NAVSIM

Non-Reactive **A**utonomous **V**ehicle **Sim**ulation and Benchmarking

Kashyap Chitta

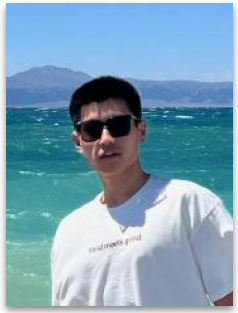
Tübingen AI Center

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Team

Daniel Dauner

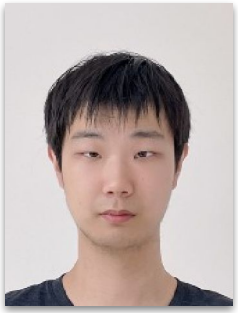
Marcel Hallgarten



Tianyu Li

Xinshuo Weng

Zhiyu Huang



Zetong Yang

Hongyang Li

Igor Gilitschenski

Boris Ivanovic

Marco Pavone

Andreas Geiger

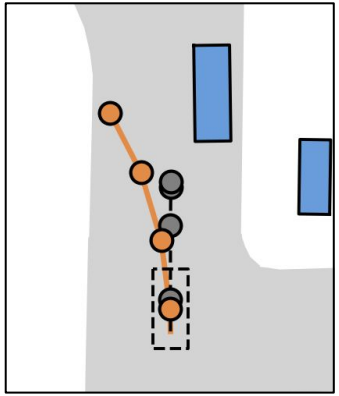
Kashyap Chitta

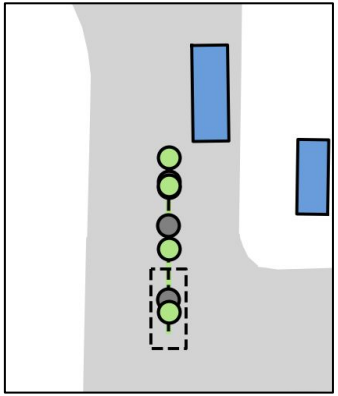
Benchmarking AVs is hard.

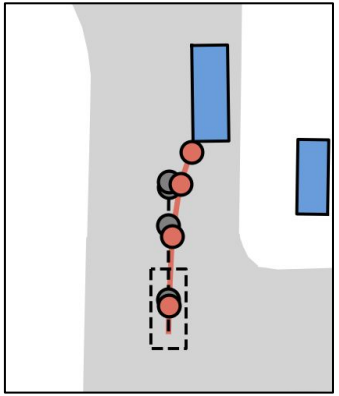
Have we made any real progress in the last year?

Which trajectory is best?

Which trajectory is best?



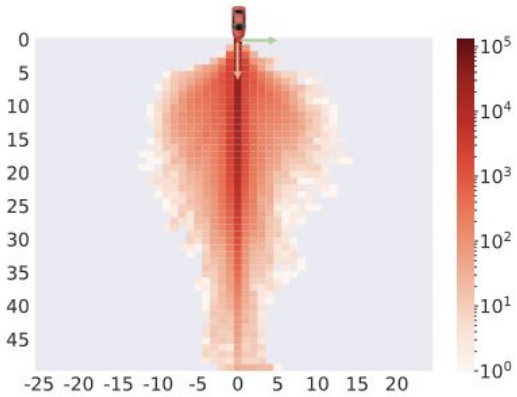




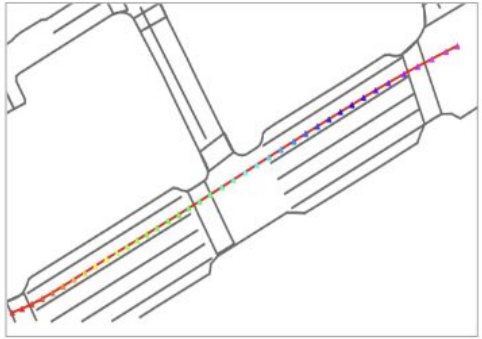
Avg. Displacement Error	2.24	1.05	**0.98**

Is Ego Status All You Need for Open-Loop End-to-End Autonomous Driving?

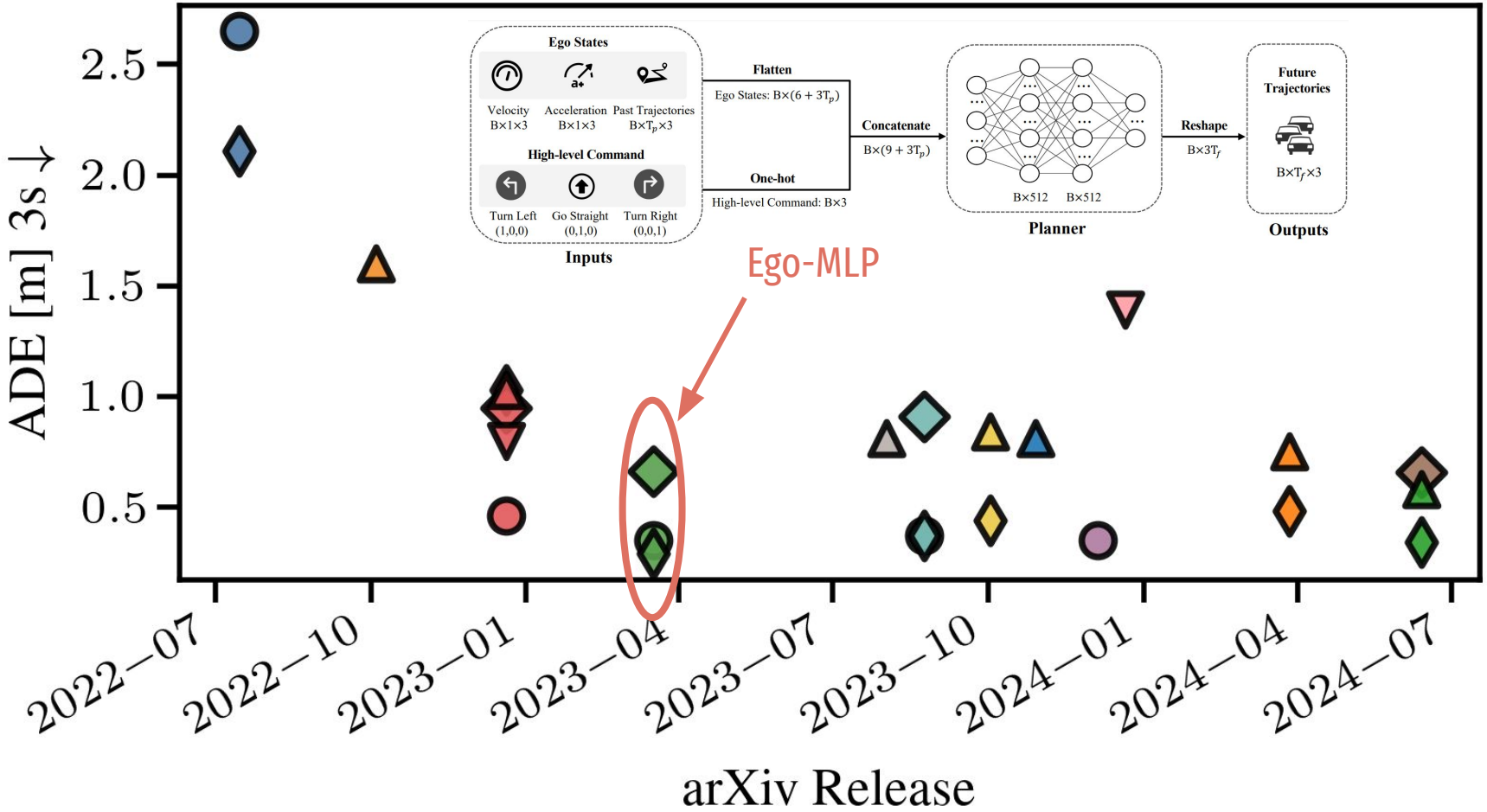
Zhiqi Li[1,2*], Zhiding Yu[2], Shiyi Lan[2], Jiahan Li[1], Jan Kautz[2], Tong Lu[1], Jose M. Alvarez[2]
[1]Nanjing University [2]NVIDIA



