

RoboTron-Drive: All-in-One Large Multimodal Model for Autonomous Driving

Supplementary Material

A. Implementation Details

In this section, we elaborate on the more detailed implementations for RoboTron-Drive and the experiments in Sec. 3.1.

Model architecture. We adapt SigLIP [13] as the vision encoder, which is pre-trained on WebLI [3] with a resolution of 384×384 . We use a 2-layer MLP [8] as the projector to project the image features into the word embedding space. For the language model, we choose Llama-3.1 [10] 8B, which uses a tokenizer with a vocabulary of 128K tokens. Our model is trained on sequences of 8,192 tokens. In particular, the vision encoder SigLIP $\mathcal{F}_e(\cdot)$ encodes the input images $X_v \in \mathbb{R}^{(n \times f) \times h \times w \times 3}$ into the visual features:

$$Z_v = \mathcal{F}_e(X_v), \quad (1)$$

where $Z_v \in \mathbb{R}^{(n \times f) \times h' \times w' \times d'}$, n and f denote the number of cameras and frames, (h, w) and (h', w') denote the size of image and feature, and d' denotes the channel dimensionality. X_v can represent the data formats mentioned above, for example, $n > 1$ and $f > 1$ for video from multi-view cameras. For LiDAR data, we project the point clouds onto the BEV or range view to convert the data into a single image format. Afterward, the projector projects MLP $\mathcal{F}_p(\cdot)$ the image features into the word embedding space:

$$H_v = \mathcal{F}_p(Z_v), \quad (2)$$

where the projector is implemented using a 2-layer MLP, $H_v \in \mathbb{R}^{(n \times f \times h \times w) \times d}$ denotes a sequence of visual tokens, and d denotes the dimensionality of the word embedding space in LLM Llama-3.1 $\mathcal{F}_l(\cdot)$. Based on the visual tokens H_v and the user instruction X_t , the LLM computes the probability of the target word step by step:

$$p(Y_t | H_v, X_t) = \prod_{i=1}^L \mathcal{F}_l(Y_{t,i} | H_v, \Phi(X_t), Y_{t,0:i-1}), \quad (3)$$

where $\Phi(\cdot)$ refers the text tokenizer, $Y_{t,i}$ and $Y_{t,0:i-1}$ represent the i th word and the preceding $i - 1$ words in Y_t , and L indicates the length of the words generated by LLM.

Dataset. This part further introduces the evaluation datasets applied in Sec. 4.1, including six datasets for general capability, three datasets for testing generalization abilities, and one dataset designed for planning tasks.

- DriveLM, OmniDrive and NuInstruct are annotated based on the nuScenes[1] dataset and contain 376,181, 374,329 and 71,842 samples respectively, suitable for multi-view

images or videos input. The CODA-LM dataset, designed for corner case question-answer pairs on the CODA [6] dataset, includes a total of 184,480 samples in both Chinese and English, as a single-view dataset. MAPLM, a multi-view image dataset, contains 94,970 samples, while LingoQA, a single-view video dataset, comprises 413,829 samples. Note the sample number is computed after our enhancement. Additionally, each dataset includes a specific number of test samples: DriveLM has 15,480 test samples, OmniDrive has 72,184 test samples, NuInstruct has 16,147 test samples, CODA-LM has 2,123 test samples, MAPLM has 6,642 test samples, and LingoQA has 500 test samples.

- In zero-shot setting, we use BDD-X, DRAMA and DriveBench dataset. BDD-X is composed of over 77 hours of driving within 6,970 videos. The test set of BDD-X consists of 698 driving videos. The videos are taken in diverse driving conditions, *e.g.* day/night, highway/city/countryside, summer/winter *etc.* On average 40 seconds long, each video contains around 3-4 actions, *e.g.* speeding up, slowing down, turning right *etc.*, all of which are annotated with a description and an explanation. DRAMA dataset is captured from a moving vehicle on highly interactive urban traffic scenes in Tokyo. It contains different annotations: Video-level Q/A, Object-level Q/A, Risk object bounding box, Free-form caption, and separate labels for ego-car intention, scene classifier and suggestions to the driver. The DRAMA test set comprises 2544 samples. DriveBench is a benchmark dataset designed to assess VLM reliability across 17 settings, including clean, corrupted, and text-only inputs. It comprises 19,200 frames, 20,498 question-answer pairs, three question types, and four mainstream driving tasks, evaluating 12 popular VLMs.
- We evaluate our model in the clean setting and compare it with all models in the dataset, which covers four mainstream driving tasks: perception, prediction, planning, and behavior. We utilize GPT-4 for GPT Score evaluation in BDD-X and DRAMA, while employing GPT-3.5-turbo for DriveBench. We use nuScenes to evaluate the planning task, which is a challenging and popular benchmark in the AD. The dataset is a multi-sensor dataset with 1,000 scenes and each scene lasts for 20 seconds. There are 6,019 validation samples.

