossroads.
	Depart at the first exit on the roundabout.
	Roll out on the second exit.
	It is necessary for you to take the third exit on the roundabout.
Non-highway Road	Maintain your course along the highway.
	Slide left and plan to hop on the highway.
	Ease on to the right and get set to quit the highway.
Non-roundabout Road	First exit.
	No sweat, just hit the second exit on the roundabout.
	It is necessary for you to take the third exit on the roundabout.
Non-intersection Road	Just up ahead, take a left.
	Your next action is a right turn, just ahead.
	Execute a left maneuver, prepare for highway entry.
	Right, ready to exit.
	Carefully navigate to the first exit as you approach the roundabout.
	You are required to take the second exit on the roundabout.
	Third exit.
T-Intersection (Left Turn Prohibited)	You'll be turning left at the next T-junction, alright?
	Transition to the left lane for travel.
	You might want to switch to the right lane.
T-Intersection (Right Turn Prohibited)	A right turn is mandatory at the upcoming T-intersection.
	Change your route to the left-hand lane.
	Just head for the right lane.
When turning	Please adjust your course to the left-most lane.
	Reposition to the right lane.

Table 11. Full list of the misleading instructions and their corresponding driving scenarios.

Driving Scenarios	**Examples of connected instructions**
Two consecutive instructions	(1) Prepare to turn right up ahead. (2) Proceed along this route.
	(1) Maintain your course along this route. (2) Left, ready to enter.
Three consecutive instructions	(1) It's critical to keep straight at the forthcoming T-intersection. (2) Keep cruising down this road. (3) At the next traffic signal, you should make a right turn.
	(1) At the forthcoming T-intersection, execute a right turn. (2) Just head for the left lane.(3) Maintain your course along this route.
	(1) Keep going on this road, you're doing great! (2) Right ahead. (3) Right, ready to exit.

Table 12. Examples of the connected navigation instructions considered in the LangAuto benchmark.