(a) Trajectory Heatmap



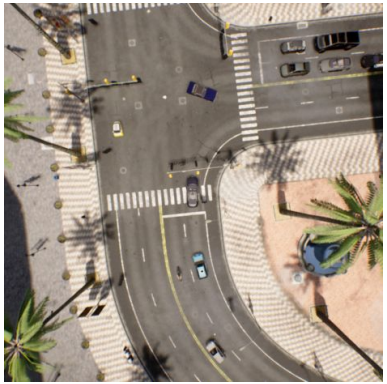
(b) Typical Scene of nuScenes


ADE [m] 3s ↓
2.5
2.0
1.5
1.0
0.5
2022−07
2022−10
2023−01
2023−04
2023−07
2023−10
2024−01
2024−04
2024−07
arXiv Release
Ego-MLP
Ego States
Velocity B×1×3
Acceleration B×1×3
Past Trajectories B×T_p×3
High-level Command
Turn Left (1,0,0)
Go Straight (0,1,0)
Turn Right (0,0,1)
Inputs
Flatten
Ego States: B×(6 + 3T_p)
One-hot
High-level Command: B×3
Concatenate
B×(9 + 3T_p)
B×512
B×512
Planner
Reshape
B×3T_f
Future Trajectories
B×T_f×3
Outputs

What about simulation?

Limited open-source options, e.g. CARLA

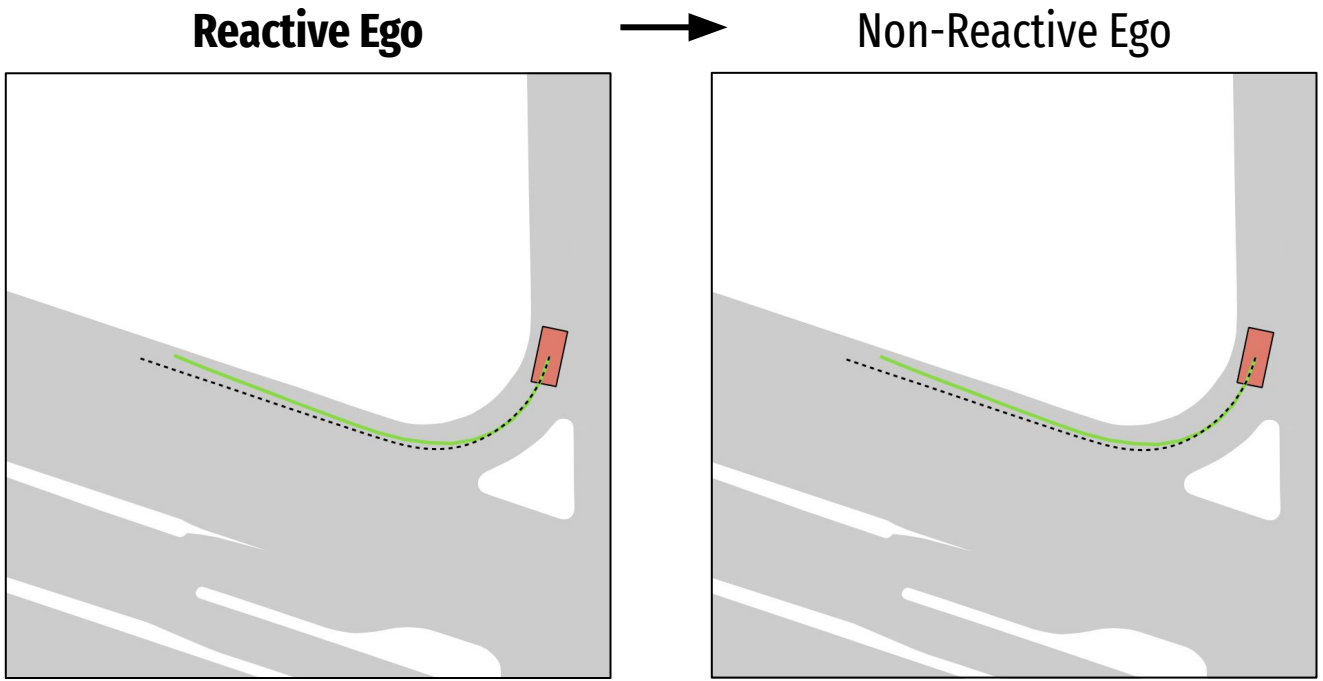
- Domain gaps
- Compute-hungry
- High variance in results



