Planning-oriented Autonomous Driving

Yihan Hu*  Jiazhi Yang*  Li Chen**  Keyu Li*
Chonghao Sima  Xizhou Zhu  Siqi Chai  Senyao Du  Tianwei Lin  Wenhai Wang
Lewei Lu  Xiaosong Jia  Qiang Liu  Jifeng Dai  Yu Qiao  Hongyang Li+
*equal contribution  +project lead

Poster: THU-AM-131
Planning-oriented Autonomous Driving

Background and Motivation
Background - Autonomous Driving (AD) Systems

Perception → Prediction → Planning

Bounding boxes → Waypoints → Trajectory

What are around? How will they go in the future? Where should I go?

Various weathers, illuminations, and scenarios
Background - Design Options for Autonomous Driving (AD) Systems

(a) Standalone Models

- Typical **Industry** solutions
- Independent teams for module developments
- Severe error accumulation
Background - Design Options for Autonomous Driving (AD) Systems

(b) Multi-task Framework

- **Shared feature** for multiple tasks
- Easily extended to more tasks, Compute-efficient
- Lack of tasks’ coordination

![Multi-task Framework Diagram]

- **Task 1**
- **Task 2**
- **Task N**

- **Object Detection**
- **Traffic Lights Detection**
- **Lane Prediction**

---

*credit to Tesla AI Day 2021*
(c.1) End-to-end Framework - Vanilla Solutions

- **Direct policy learning** from sensor inputs, bypassing intermediate tasks
  - ✔️ Simple design with good performance in the simulator
  - ❌ Deficient in interpretability
(c.2) End-to-end Framework - Explicit / Interpretable Design

- Introducing **intermediate tasks** to assist planning
- Better interpretability (e.g. Bird’s-eye-view, BEV)
- Lack some crucial components¹

1. The necessities of each component is mentioned in Appendix.
Motivation - Towards Reliable Planning

Ours: **Planning-oriented** Autonomous Driving

What do we want:
- Unify full-stack AD tasks
- Coordinate all task towards safe planning
UniAD - Overview

Which tasks?

How to construct?

How to train?
Planning-oriented Autonomous Driving

Delving into Details
UniAD - Which Tasks?

- Object Detection
- Object Tracking
- Online Mapping
  - ...

Incorporate all tasks in a \textit{hierarchical} manner

- Motion Forecasting
- Occupancy Prediction
  - ...

- Track agents behavior
- Guide with map (lane)

- Interact with environment
- Find free space

- Five \textit{safety-critical tasks}: Model the static and dynamic information
- Task hierarchy: Tasks are \textit{well-organized} to optimize information flow to the planner
UniAD - How to Construct?

**Pipeline**

- **Ego-vehicle Query**
  - Multi-view Vision-only Input

- **Backbone**
  - Perception

- **Track agents across time**

- **Predict long-term trajectory**

- **Plan final trajectory**
  - **Avoid collision**

- **MapFormer**
  - **TrackFormer**
  - **Motion Former**
    - Agent-level Feature
    - Scene-level Feature

- **OccFormer**
  - **Planner**

- **Segment map elements**

- **Scene-level representation**

- **Prediction**
UniAD - How to Construct?

- Entire pipeline connected by queries
- Tasks coordinated with queries
- Interactions modeled by attention

Unified Query

Transformer-based

First time to unify full-stack AD tasks!
UniAD - How to Construct?

**TrackFormer - MOTR (ECCV 2022)**

- **End-to-end trainable tracking without post-association**

**MapFormer - Panoptic SegFormer (CVPR 2022)**

- Each query represents a map element
UniAD - How to Construct?

MotionFormer (Proposed in UniAD)

- Diverse relation modelings via attentions: Agent-agent, agent-map, agent-goal

Non-linear optimization:
Adjust ground-truth trajectory based on upstream predictions
OccFormer (Proposed in UniAD)

- **Encode agent-wise knowledge** into the scene representation
- **Predict occupancy as attention mask** to restrict the interactions between the agents and their corresponding BEV features.
UniAD - How to Construct?

**Perception**

**Prediction**

**Planning**

**Planner** (Proposed in UniAD)

- **Ego-vehicle query**: consistently models the ego-vehicle
- **Collision optimization**: Steer the predicted trajectories clear of predicted occupancy.
The Recipe - How to Train?

