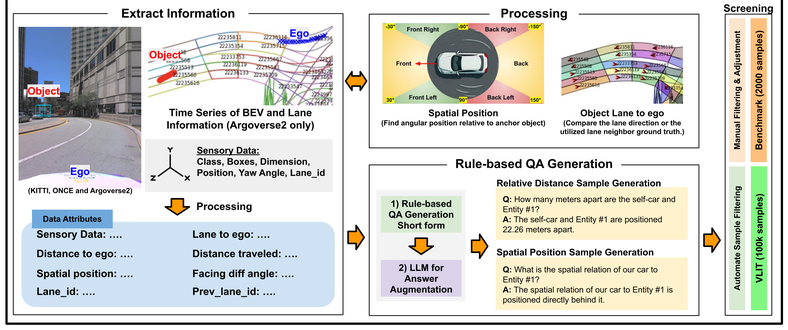
TB-Bench: Training and Testing Multi-Modal AI for Understanding Spatio-Temporal Traffic Behaviors from Dashcam Images/Videos

Korawat Charoenpitaks[1], Van-Quang Nguyen[2], Masanori Suganuma[1], Kentaro Arai[3], Seiji Totsuka[3], Hiroshi Ino[3], Takayuki Okatani[1,2]
[1]Tohoku University [2]RIKEN AIP [3]DENSO Corporation

Motivation & Tasks

- MLLMs in the autonomous driving domain struggle with **spatiotemporal understanding,** lacking traffic-specific data and dedicated benchmarks.
- Existing benchmarks focus on **static spatial tasks** (i.e., object detection) or single-image evaluations, and thus fail to capture **temporal aspects** of traffic behavior.
- **Contributions**:
  - **Closing the dataset gap:** Create eight spatiotemporal tasks covering critical traffic behaviors* and provide two training sets (TB-100k and TB-250k) and the TB-Bench benchmark.
  - **Findings:** Off-the-shelf proprietary models exhibit poor spatiotemporal performance; fine-tuning substantially improves results.
  - **Cross-dataset transfer:** Co-training with our training set boosts other performance.

Data Generation Pipeline



SCAN ME

Do MLLMs Have Spatio-Temporal Capabilities?

Relative Distance (RD)

Question: How many meters away is Entity #2 from Entity #1?

Answer: Entity #2 is situated **15.53 meters** away from Entity #1.

Spatial Reasoning (SR)

Question: What is the spatial position of Entity #1 relative to Entity #2?

Answer: Entity #1 is positioned at the **back right** relative to Entity #2.

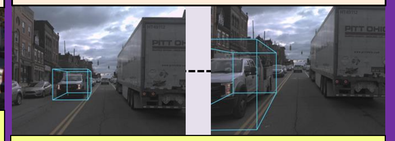
Orientation Reasoning (SR)

Question: What is the orientation of Entity #2 relative to Entity #1, similar, opposite or perpendicular?

Answer: Entity #2 is oriented in a direction that is **diametrically opposed** to the orientation of Entity #1.

Lane relative to ego-car (EGO-LANE)

Question: How would you describe the lane position of Entity #1? Options: front lane, front left lane, front right lane, or oncoming traffic lane.



Answer: Entity #1 is positioned in the lane designated for **oncoming traffic**.

Other lane change (OBJ-LANE)

Question: How would you describe the driving scene involving Entity #1? Please explain, focusing on the vehicle's lane change maneuver.

Answer: Entity #1 smoothly performs a **right lane change**.

Other object turning (OBJ-TURN)

Question: How would you describe the driving scene involving Entity #1? Please explain, focusing on the vehicle's turning maneuver.

Answer: Entity #1 executes a **right-turn** maneuver, smoothly navigating the vehicle in a clockwise direction.

Ego-vehicle turning (EGO-TURN)

Question: How would you describe the driving scene involving our car? Please explain, focusing on our car's turning maneuver.

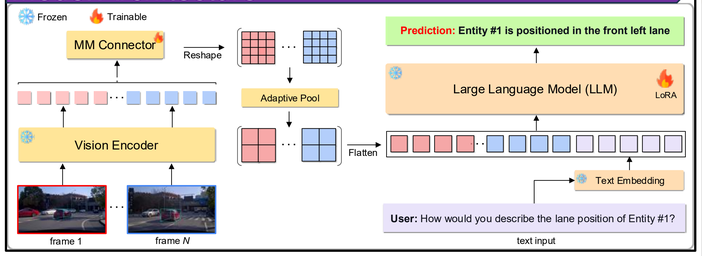
Answer: Our car executed a precise **right-turn** maneuver, smoothly navigating the corner with adept control.

Ego-vehicle traverse distance (EGO-TRA)

Question: How far has our car driven and what kind of steering maneuver did it perform in the current scene?

Answer: The car has driven **11.97 meters** with a straight steering maneuver.

Model Architecture



*Based on Pre-crash Scenarios typology from the National Automotive Sampling System (NASS)

Data Statistics

Statistics of TB-Bench, TB-100k, and TB-250K.
Source datasets: K (KITTI), O (ONCE), Arv2 (Argoverse2).

Task Type	Sources/ Frames	TB-Bench	TB-250k	TB-100k
Spatial Information:				
Relative Distance	[K,O]/ 1	250	35k	10k
Spatial Reasoning	[K,O]/ 1	250	70k	30k
Orientation Reasoning	[K,O]/ 1	250	70k	30k
Object Behavior:				
Other Lane to Ego	[Arv2]/ 8	250	50k	20k
Other Lane Changing	[Arv2]/ 8	250	1.5k	1.5k
Other Turning	[Arv2]/ 8	250	1.5k	1.5k
Ego Behavior:				
Ego Turning	[Arv2]/ 8	250	1.5k	1.5k
Ego Traverse Distance	[Arv2]/ 8	250	25k	15.5k
Total		2000	254k	110k

Results on TB-Bench

Model	RD	SR	OR	EGO-LANE	OBJ-LANE	OBJ-TURN	EGO-TURN	EGO-TRA	Avg
Zero-shot: Bunny-v1.1-4B	24.4	20.4	19.6	28.4	16.0	20.0	34.4	0.0	**20.4**
Zero-shot: Mini-InternVL2-1B-DriveLM	0.0	31.2	20.0	28.4	24.8	47.2	41.6	0.0	**24.2**
Zero-shot: GPT-4o-2024-08-06	8.4	32.0	40.8	54.4	39.6	43.2	40.4	16.0	**34.4**
Fine-tuned: Ours with TB-100k	80.4	74.8	88.8	93.6	65.2	76.4	80.0	60.4	**77.5**
Fine-tuned: Ours with TB-250k	91.2	83.2	94.8	99.6	69.6	80.4	82.8	78.8	**85.1**

More than 50%

Cross-dataset results

Metric	Standard training (only BDD-X)	With co-training (Sampling: BDD-X=20, TB-100k=1)
Speed RMSE↓	1.40	1.38
Speed $A_{0.1}$↑	26.1	26.3
Speed $A_{0.5}$↑	55.7	57.6
Turning RMSE↓	11.2	11.3
Turning $A_{0.1}$↑	44.2	44.5
Turning $A_{0.5}$↑	62.2	63.7