Non-reactive simulation

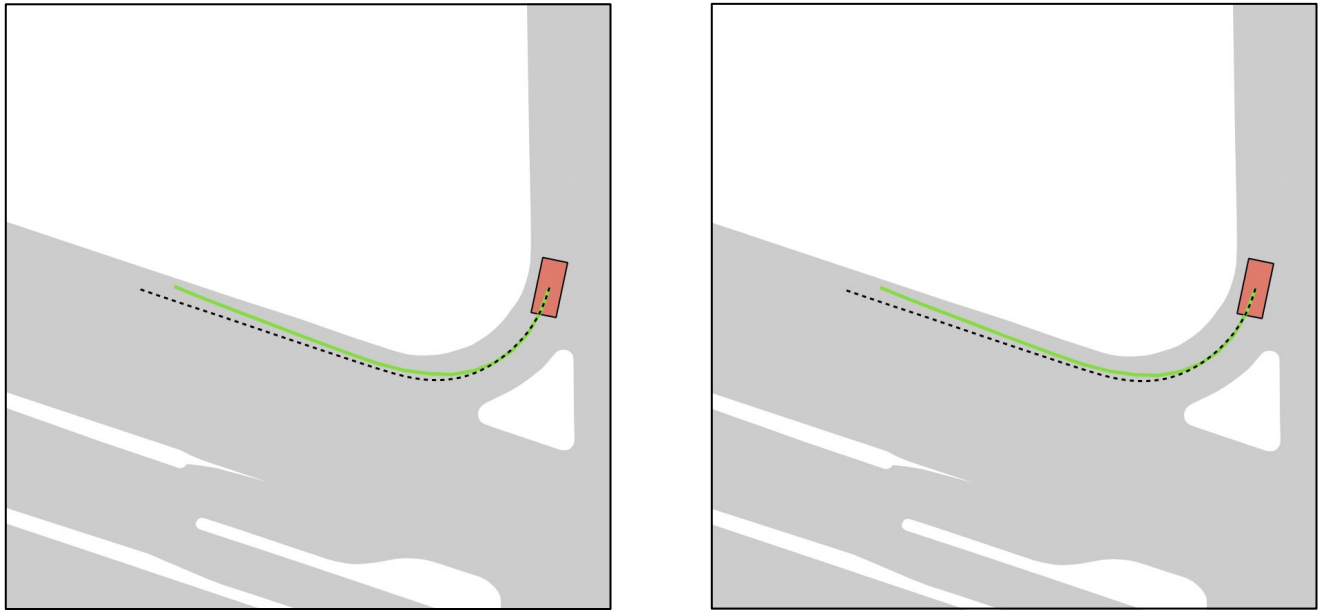
Bypassing the challenges of simulation

Non-reactive ego-vehicle: no sensor simulation



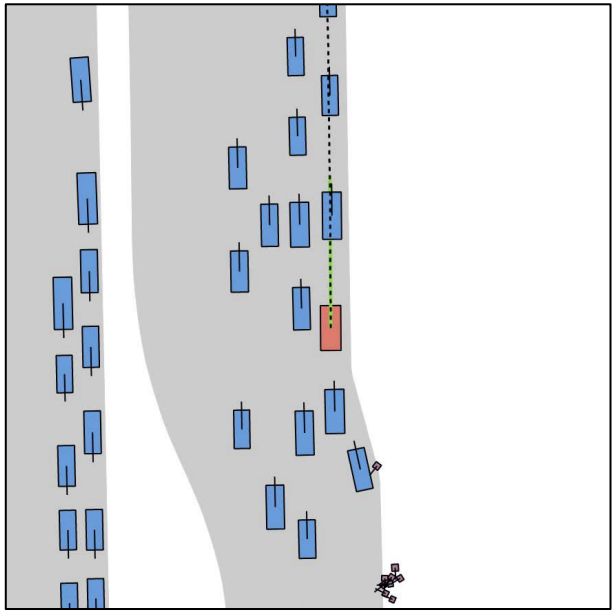
Non-reactive ego-vehicle: no sensor simulation

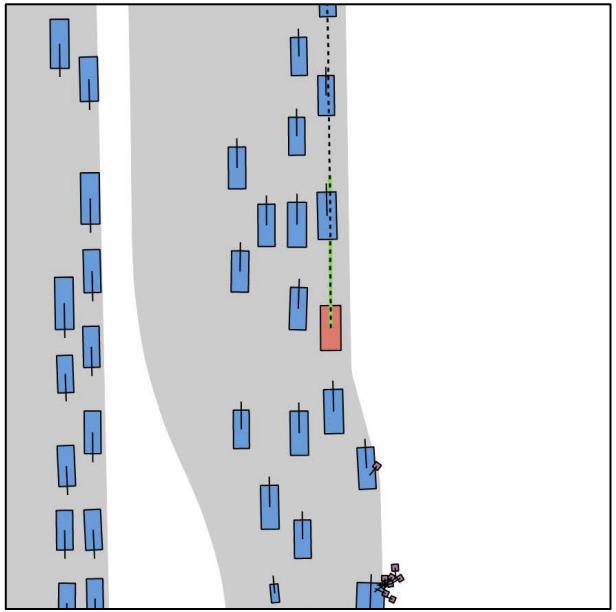
Reactive Ego → **Non-Reactive Ego**



Non-reactive background: no traffic simulation

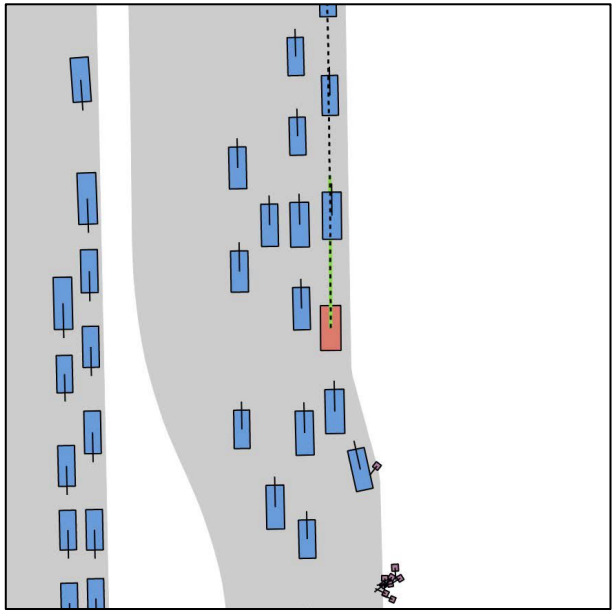
Reactive Background → Non-Reactive Background

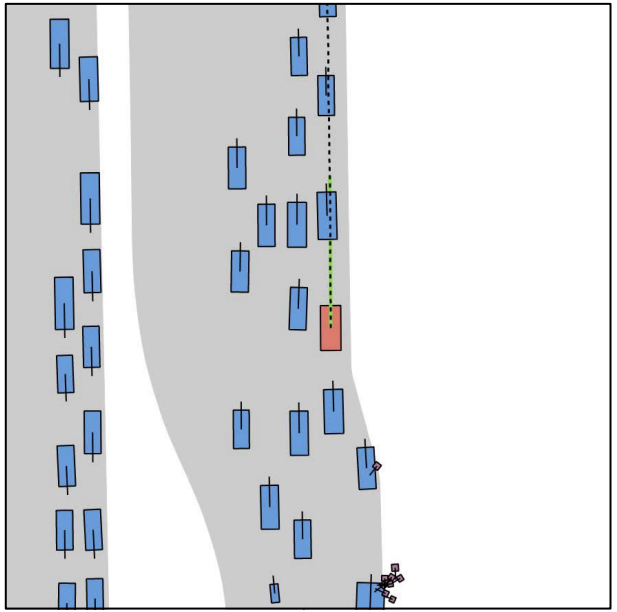




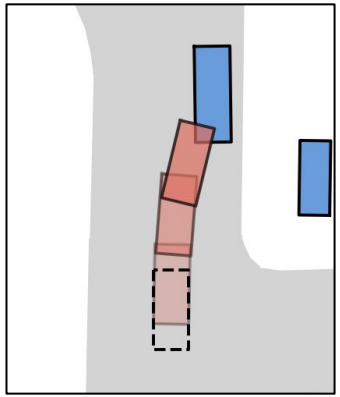
Non-reactive background: no traffic simulation

