NYU

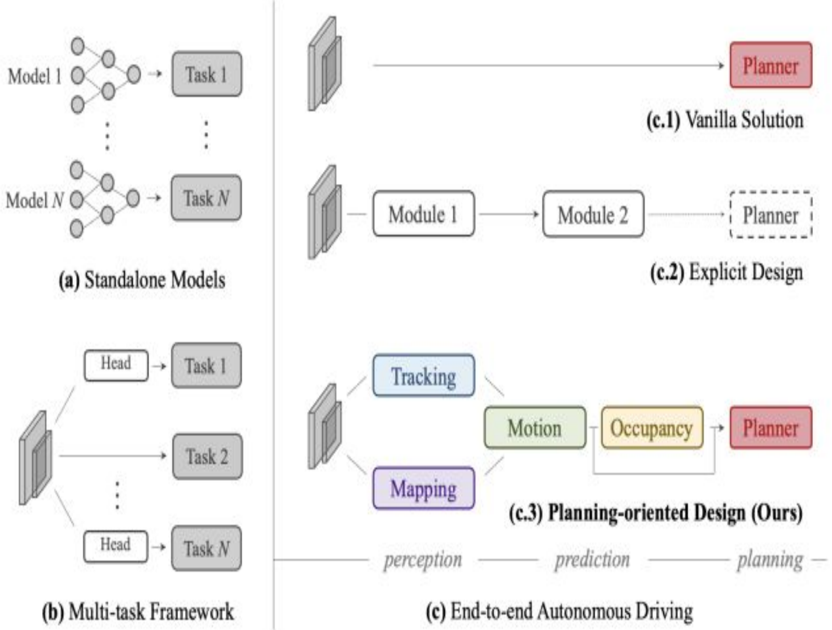
Planning-oriented Autonomous Driving

Presented by Raman Jha

04/3/2025

Introduction

- **UniAD** is a unified framework for autonomous driving that **integrates perception, prediction, and planning** tasks into a single end-to-end system.
- Unlike traditional modular approaches, **UniAD adopts a planning-oriented philosophy**, ensuring that all preceding tasks contribute directly to safe and efficient driving decisions.
- The framework uses **query-based interfaces to connect modules**, enabling flexible feature sharing and robust task coordination