Experiment setting. Here we describe the training details of RoboTron-Drive. We adopt a curriculum learning approach to progressively train RoboTron-Drive as introduced in Sec. 3.4.

- *Stage-1: Language-image alignment.* We use LCS-558K [9] to align the visual patch features into the word embedding space. During this stage, we train only the projector while keeping the other components frozen. The learning rate is set to 1×10^{-3} and the training is conducted for 1 epoch with a batch size of 512.
- *Stage-2: Single-image pre-training.* In this stage, we use single-image data to improve model’s image comprehension capability. We utilize the recaptioned BLIP558K [9], COCO118K [4] and CC3M [12] datasets to improve the model. Meanwhile, we use language data Evo-Instruct [2] to balance the model’s language understanding ability. At this stage, the dataset comprises 3M single-image data and 143K language data. We fine-tune the entire model using a batch size of 256 for 1 epoch. The learning rate for the vision encoder is set to 2×10^{-6} , while the learning rate for both the projector and LLM is 1×10^{-5} .
- *Stage-3: Multi-capacity pre-training.* In this stage, our primary objective is to enhance the model’s reasoning and perception capabilities, and equip the model with the ability to handle diverse data formats. To achieve this, we use various multimodal data and perception data, including 1.5M single images, 760K multi-view images, 501K single videos, and 145K multi-view videos. Specifically, the single-image data consists of the multimodal data from [4] and the perception data from COCO [7] and Object365 [11]. The multi-view image data includes the multimodal data LLaVA-NeXT-Interleave [5] and the perception data nuScenes [1]. The single-view video data is derived from the works [5, 14]. Given the scarcity of multi-view videos in the available data, we generate the multi-view perception data using nuScenes [1], with each view consisting of 5 frames. We fine-tune the entire model with a batch size of 256 for 1 epoch, maintaining the same learning rates as Stage-2.
- *Stage-4: Driving fine-tuning.* In the final stage, we employ a diverse array of high-quality driving datasets to fine-tune RoboTron-Drive for AD tasks. We compile six public AD datasets, which include single image (CODA-LM), multi-view images (MAPLM, DriveLM), single video (LingoQA), and multi-view videos (OmniDrive, NuInstruct), amounting to a total of 1.5M. Specifically, we augment the CODA-LM dataset, expanding it from 36,896 to 184,480 samples, and the MAPLM dataset from 47,485 to 94,970 samples. Additionally, we standardize the DriveLM and NuInstruct datasets to ensure uniformity across the data. In this stage, we use the same batch size and learning rate as Stage-2.

B. Prompt Design

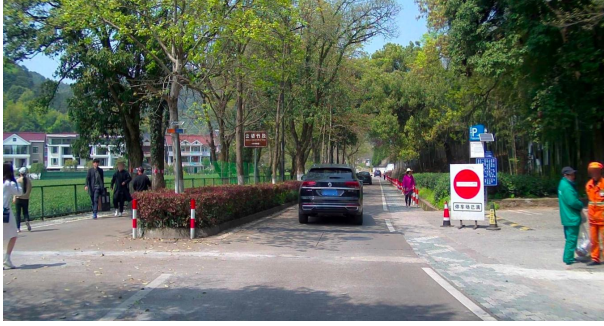
```
Messages = [
  {"role": "system", "content": f" " "You are an English improver." " "},
  {"role": "user", "content": f" " "I have a question and its corresponding answer. I need your assistance in revising and refining them. Please make some changes to the written content while preserving the meaning. The question and answer that require modifications are: {QA}. Please provide the revised question and answer in the format: Question: <question> Answer: <answer>." " "}]
```

Table S1. The prompt for question-answer augmentation.

In the question-answer augmentation, we use the prompt presented in Tab. S1 for GPT-4o-mini to enhance and improve the question-answer pairs.