Reactive Background → **Non-Reactive Background**

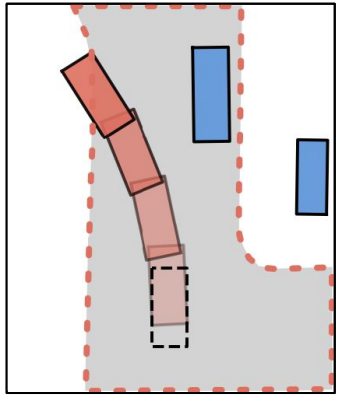




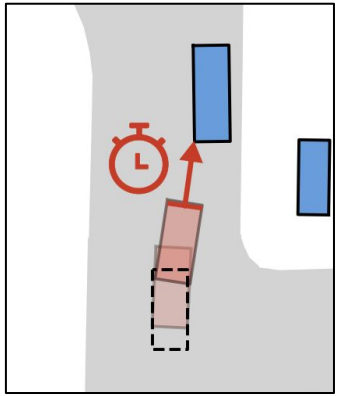
No Collision



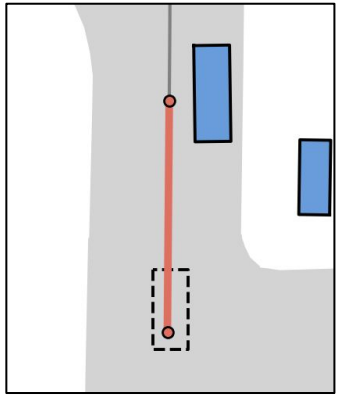
Drivable Area Compliance



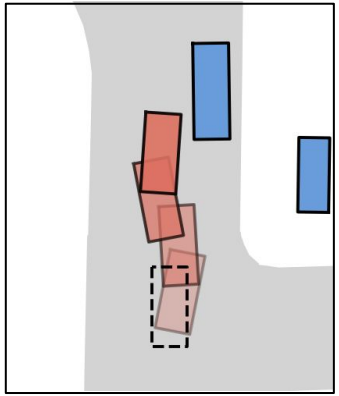
Time to Collision



Ego Progress

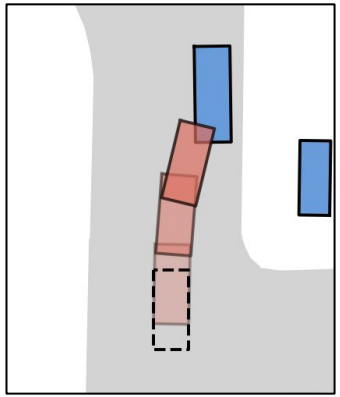


Comfort

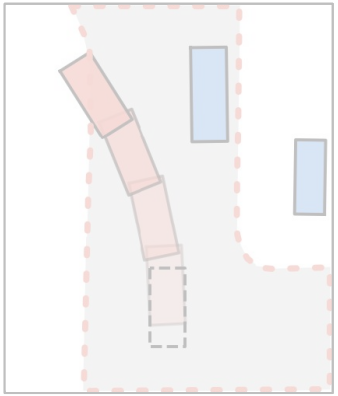


NAVSIM includes five **simulation-based** metrics.

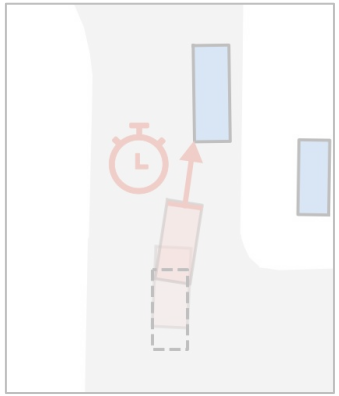
No Collision



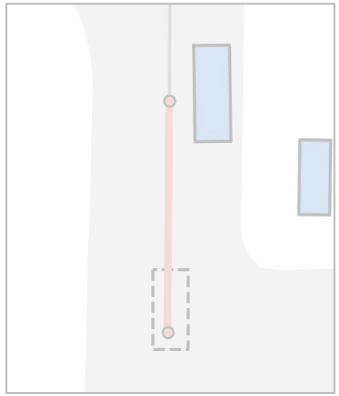
Drivable Area Compliance



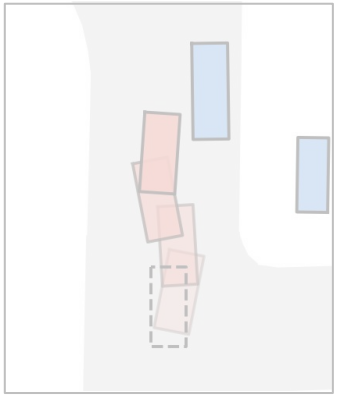
Time to Collision



Ego Progress

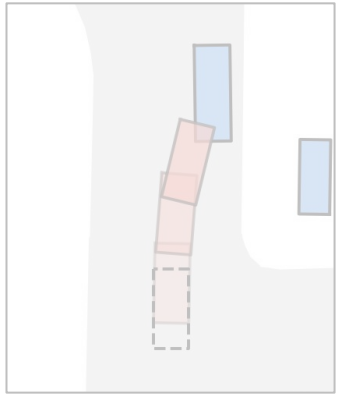


Comfort

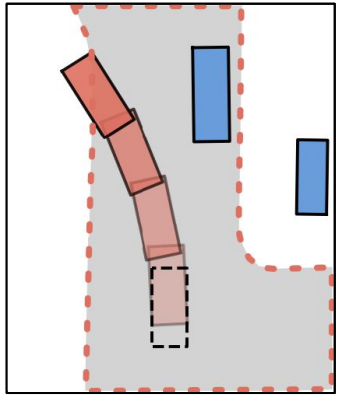


No Collision (NC) for bounding box intersections that are **not "at fault"**.

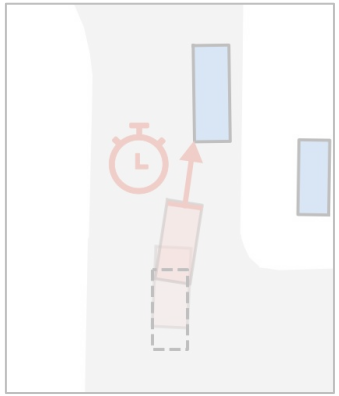
No Collision



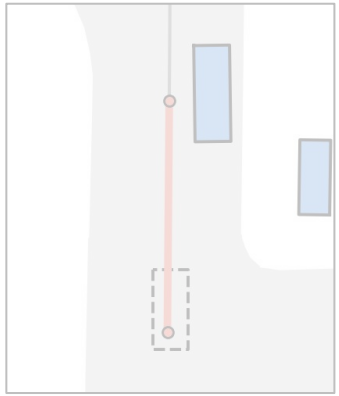
Drivable Area Compliance



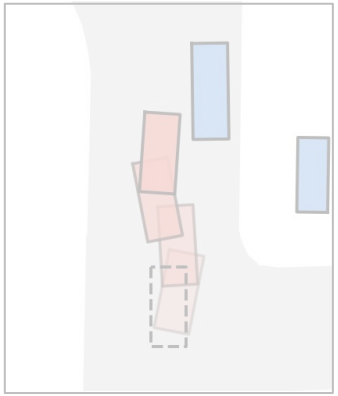
Time to Collision



Ego Progress

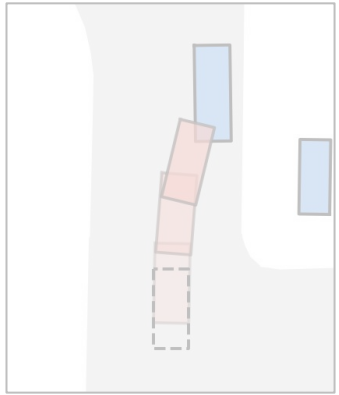


Comfort

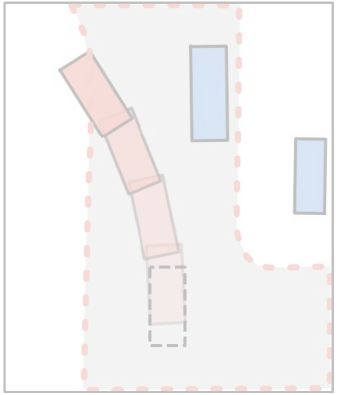


Drivable Area Compliance (DAC) for **staying within** lanes, intersections, parking areas.