NYU

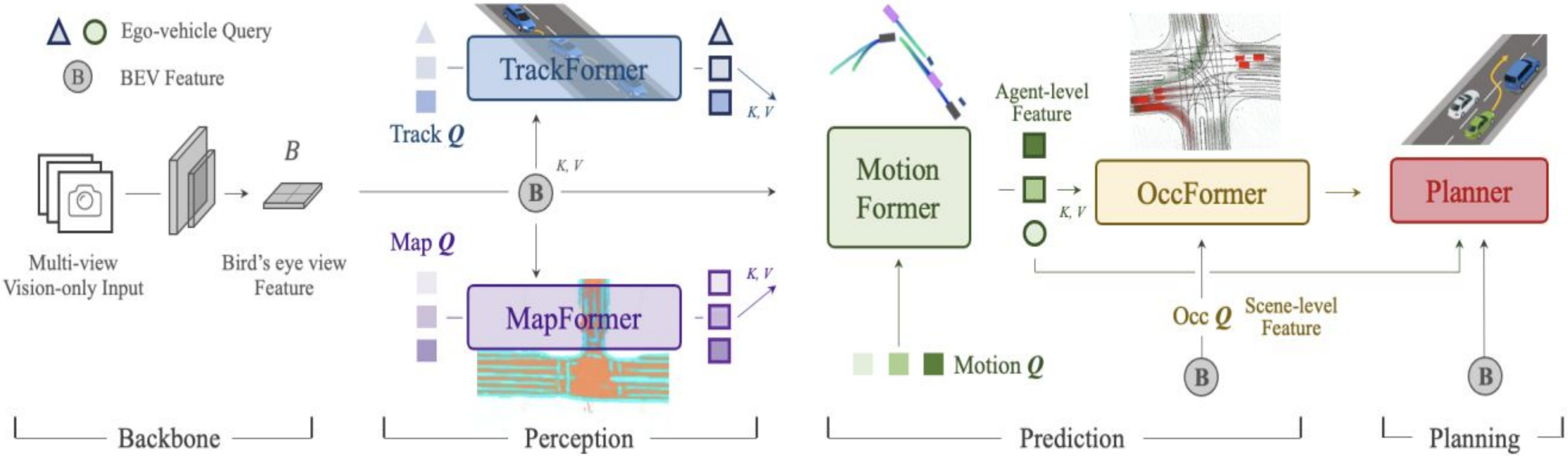
Background

- **Traditional autonomous driving systems** often rely on standalone models for individual tasks or multi-task learning paradigms with separate heads, which can lead to cascading errors and poor task coordination.
- **End-to-end approaches** have emerged to unify perception, prediction, and planning but often lack interpretability and robustness in dynamic urban environments.
- **UniAD addresses these challenges** by explicitly modeling intermediate representations (e.g., occupancy maps, agent trajectories) and optimizing the system for planning as the ultimate goal.

Design	Approach	Perception			Prediction		Plan
		Det.	Track	Map	Motion	Occ.	
(b)	NMP [101]	✓			✓		✓
	NEAT [19]			✓			✓
	BEVerse [105]	✓		✓		✓	
(c.1)	[14, 16, 78, 97]						✓
(c.2)	PnPNet† [57]	✓	✓		✓		
	ViP3D† [30]	✓	✓		✓		
	P3 [82]					✓	✓
	MP3 [11]			✓		✓	✓
	ST-P3 [38]			✓		✓	✓
	LAV [15]	✓		✓	✓		✓
(c.3)	**UniAD** (ours)	✓	✓	✓	✓	✓	✓