C. Qualitative Results

Here we present the qualitative results for analyzing the distinctions among various models. In Fig. S1-S16, we provide the visualization examples from each dataset, illustrating the general performance on the six benchmarks and the generalization ability in BDD-X. The specialist model is trained on individual dataset. The all-in-one model RoboTron-Drive is able to accurately identify and analyze key objects related to driving in images and videos, thereby enhancing its ability to handle diverse autonomous driving data and tasks.

**User**

<image>. There is an image of traffic captured from the front view of the ego vehicle. Focus on objects influencing the ego car's driving behavior: vehicles (cars, trucks, buses, etc.), vulnerable road users (pedestrians, cyclists, motorcyclists), traffic signs (no parking, warning, directional, etc.), traffic lights (red, green, yellow), traffic cones, barriers, miscellaneous(debris, dustbin, animals, etc.). You must not discuss any objects beyond the seven categories above. Please describe each object's appearance, position, direction, and explain why it affects the ego car's behavior.

RoboTron-Drive

In the traffic image, there is a **black SUV** directly in front of the ego car, occupying the same lane and moving in the same direction. The presence of this SUV is significant as it dictates the following distance and speed that the ego car must maintain to ensure safe driving conditions. Additionally, there are **pedestrians on the sidewalk** to the right of the ego car, with one individual wearing a bright orange outfit and standing close to the curb. The proximity of these pedestrians to the road means that the ego car must remain vigilant and be prepared to stop or slow down if any of them unexpectedly enter the roadway. On the right side of the road, there is a **'No Entry' sign** for vehicles, accompanied by a blue sign indicating a detour or alternative route. This sign is crucial as it informs the ego car that it cannot proceed straight ahead in its current lane and must instead follow the detour to continue its journey. Furthermore, there are **orange traffic cones** placed on the right side of the road, partially obstructing the sidewalk and the edge of the road. These cones suggest that there may be road work or an obstruction ahead, which could affect the available driving space and necessitate caution from the ego car. There are no traffic lights, barriers, or other objects reported in the image that affect the driving behavior of the ego car.

Specialist

In the traffic image, there is a **dark-colored SUV** directly in front of the ego car, occupying the same lane and moving in the same direction. The presence of this vehicle requires the ego car to maintain a safe following distance and be prepared to adjust speed or change lanes if necessary. On the right side of the road, there are **pedestrians** walking parallel to the road. The ego car must be cautious and ready to stop or slow down in case any pedestrian steps onto the road. Additionally, there is a **cyclist** on the right side of the road, moving in the same direction as the ego car. The ego car should give the cyclist enough space and be prepared to slow down or change lanes if the cyclist's path intersects with the car's path. There is also a **'No Entry' sign** on the right side of the road, indicating that the road ahead is closed to traffic. The ego car must not proceed forward and should look for an alternative route. Furthermore, there are traffic cones placed on the right side of the road, indicating a work zone or a closed area. The ego car should be cautious and may need to merge left if the right lane is obstructed. There are no traffic lights, barriers, or other objects present in the image that affect the driving behavior of the ego car.

GPT-4o

1. **Vehicle (Car)** - **Appearance**: **Dark-colored SUV**. - **Position**: Directly ahead of the ego car in the same lane. - **Direction**: Moving forward. - **Influence**: The ego car must maintain a safe following distance and adjust its speed according to the SUV's speed. 2. **Vulnerable Road Users (Pedestrians)** - **Appearance**: Several pedestrians, some walking and some standing. - **Position**: On the left side of the road, near the sidewalk, and one pedestrian on the right side near the "Do Not Enter" sign. - **Direction**: Walking along the sidewalk and crossing the road. - **Influence**: The ego car must be cautious and prepared to stop if any pedestrian decides to cross the road. 3. **Traffic Sign (Do Not Enter)** - **Appearance**: Red circle with a white horizontal line and the text "DO NOT ENTER." - **Position**: On the right side of the road. - **Direction**: Facing the ego car. - **Influence**: The ego car must not enter the road indicated by the sign and should continue straight or follow any other permissible route. 4. **Miscellaneous (Debris)** - **Appearance**: Small debris scattered on the road. - **Position**: On the road surface, near the center. - **Direction**: Stationary. - **Influence**: The ego car should avoid the debris to prevent potential damage to the vehicle and ensure a smooth drive.

Figure S1. Visualization of CODA-LM. Key information is highlighted in green, while errors are marked in red.