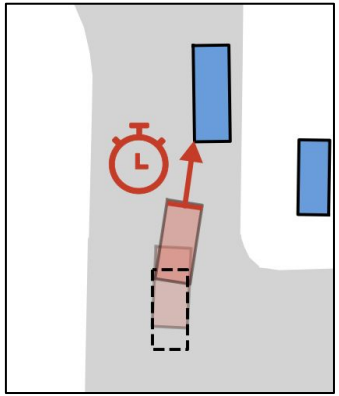
No Collision



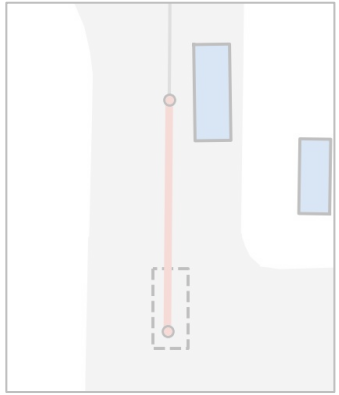
Drivable Area Compliance



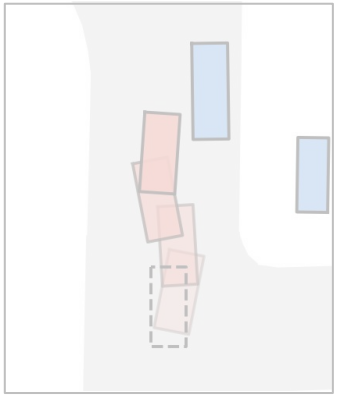
Time to Collision



Ego Progress

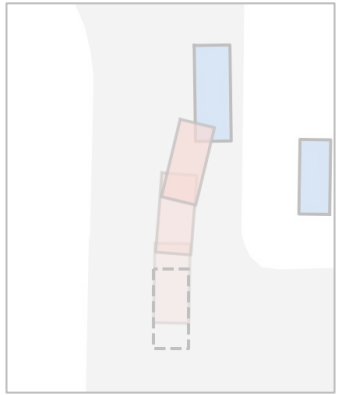


Comfort

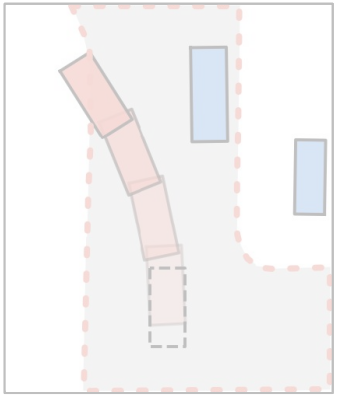


Time-to-Collision (TTC) penalizing **near-collisions** within one second.

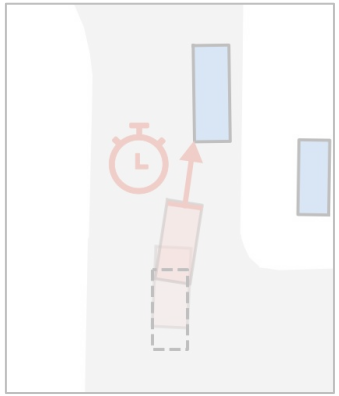
No Collision



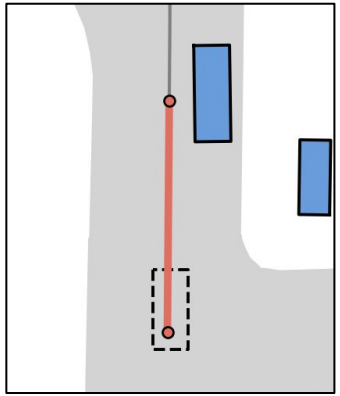
Drivable Area Compliance



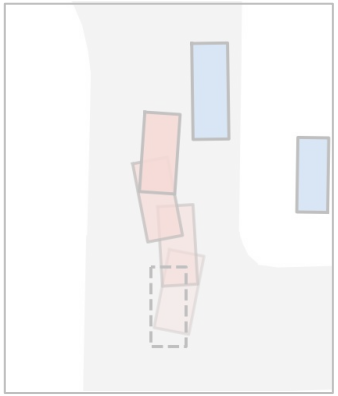
Time to Collision



Ego Progress

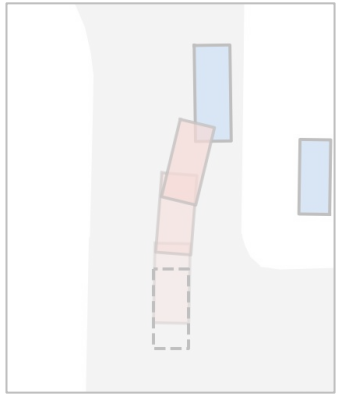


Comfort

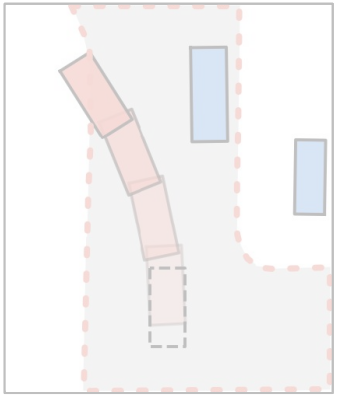


Ego Progress (EP) **relative to a privileged MPC planner**.