NYU

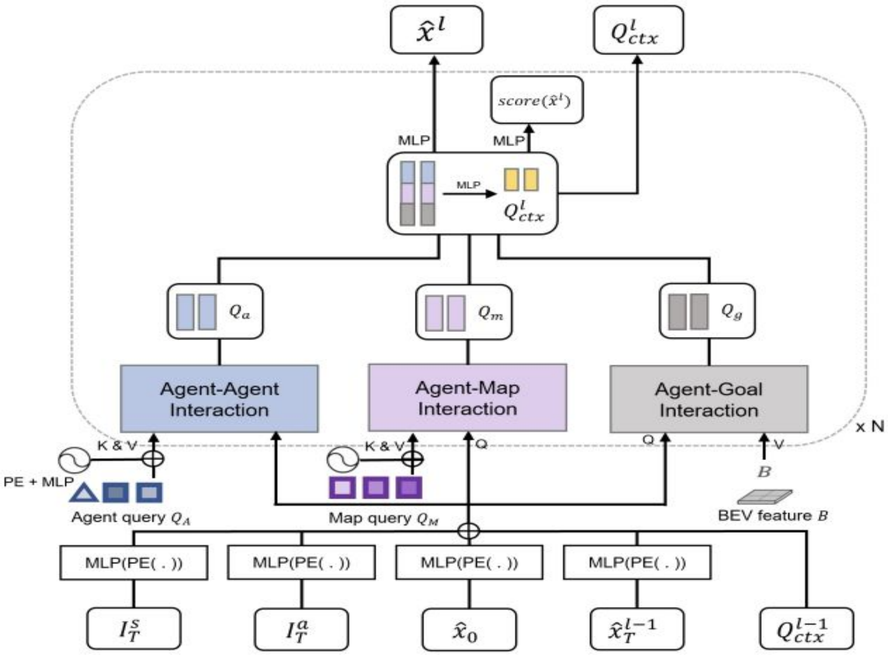
Model Architecture



NYU

Motionformer

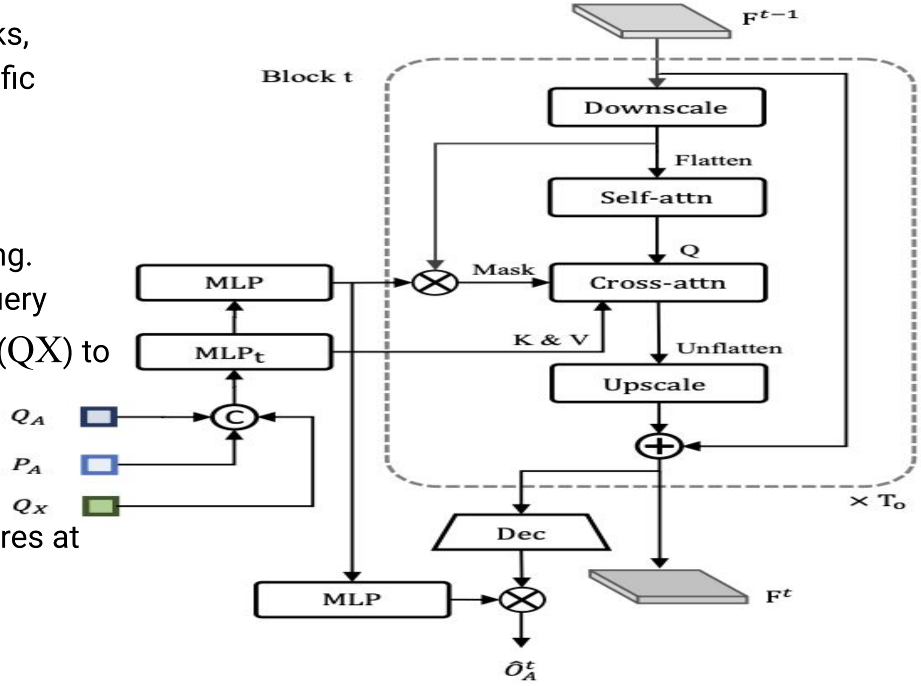
- **Structure**: MotionFormer consists of N stacked transformer layers for agent-agent, agent-map, and agent-goal interactions.
- **Modules**:
  - Agent-agent and agent-map interactions use standard transformer decoder layers.
  - Agent-goal interaction is based on the deformable cross-attention module.
- **Inputs:**
- ITs: Scene-level anchor endpoint.
- ITa: Clustered agent-level anchor endpoint.
- x^00: Current position of the agent.
- x^Tl−1: Predicted goal point from the previous layer.
- Qctxl−1: Query context from the preceding layer.