CODA-LM



User

<image>. There is an image of traffic captured from the front view of the ego vehicle. Please describe the object inside the red rectangle in the image and explain why it affect ego car driving.

RoboTron-Drive

This object is a **traffic sign displaying a red circle with a white bar**, indicating a prohibition of entry for vehicles in the direction the sign is facing. The ego car must not proceed forward or turn in the direction indicated by this sign to comply with traffic rules.

Specialist

This is a traffic sign, specifically indicating **a no parking or stopping zone**. It is important for the ego car to recognize this sign to avoid stopping or parking in this area, which could result in a traffic violation.

GPT-4o

The object inside the red rectangle is **a pedestrian**. The presence of the pedestrian affects the ego car's driving because the car must slow down or stop to ensure the safety of the pedestrian, adhering to traffic laws and preventing potential accidents.

Figure S2. Visualization of CODA-LM.

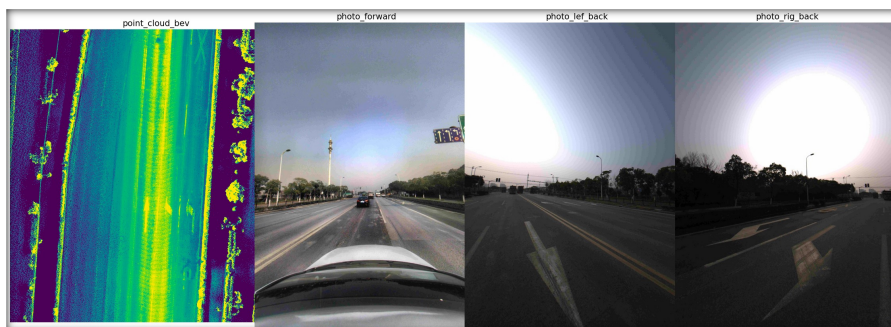
CODA-LM



User	<image>. There is an image of traffic captured from the front view of the ego vehicle. Focus on objects influencing the ego car's driving behavior: vehicles (cars, trucks, buses, etc.), vulnerable road users (pedestrians, cyclists, motorcyclists), traffic signs (no parking, warning, directional, etc.), traffic lights (red, green, yellow), traffic cones, barriers, miscellaneous(debris, dustbin, animals, etc.). You must not discuss any objects beyond the seven categories above. Please provide driving suggestions for the ego car based on the current scene.
RoboTron-Drive	The ego car should maintain a safe following distance from the vehicle ahead, adhere to the 20 km/h speed limit, and be prepared to slow down or stop for the road workers. It should also be ready to merge left if necessary due to the traffic cones indicating a potential lane closure.
Specialist	The ego car should maintain a safe following distance from the vehicles ahead, be prepared to slow down or stop for the road workers, and adhere to the 20 km/h speed limit. It should also be ready to change lanes if necessary, while being cautious of the cyclist and giving them ample space.
GPT-4o	Based on the image, here are the driving suggestions for the ego car: 1. Vehicles : There is a car ahead in the same lane. Maintain a safe following distance and be prepared to adjust speed based on the car's movements. 2. Vulnerable Road Users : There are pedestrians on the right side of the road. Drive cautiously and be prepared to stop if any pedestrian steps onto the road. 3. Traffic Signs : There are no visible traffic signs directly influencing the ego car's behavior. 4. Traffic Lights : There are no traffic lights visible in the image. 5. Traffic Cones : There are traffic cones on the right side of the road, indicating a potential work zone or hazard. Slow down and proceed with caution. 6. Barriers : There are no barriers directly influencing the ego car's behavior. 7. Miscellaneous : There are no miscellaneous objects directly influencing the ego car's behavior. Driving Suggestions : - Maintain a safe following distance from the car ahead. - Slow down and be prepared to stop for pedestrians. - Proceed with caution near the traffic cones, as they may indicate a work zone or hazard.

Figure S3. Visualization of CODA-LM.

MAPLM



User 1: <image> 2: <image> 3: <image> 4: <image>. These four images are the point cloud bev, camera front view, camera back left view, camera right view of the ego vehicle. How many lanes on the current road?