**Two-phase training.** Perception stage + End-to-end stage
- The stabilized perception capability helps the end-to-end stage converge faster

**Shared matching.** Matching results of tracking reused in motion and occupancy
- Consistent learning of agent identities
- Converging faster
Planning-oriented Autonomous Driving

Experiments
## UniAD - Ablation Results

### Tasks benefit each other and contribute to safe planning

<table>
<thead>
<tr>
<th>ID</th>
<th>Track</th>
<th>Map</th>
<th>Motion</th>
<th>Occ.</th>
<th>Plan</th>
<th>AMOTA↑</th>
<th>AMOTP↓</th>
<th>IDS↓</th>
<th>Tracking</th>
<th>Mapping</th>
<th>Motion Forecasting</th>
<th>Occupancy Prediction</th>
<th>Planning</th>
</tr>
</thead>
<tbody>
<tr>
<td>0°</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>0.356</td>
<td>1.328</td>
<td>893</td>
<td>0.302</td>
<td>0.675</td>
<td>0.858</td>
<td>1.270</td>
<td>0.186</td>
</tr>
<tr>
<td>1</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td>0.348</td>
<td>1.333</td>
<td>791</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td>0.355</td>
<td>1.336</td>
<td>785</td>
<td>0.305</td>
<td>0.674</td>
<td>0.815</td>
<td>1.224</td>
<td>0.182</td>
</tr>
<tr>
<td>3</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td>0.354</td>
<td>1.339</td>
<td>820</td>
<td>0.303</td>
<td>0.672</td>
<td>0.751</td>
<td>1.109</td>
<td>0.162</td>
</tr>
<tr>
<td>4</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td>0.355</td>
<td>1.336</td>
<td>785</td>
<td>0.305</td>
<td>0.674</td>
<td>0.815</td>
<td>1.224</td>
<td>0.182</td>
</tr>
<tr>
<td>5</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td>0.360</td>
<td>1.350</td>
<td>919</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td>0.354</td>
<td>1.339</td>
<td>820</td>
<td>0.303</td>
<td>0.672</td>
<td>0.736(-9.7%)</td>
<td>1.066(-12.9%)</td>
<td>0.158</td>
</tr>
<tr>
<td>7</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td>0.359</td>
<td>1.359</td>
<td>1057</td>
<td>0.304</td>
<td>0.675</td>
<td>0.710(-5.8%)</td>
<td>1.005(-5.8%)</td>
<td>0.146</td>
</tr>
<tr>
<td>8</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td>0.360</td>
<td>1.322</td>
<td>809</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td>0.359</td>
<td>1.359</td>
<td>1057</td>
<td>0.304</td>
<td>0.675</td>
<td>0.710(-5.8%)</td>
<td>1.005(-5.8%)</td>
<td>0.146</td>
</tr>
<tr>
<td>10</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td>0.366</td>
<td>1.337</td>
<td>889</td>
<td>0.303</td>
<td>0.672</td>
<td>0.741</td>
<td>1.077</td>
<td>0.157</td>
</tr>
<tr>
<td>11</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td>0.358</td>
<td>1.334</td>
<td>641</td>
<td>0.302</td>
<td>0.672</td>
<td>0.728</td>
<td>1.054</td>
<td>0.154</td>
</tr>
<tr>
<td>12</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td>0.358</td>
<td>1.334</td>
<td>641</td>
<td>0.302</td>
<td>0.672</td>
<td>0.728</td>
<td>1.054</td>
<td>0.154</td>
</tr>
</tbody>
</table>

### Conclusion:

- **ID. 4-6:** Track & Map $\rightarrow$ Motion
- **ID. 7-9:** Motion $\leftrightarrow$ Occupancy
- **ID. 10-12:** Motion & Occupancy $\rightarrow$ Planning
UniAD - Results

Even outperforms LiDAR-based counterparts on planning

### Planning

<table>
<thead>
<tr>
<th>Method</th>
<th>L2(m) (_\downarrow)</th>
<th>Col. Rate(%) (_\downarrow)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1s</td>
<td>2s</td>
</tr>
<tr>
<td>NMP(^\dagger) [88]</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SA-NMP(^\dagger) [88]</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>FF(^\dagger) [36]</td>
<td>0.55</td>
<td>1.20</td>
</tr>
<tr>
<td>EO(^\dagger) [42]</td>
<td>0.67</td>
<td>1.36</td>
</tr>
<tr>
<td>ST-P3 [37]</td>
<td>1.33</td>
<td>2.11</td>
</tr>
<tr>
<td>UniAD</td>
<td><strong>0.48</strong></td>
<td><strong>0.96</strong></td>
</tr>
</tbody>
</table>

\(^\dagger\): LiDAR-based

Camera-based
# UniAD - Results

## SOTA performance on all investigated tasks

### Multi-object Tracking

<table>
<thead>
<tr>
<th>Method</th>
<th>AMOTA↑</th>
<th>AMOTP↓</th>
<th>Recall↑</th>
<th>IDS↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>Immortal Tracker† [82]</td>
<td>0.378</td>
<td>1.119</td>
<td>0.478</td>
<td>936</td>
</tr>
<tr>
<td>ViP3D [30]</td>
<td>0.217</td>
<td>1.625</td>
<td>0.363</td>
<td>-</td>
</tr>
<tr>
<td>QD3DT [35]</td>
<td>0.242</td>
<td>1.518</td>
<td>0.399</td>
<td>-</td>
</tr>
<tr>
<td>MUTR3D [91]</td>
<td>0.294</td>
<td>1.498</td>
<td>0.427</td>
<td>3822</td>
</tr>
<tr>
<td>UniAD</td>
<td>0.359</td>
<td>1.320</td>
<td>0.467</td>
<td>906</td>
</tr>
</tbody>
</table>