NYU

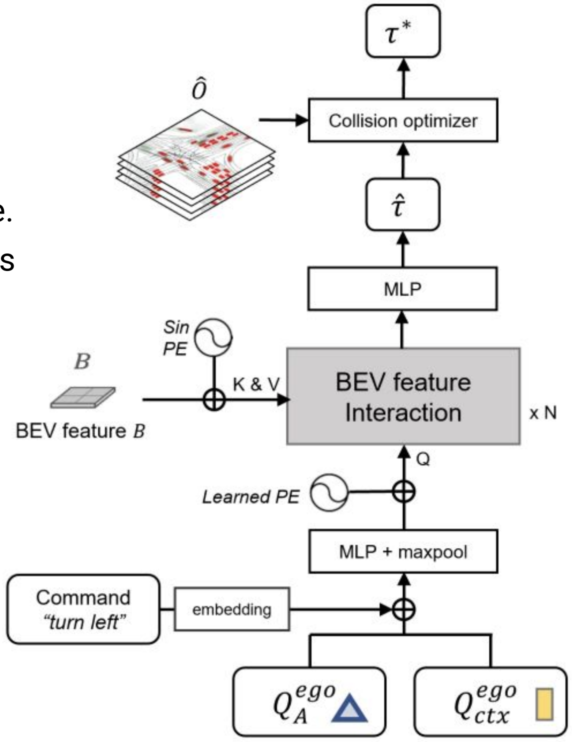
OccFormer

- **Structure**: OccFormer comprises To sequential blocks, where each block predicts the occupancy for a specific frame within the temporal horizon.
- **Features Incorporated:**
    - **Dense Scene Features:** Encoded from BEV representations for global scene understanding.
    - **Sparse Agent Features:** Derived from track query (QA), agent position (PA), and motion query (QX) to inject agent-level knowledge.
- **Instance-Level Occupancy:**
    - Generated via matrix multiplication between agent-level features and decoded dense features at the end of each block (O^At)



Planner

- **Inputs:**
  - QegoA: Ego-vehicle query from the tracking module.
  - Qegoctx: Ego-vehicle query from the motion forecasting module.
  - High-level command embeddings indicating navigation directions (e.g., turn left, go straight).
- **Processing:**
  - Queries are encoded via MLP layers and aggregated using max-pooling to select salient modal features.
  - BEV feature interaction is performed using stacked transformer decoder layers (N layers).
- **Output:**
  - Predicts future waypoints (τ^) for ego-vehicle planning while optimizing trajectories to avoid collisions based on predicted occupancy maps (O^).



Loss Function

$$L_1 = L_{\text{track}} + L_{\text{map}}.$$

$$L_2 = L_{\text{track}} + L_{\text{map}} + L_{\text{motion}} + L_{\text{occ}} + L_{\text{plan}}.$$

Stage One Loss Function

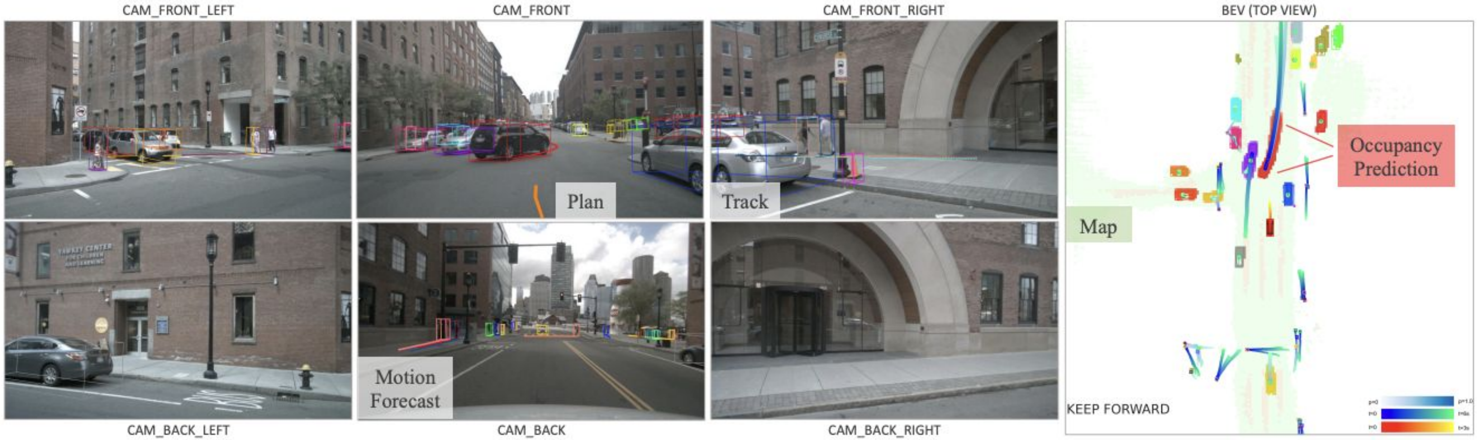
Combines tracking loss (Hungarian loss with Focal and L1 components) and mapping loss (Focal, L1, GIoU, and Dice losses) to pre-train perception tasks:

Stage Two Loss Function

Integrates all task-specific losses (tracking, mapping, motion forecasting, occupancy prediction, and planning) for end-to-end training

NYU

Qualitative Results



- **Task Results:** Predictions from motion and occupancy modules are consistent, visualized in surround-view images and BEV.
- **Ego-Vehicle Behavior:** Ego vehicle yields to a front black car, demonstrating safe decision-making.
- **Agent Representation:** Each agent is illustrated with a unique color for clarity.
- **Trajectory Visualization:**
  - Image View: Displays top-1 trajectory from motion forecasting.
  - BEV View: Shows top-3 trajectories for better spatial understanding.

NYU

Quantitative Results

Method	AMOTA↑	AMOTP↓	Recall↑	IDS↓
Immortal Tracker† [93]	0.378	1.119	0.478	936
ViP3D [30]	0.217	1.625	0.363	-
QD3DT [36]	0.242	1.518	0.399	-
MUTR3D [104]	0.294	1.498	0.427	3822
UniAD	**0.359**	**1.320**	**0.467**	**906**

Multi-object tracking

Method	Lanes↑	Drivable↑	Divider↑	Crossing↑
VPN [72]	18.0	76.0	-	-
LSS [76]	18.3	73.9	-	-
BEVFormer [55]	23.9	**77.5**	-	-
BEVerse† [105]	-	-	**30.6**	**17.2**
UniAD	**31.3**	69.1	25.7	13.8

Online mapping

- **UniAD Performance:** Outperforms previous end-to-end MOT techniques with image inputs on all metrics.
- **Comparison Note:** Tracking-by-detection methods with post-association are implemented using BEVFormer for fair evaluation.

- **Performance:** UniAD achieves competitive results against state-of-the-art perception-oriented methods with comprehensive road semantics.
- **Segmentation Metric:** Reports segmentation IoU (%) for lanes, drivable areas, dividers, and crossings.
- **Comparison Note:** Methods are implemented with BEVFormer for fair evaluation

NYU

Quantitative Results

Method	minADE(m)↓	minFDE(m)↓	MR↓	EPA↑
PnPNet† [57]	1.15	1.95	0.226	0.222
ViP3D [30]	2.05	2.84	0.246	0.226
Constant Pos.	5.80	10.27	0.347	-
Constant Vel.	2.13	4.01	0.318	-
UniAD	**0.71**	**1.02**	**0.151**	**0.456**

Motion forecasting.

Method	IoU-n.↑	IoU-f.↑	VPQ-n.↑	VPQ-f.↑
FIERY [35]	59.4	36.7	50.2	29.9
StretchBEV [1]	55.5	37.1	46.0	29.0
ST-P3 [38]	-	38.9	-	32.1
BEVerse† [105]	61.4	**40.9**	54.3	**36.1**
UniAD	**63.4**	40.2	**54.7**	33.5

Occupancy prediction

- **Performance:** UniAD significantly outperforms prior vision-based end-to-end methods across all metrics.
- **Comparative Settings:** Evaluated with two vehicle modeling settings—constant positions and constant velocities.
- **Reimplementation:** Prior methods reimplemented with BEVFormer for fair comparisons.

- **Improvement in Nearby Areas:** UniAD achieves significant gains in near evaluation ranges (30×30m), critical for planning accuracy.
- **Evaluation Ranges:** Results are reported for "n." (near) and "f." (far, 50×50m) evaluation ranges.
- **Training Note:** Models trained with heavy augmentations yield improved occupancy prediction metrics.