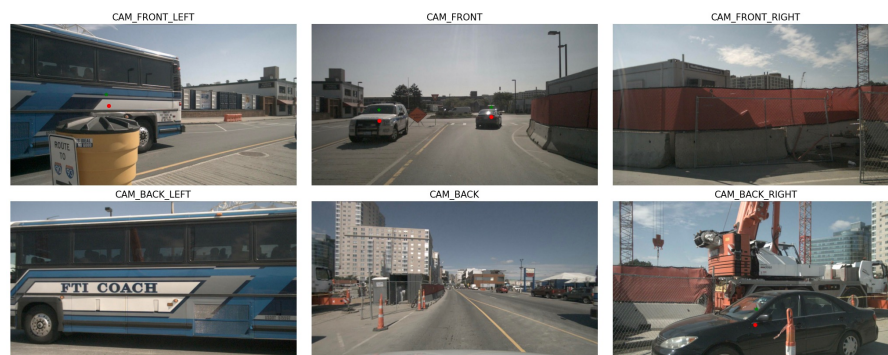
RoboTron-Drive 3.

Specialist 4.

GPT-4o The current road has **two lanes**.

Figure S4. Visualization of MAPLM.

DriveLM



User 1: <image> 2: <image> 3: <image> 4: <image> 5: <image> 6: <image>. These six images are the front view, front left view, front right view, back view, back left view and back right view of the ego vehicle. What are the important objects in the current scene? Those objects will be considered for the future reasoning and driving decision.

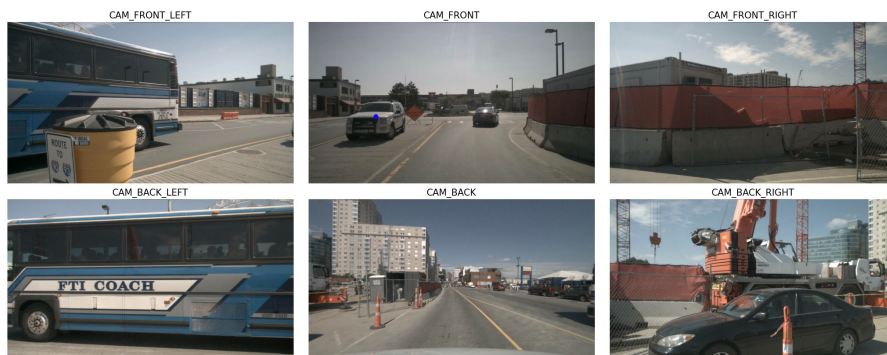
RoboTron-Drive There is a black car to the back right of the ego vehicle, a black sedan to the front of the ego vehicle, a white SUV to the front of the ego vehicle, and a blue and white bus to the front left of the ego vehicle. The IDs of these objects are <c1,CAM_BACK_RIGHT,800.8,599.288>, <c2,CAM_FRONT,1004.192,463.736>, <c3,CAM_FRONT,375.008,479.288>, and <c4,CAM_FRONT_LEFT,535.792,393.336>.

Specialist There is a black car to the back right of the ego vehicle, a black sedan to the front of the ego vehicle, a white SUV to the front of the ego vehicle, and a blue and white bus to the front left of the ego vehicle. The IDs of these objects are <c1,CAM_BACK_RIGHT,794.2,687.5>, <c2,CAM_FRONT,1004.2,517.5>, <c3,CAM_FRONT,377.5,539.2>, and <c4,CAM_FRONT_LEFT,548.3,455.0>.

GPT-4o The important objects in the current scene that should be considered for future reasoning and driving decisions are: 1. Vehicles (including the **white car, black car, and buses**) 2. Construction barriers and cones 3. Construction equipment (e.g., crane) 4. Road signs and traffic signs 5. Buildings and structures along the road These objects are crucial for understanding the driving environment and making safe driving decisions.

Figure S5. Visualization of DriveLM.

DriveLM



User 1: <image> 2: <image> 3: <image> 4: <image> 5: <image> 6: <image>. These six images are the front view, front left view, front right view, back view, back left view and back right view of the ego vehicle. Would <c3,CAM_FRONT,373.7,526.2> be in the moving direction of the ego vehicle?

RoboTron-Drive No.

Specialist No.

GPT-4o Yes, <c3,CAM_FRONT,373.7,526.2> would be in the moving direction of the ego vehicle. The CAM_FRONT camera captures the front view, which aligns with the direction the vehicle is moving.

Figure S6. Visualization of DriveLM.

LingoQA



User <image>. There is a video of traffic captured from the front view of the ego vehicle. Is there a traffic light in the vicinity? If so, what color is it displaying?