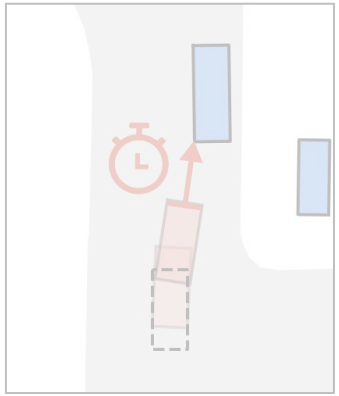
No Collision



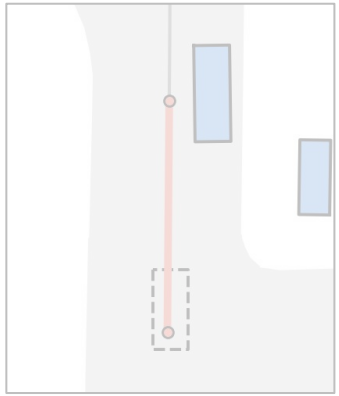
Drivable Area Compliance



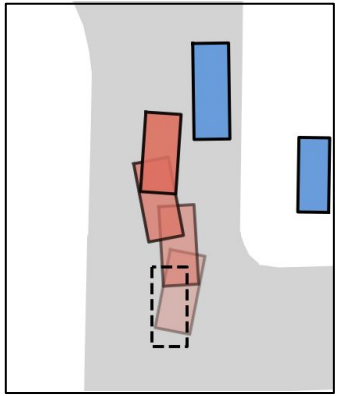
Time to Collision



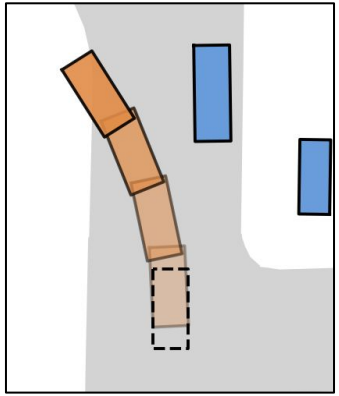
Ego Progress

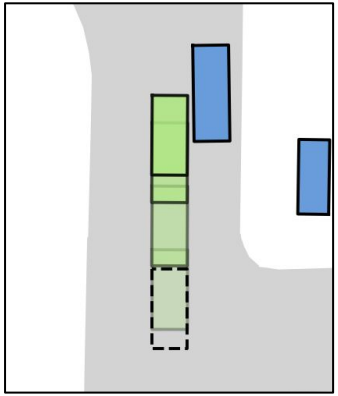


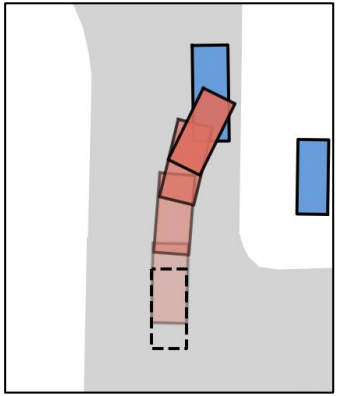
Comfort



Comfort (C) inspecting that **acceleration and jerk** are within human-like thresholds.







1. No at-fault Collision	1.0	1.0	0.0
2. Drivable Area Compliance	0.0	1.0	1.0
3. Time to Collision	1.0	1.0	0.0
4. Ego Progress	1.0	0.93	0.97
5. Comfort	0.0	1.0	1.0

The Predictive Driver Model (PDM) Score

Penalties

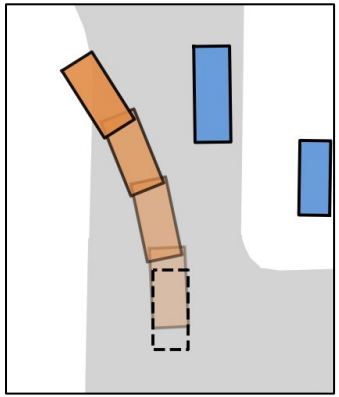
Trade-offs

$$\text{PDMS} = \left(\prod_{m \in \{\text{NC},\text{DAC}\}} \texttt{score}_m \right) \times \left(\frac{\sum_{w \in \{\text{EP},\text{TTC},\text{C}\}} \texttt{weight}_w \times \texttt{score}_w}{\sum_{w \in \{\text{EP},\text{TTC},\text{C}\}} \texttt{weight}_w} \right)$$

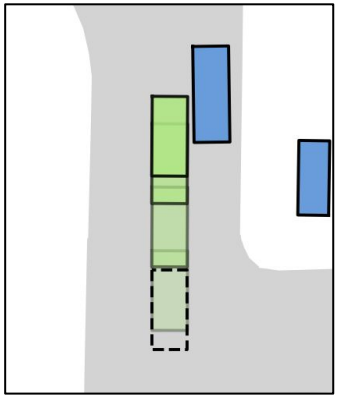
$$= \left(\texttt{score}_{\text{NC}} \times \texttt{score}_{\text{DAC}} \right) \times \left(\frac{5 \times \texttt{score}_{\text{EP}} + 5 \times \texttt{score}_{\text{TTC}} + 2 \times \texttt{score}_{\text{C}}}{12} \right)$$

The Predictive Driver Model (PDM) Score

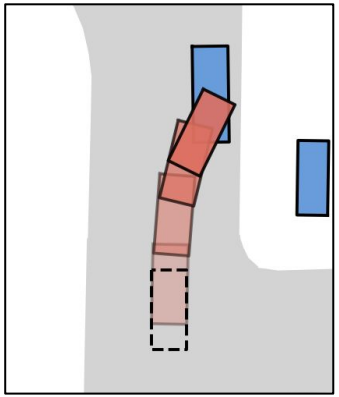
PDM Score (4s)



0.0



0.97

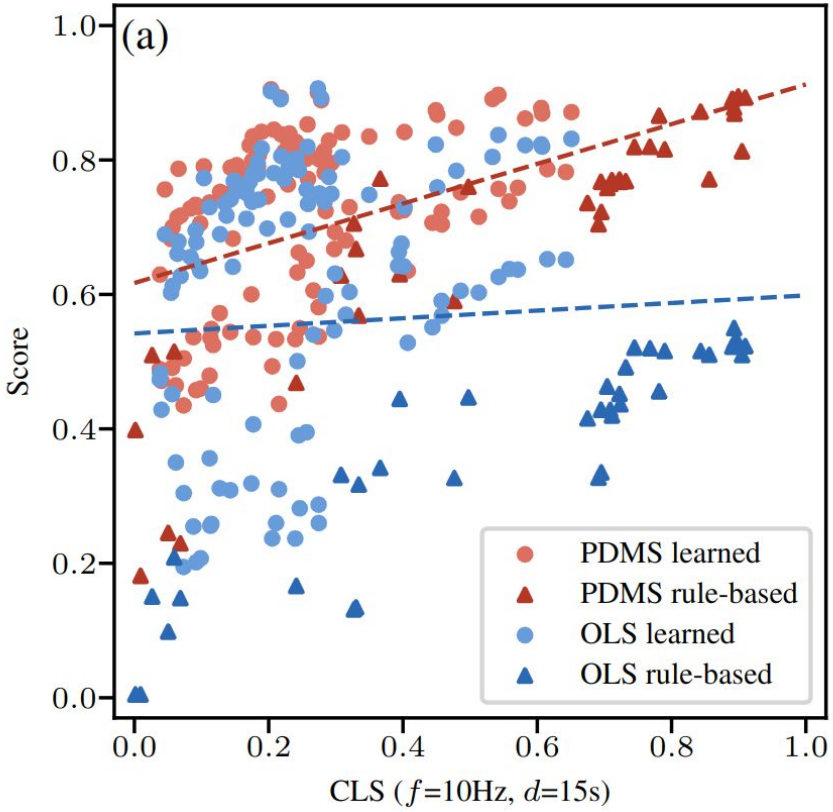


0.0

Does it work?