### Motion Forecasting

<table>
<thead>
<tr>
<th>Method</th>
<th>minADE(m)↓</th>
<th>minFDE(m)↓</th>
<th>MR↓</th>
<th>EPA↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>PnPNet† [50]</td>
<td>1.15</td>
<td>1.95</td>
<td>0.226</td>
<td>0.222</td>
</tr>
<tr>
<td>ViP3D [30]</td>
<td>2.05</td>
<td>2.84</td>
<td>0.246</td>
<td>0.226</td>
</tr>
<tr>
<td>Constant Pos.</td>
<td>5.80</td>
<td>10.27</td>
<td>0.347</td>
<td>-</td>
</tr>
<tr>
<td>Constant Vel.</td>
<td>2.13</td>
<td>4.01</td>
<td>0.318</td>
<td>-</td>
</tr>
<tr>
<td>UniAD</td>
<td>0.71</td>
<td>1.02</td>
<td>0.151</td>
<td>0.456</td>
</tr>
</tbody>
</table>

### Mapping

<table>
<thead>
<tr>
<th>Method</th>
<th>Lanes↑</th>
<th>Drivable↑</th>
<th>Divider↑</th>
<th>Crossing↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>VPN [63]</td>
<td>18.0</td>
<td>76.0</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>LSS [66]</td>
<td>18.3</td>
<td>73.9</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>BEFormer [48]</td>
<td>23.9</td>
<td>77.5</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>BEVerse† [92]</td>
<td>-</td>
<td>-</td>
<td>30.6</td>
<td>17.2</td>
</tr>
<tr>
<td>UniAD</td>
<td>31.3</td>
<td>69.1</td>
<td>25.7</td>
<td>13.8</td>
</tr>
</tbody>
</table>

### Occupancy Prediction

<table>
<thead>
<tr>
<th>Method</th>
<th>IoU-n.↑</th>
<th>IoU-f.↑</th>
<th>VPQ-n.↑</th>
<th>VPQ-f.↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>FIERY [34]</td>
<td>59.4</td>
<td>36.7</td>
<td>50.2</td>
<td>29.9</td>
</tr>
<tr>
<td>StretchBEV [1]</td>
<td>55.5</td>
<td>37.1</td>
<td>46.0</td>
<td>29.0</td>
</tr>
<tr>
<td>ST-P3 [37]</td>
<td>-</td>
<td>38.9</td>
<td>-</td>
<td>32.1</td>
</tr>
<tr>
<td>BEVerse† [92]</td>
<td>61.4</td>
<td>40.9</td>
<td>54.3</td>
<td>36.1</td>
</tr>
<tr>
<td>UniAD</td>
<td>63.4</td>
<td>40.2</td>
<td>54.7</td>
<td>33.5</td>
</tr>
</tbody>
</table>
UniAD - Visualizations

Planner attends to crucial areas in complex scenes
UniAD - Recover from Upstream Errors

Planner could still attend to ‘undetected’ regions/objects
One-page Summary

- **Planning-oriented Philosophy**: An end-to-end autonomous driving (AD) framework in pursuit of safe planning, equipped with a wide span of AD tasks.

- **Unified Query design**: Queries as interfaces to connect and coordinate all tasks.

- **State-of-the-art (SOTA) Performance** with vision-only input.

- **First Step towards Autonomous Driving Foundation Models**

![Diagram of autonomous driving framework](image-url)
What’s next? beyond UniAD

Embracing Foundation Models for Autonomous Driving
UniAD v2

Data & Training Strategy
- Multiple datasets with labels for various tasks?

Shippable Algorithm
- More modules integration, extensible to applications (e.g. V2X)

Closed-loop System
- Closed-loop training and testing in simulator & real world

Check out the latest Survey Paper!
https://github.com/OpenDriveLab/End-to-end-Autonomous-Driving
Beyond UniAD: DriveAGI

Data-centric Pipeline

**Data Collection**

- Motional
- Lyf
- YouTube

**Data Generation**

*Partial photo by courtesy of online resources.*

Pre-training DriveCore

Universal Foundation Model for autonomous driving

How to formulate?
What's the objective goal?

Applications

**Autonomous Driving**

**Broader Impact**

*Partial photo by courtesy of online resources.*