RoboTron-Drive Yes, there is a traffic light in the vicinity and it is displaying red.

Specialist Yes, there is a traffic light and it is currently displaying red.

GPT-4o Yes, there is a traffic light in the vicinity. It is displaying red.

Figure S7. Visualization of LingoQA.

LingoQA



User

<image>. There is a video of traffic captured from the front view of the ego vehicle. What is the current action and its justification? Answer in the form \"action, justification\".

RoboTron-Drive

I am **maintaining speed** and steering right to keep a safe distance to the adjacent lane on the left.

Specialist

I am **accelerating** because there is a gap in the traffic behind the car ahead, and it is safe to do so.

GPT-4o

Driving, clear road ahead.

Figure S8. Visualization of LingoQA.

OmniDrive



User

1: <image> 2: <image> 3: <image> 4: <image> 5: <image> 6: <image>. These six videos are the front view, front left view, front right view, back view, back left view and back right view of the ego vehicle. Is there any indication of a special event happening near the dome-like structure?

RoboTron-Drive

There is no clear indication of a special event happening near the dome-like structure, but the presence of **tents and a parked truck** suggests some activity.

Specialist

Here is no clear indication of a special event happening near the dome-like structure, but the presence of **tents and parked vehicles** suggests some sort of activity or event

GPT-4o

Yes, there is an indication of a special event happening near the dome-like structure. The presence of **large tents and temporary structures** suggests that an event might be taking place.

Figure S9. Visualization of OmniDrive.

OmniDrive



User 1: <image> 2: <image> 3: <image> 4: <image> 5: <image> 6: <image>. These six videos are the front view, front left view, front right view, back view, back left view and back right view of the ego vehicle. Given the proximity of other vehicles, what would be a safe driving practice right now?

RoboTron-Drive A safe driving practice would be to **remain stationary**, maintain a safe distance from the vehicle ahead, and be prepared to proceed when the traffic light changes.

Specialist A safe driving practice would be to maintain a safe following distance from the vehicle ahead, **stay alert** for any sudden stops or changes in traffic flow, and be prepared to react accordingly.

GPT-4o Given the proximity of other vehicles, a safe driving practice right now would be to **remain stationary** and wait for a clear and safe opportunity to move. This will help avoid potential collisions with the nearby vehicles.

Figure S10. Visualization of OmniDrive.

NuInstruct



User 1: <image> 2: <image> 3: <image> 4: <image> 5: <image> 6: <image>. These six videos are the front view, front left view, front right view, back view, back left view and back right view of the ego vehicle. What is the closest object in the right of the ego car? If any, please list all objects as format <class>[c,x1,y1,x2,y2], class is the class label, c indicates the camera view, and x1, y1, x2, y2 are the coordinates of the top-left and bottom-right corners of the bounding box.

RoboTron-Drive <car>[CAM_FRONT_RIGHT,33.000,51.000,66.000,80.000];

Specialist <pedestrian>[CAM_FRONT_RIGHT,84.000,45.000,91.000,68.000];

GPT-4o <Car><3, 264, 144, 432, 288>

Figure S11. Visualization of NuInstruct.

NuInstruct



User

1: <image> 2: <image> 3: <image> 4: <image> 5: <image> 6: <image>. These six videos are the front view, front left view, front right view, back view, back left view and back right view of the ego vehicle. How many <vehicle>s in the back of the ego car?

RoboTron-Drive

1

Specialist

2

GPT-4o

There is **one vehicle** in the back of the ego car.

Figure S12. Visualization of NuInstruct.

NuInstruct



User

1: <image> 2: <image> 3: <image> 4: <image> 5: <image> 6: <image>. These six videos are the front view, front left view, front right view, back view, back left view and back right view of the ego vehicle. Does any objects approach the ego car? If any, please list all objects as format <class>[c,x1,y1,x2,y2], class is the class label, c indicates the camera view, and x1, y1, x2, y2 are the coordinates of the top-left and bottom-right corners of the bounding box.

RoboTron-Drive

<truck>[CAM_BACK_RIGHT,56.250,0.000,100.000,100.000];

Specialist

There are **no objects** that would have risks for the ego car.

GPT-4o

1. **Person**[3, 200, 150, 300, 400] 2. **Person**[3, 300, 150, 400, 400]

Figure S13. Visualization of NuInstruct.