Benchmarking 150+ planners using their
CLS (Closed-Loop Score)

- Simulation @ 10Hz
- 15 second horizon
- OLS: prior open-loop metric
- Both used in 2023 nuPlan Challenge
- **PDMS and CLS much better correlated**



Entry bottlenecks

Making E2E driving research more accessible

Storage bottleneck of large-scale benchmarking

Storage requirements seldom feasible, e.g. nuPlan

10Hz

20+ TB

Storage bottleneck of large-scale benchmarking

Storage requirements seldom feasible, e.g. nuPlan

OpenScene:

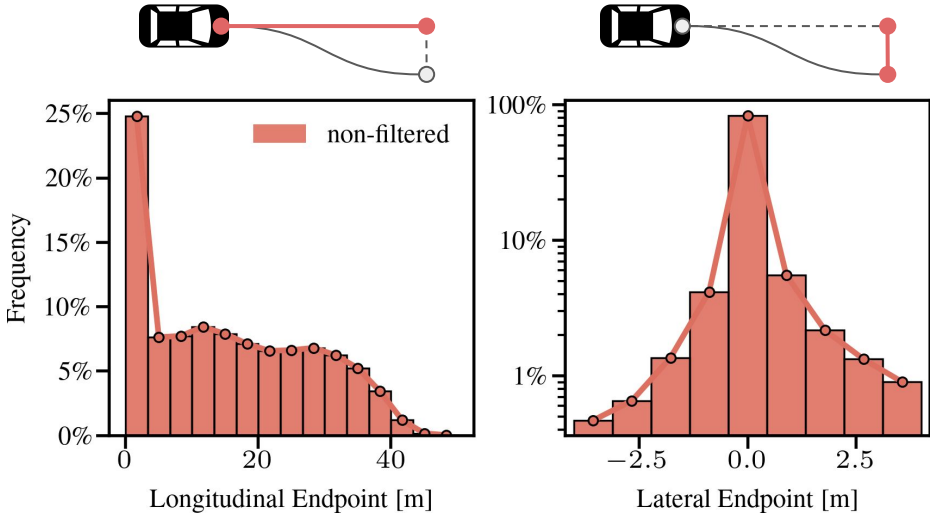
- Redistribution with 2Hz (< 3TB)
- Standardized train (100k) & test (12k) splits
- Private data for evaluation server

10Hz → 2Hz

20+ TB → 3 TB

Improving the test distribution

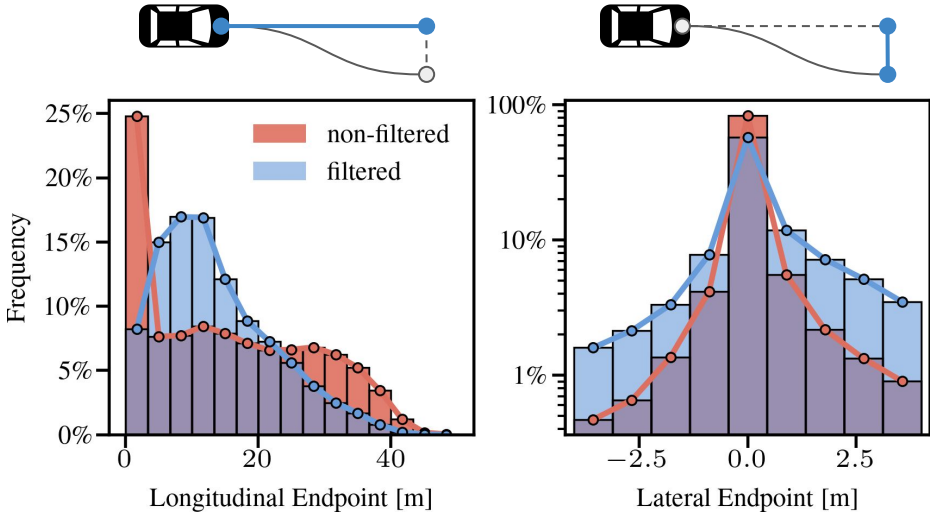
	Agent	NC	DAC	PDMS
unfiltered	Straight	93	90	79
unfiltered	Human	99	97	91
filtered	Straight	69	59	22
filtered	Human	100	100	95



Unfiltered recordings are mostly static or straight driving scenes.

Improving the test distribution

	Agent	NC	DAC	PDMS
unfiltered	Straight	93	90	79
unfiltered	Human	99	97	91
filtered	Straight	69	59	22
filtered	Human	100	100	95



Filtered data results in more diverse and challenging scenes.

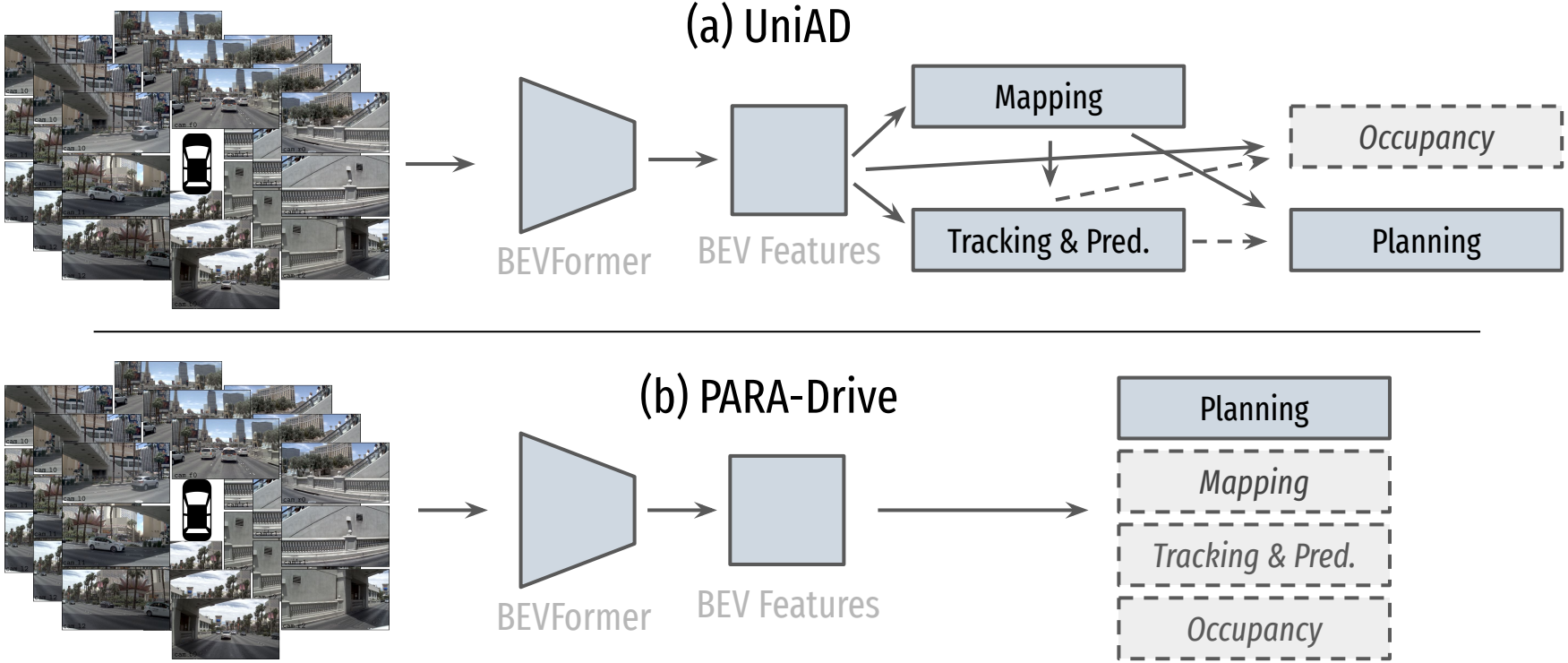
Agent Interface in NAVSIM

Task: predict 4-second trajectory

- 8 x surround-view cameras
- 5 x merged LiDAR
- Ego velocity & acceleration
- Navigation goal