NuInstruct



User

1: <image> 2: <image> 3: <image> 4: <image> 5: <image> 6: <image>. These six videos are the front view, front left view, front right view, back view, back left view and back right view of the ego vehicle. What is the next motion for the ego? And give the reason. If any, please list all objects as format <class>[c,x1,y1,x2,y2], class is the class label, c indicates the camera view, and x1, y1, x2, y2 are the coordinates of the top-left and bottom-right corners of the bounding box.

RoboTron-Drive

There are<car>[CAM_FRONT,70.812,51.000,100.000,100.000];<car>[CAM_FRONT_RIGHT,0.000,50.556,12.750,100.000]; approach the ego car. However, they all keep a safe distance from ego vehicles. Hence, The ego would keep speed and go straightforward.

Specialist

There are no objects that would have risks for the ego car. Hence, The ego would keep speed and go straightforward.

GPT-4o

The next motion for the ego vehicle should be to proceed straight. Reason: The front view image shows a green traffic light, indicating that it is safe and legal for the ego vehicle to move forward.

Figure S14. Visualization of NuInstruct.

BDD-X



User

<image>. There is a video of traffic captured from the front view of the ego vehicle. Describe the current action of the ego car, and explain the cause of this car's action.

RoboTron-Drive

The ego car is currently **stationary** because the **traffic light is red**.

Specialist

CODA-LM: The ego car is halted at a **red traffic light**, which is clearly visible in the upper left corner of the image. The red light indicates that vehicles must stop and wait until the signal changes to green before proceeding. The ego car is positioned at an intersection, and the presence of other vehicles and a pedestrian in the vicinity suggests that it is a busy area. The ego car must remain stationary to comply with traffic regulations and to ensure the safety of pedestrians and other road users.

MAPLM: There are **3 lanes** in the image, lane attributes from left to right are: motorway|motorway|bicycle lane.

DriveLM: The action is **to remain stationary**, the cause is to follow the traffic rules.

LingoQA: The current action of the ego car is **accelerating** because the traffic light is green and the zebra crossing is clear.

Omnidrive: The vehicle is currently stopped at an intersection, waiting for the traffic light to turn green. This is because the **traffic light is red**, indicating that vehicles must halt until it changes to green.

NuInstruct: the ego no turn and **keep speed**

Figure S15. Visualization of BDD-X.

BDD-X



User

<image>. There is a video of traffic captured from the front view of the ego vehicle. Describe the current action of the ego car, and explain the cause of this car's action.

RoboTron-Drive

The ego car is **accelerating** because there is **a gap in the traffic behind the car ahead**.

Specialist

CODA-LM: The ego car is positioned at a **red traffic light**, which requires it to stop and wait for the signal to change to green before it can proceed.

MAPLM: There are **3 lanes in the image**, lane attributes from left to right are: motorway|motorway|bicycle lane.

DriveLM: The current action of the ego car is to **keep going at the same speed**. The cause of this car's action is to follow the traffic rules.

LingoQA: The current action of the ego car is **following the lane** and maintaining a safe distance from the adjacent lane on the right.

Omnidrive: The vehicle is currently proceeding straight ahead at a moderate speed, as the **traffic light is green** and the road ahead is clear.

Nulnstruct: the ego no turn and **keep speed**

Figure S16. Visualization of BDD-X.

References

- [1] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, pages 11621–11631, 2020. [1](#), [2](#)
- [2] Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. Allava: Harnessing gpt4v-synthesized data for a lite vision-language model. *arXiv preprint arXiv:2402.11684*, 2024. [2](#)
- [3] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022. [1](#)
- [4] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. [2](#)
- [5] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024. [2](#)
- [6] Yanze Li, Wenhua Zhang, Kai Chen, Yanxin Liu, Pengxiang Li, Ruiyuan Gao, Lanqing Hong, Meng Tian, Xinhai Zhao, Zhenguo Li, et al. Automated evaluation of large vision-language models on self-driving corner cases. *arXiv preprint arXiv:2404.10595*, 2024. [1](#)
- [7] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014. [2](#)
- [8] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. [1](#)
- [9] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 36, 2024. [2](#)
- [10] AI Meta. Introducing llama 3.1: Our most capable models to date. *Meta AI Blog*, 2024. [1](#)
- [11] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *ICCV*, pages 8430–8439, 2019. [2](#)
- [12] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018. [2](#)
- [13] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, pages 11975–11986, 2023. [1](#)
- [14] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024. [2](#)