1.5s history

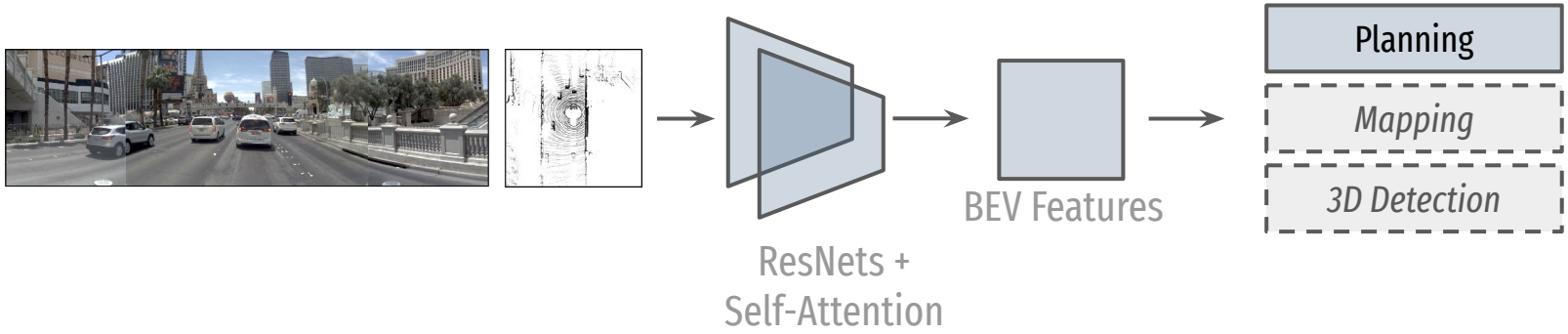
Baselines taken from nuScenes



Training budget: 5000 GPU hours

Baselines taken from CARLA

(c) TransFuser



Training budget: 24 GPU hours

Current state of the field

What does the new benchmark show us?

Benchmarking on filtered test scenarios

Method	**NC↑**	**DAC↑**	**TTC↑**	**Comf↑**	**EP↑**	**PDMS↑**
Ego-MLP	93	77	84	**100**	63	66
(a) UniAD	**98**	92	**93**	**100**	**79**	83
(b) PARA-Drive	**98**	92	**93**	**100**	**79**	**84**

Clear gap between sensor agents and “blind” Ego-MLP

Benchmarking on filtered test scenarios

Method	NC↑	DAC↑	TTC↑	Comf↑	EP↑	PDMS↑
Ego-MLP	93	77	84	**100**	63	66
(a) UniAD	**98**	92	**93**	**100**	**79**	83
(b) PARA-Drive	**98**	92	**93**	**100**	**79**	**84**
(c) TransFuser	**98**	**93**	**93**	**100**	**79**	**84**

TransFuser on par with nuScenes baselines, despite less compute (1 vs. 80 GPUs)

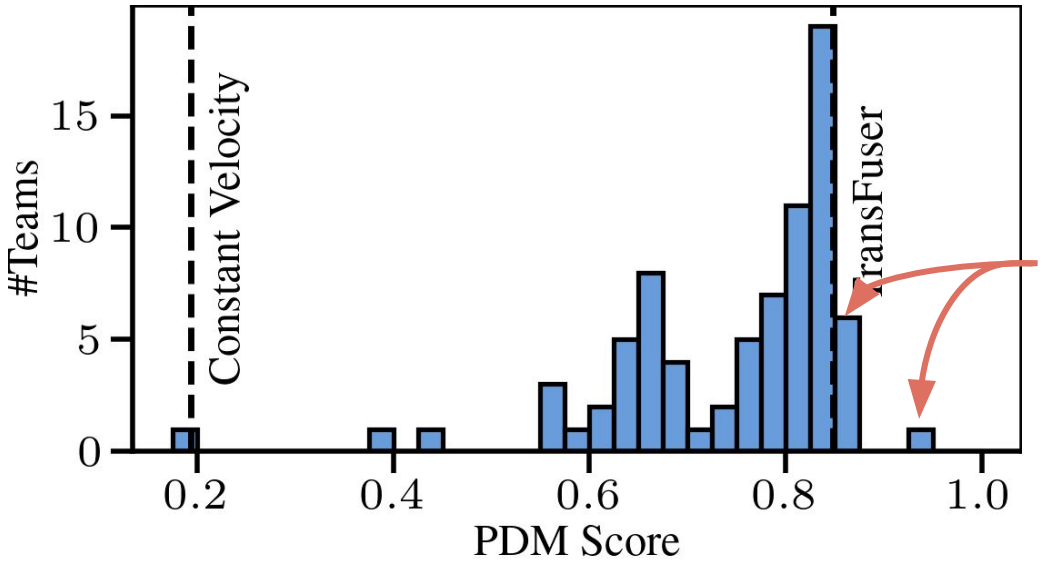
Benchmarking on filtered test scenarios

Method	**NC↑**	**DAC↑**	**TTC↑**	**Comf↑**	**EP↑**	**PDMS↑**
Ego-MLP	93	77	84	**100**	63	66
(a) UniAD	**98**	92	**93**	**100**	**79**	83
(b) PARA-Drive	**98**	92	**93**	**100**	**79**	**84**
(c) TransFuser	**98**	**93**	**93**	**100**	**79**	**84**
Human	*100*	*100*	*100*	*99.9*	*87.5*	*95*

Human Trajectories 11% better than all sensor agents.

2024 NAVSIM Challenge

463 submissions, 78 on leaderboard 🤗



Rank	🗺️	Institution	PDM Score (primary) ▼	Team Name
1	US	NVIDIA	**0.9274**	Team NVIDIA
2	CN	ZERON 零一汽车	**0.8747**	ZERON

Limitations

We still recommend complementing NAVSIM with CARLA:

- Longer evaluations (~10km, several minutes of driving)
- Considers more infractions (rear-end collisions, running red lights)
- However, simulation much more compute intensive

Next steps

Devkit available, paper out soon!

- Better metrics
- More metrics
- New datasets
- More challenges!

https://github.com/autonomousvision/navsim