Quantitative Results

Method	L2(m)↓				Col. Rate(%)↓			
	$1s$	$2s$	$3s$	Avg.	$1s$	$2s$	$3s$	Avg.
NMP† [101]	-	-	2.31	-	-	-	1.92	-
SA-NMP† [101]	-	-	2.05	-	-	-	1.59	-
FF† [37]	0.55	1.20	2.54	1.43	0.06	0.17	1.07	0.43
EO† [47]	0.67	1.36	2.78	1.60	0.04	0.09	0.88	0.33
ST-P3 [38]	1.33	2.11	2.90	2.11	0.23	0.62	1.27	0.71
UniAD	**0.48**	**0.96**	**1.65**	**1.03**	**0.05**	**0.17**	**0.71**	**0.31**

Planning

- **Performance:** UniAD achieves the lowest L2 error and collision rate across all time intervals.
- **Comparison:** Outperforms LiDAR-based methods in most cases, demonstrating superior safety.
- **Validation:** Results verify the effectiveness of integrating motion and occupancy prediction for safe planning.

Ablation Study

ID	Scene-l. Anch.	Goal Inter.	Ego Q	NLO.	minADE↓	minFDE↓	MR↓	minFDE-mAP*↑
1					0.844	1.336	0.177	0.246
2	✓				0.768	1.159	0.164	0.267
3	✓	✓			0.755	1.130	0.168	0.264
4	✓	✓	✓		0.747	1.096	0.156	0.266
5	✓	✓	✓	✓	**0.710**	**1.004**	**0.146**	**0.273**

ID	Cross. Attn.	Attn. Mask	Mask Feat.	IoU-n.↑	IoU-f.↑	VPQ-n.↑	VPQ-f.↑
1				61.2	**39.7**	51.5	31.8
2	✓			61.3	39.4	51.0	31.8
3	✓	✓		62.3	**39.7**	52.4	32.5
4	✓	✓	✓	**62.6**	39.5	**53.2**	**32.8**

ID	BEV Att.	Col. Loss	Occ. Optim.	L2↓			Col. Rate↓		
				1s	2s	3s	1s	2s	3s
1				**0.44**	**0.99**	**1.71**	0.56	0.88	1.64
2	✓			**0.44**	1.04	1.81	0.35	0.71	1.58
3	✓	✓		**0.44**	1.02	1.76	0.30	0.51	1.39
4	✓	✓	✓	0.54	1.09	1.81	**0.13**	**0.42**	**1.05**

Ablation for designs in the **motion forecasting module**

Ablation for designs in the **occupancy prediction module**

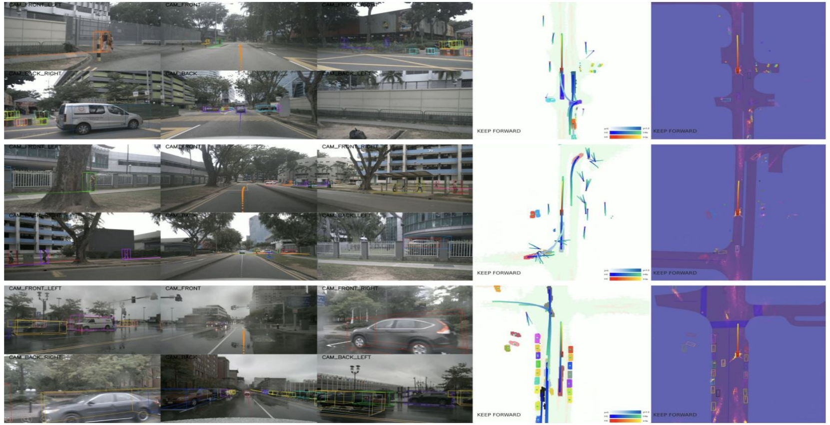
Ablation for designs in the **planning module**

NYU

Strengths

- **UniAD** integrates **perception, prediction, and planning** into a unified end-to-end framework for enhanced coordination.
- **Query-based design enables flexible feature sharing** across tasks, improving accuracy and task interaction.
- Achieved **state-of-the-art performance** in motion forecasting, occupancy prediction, and safe planning metrics.
- **Reduces cascading errors and enhances interpretability** through explicit intermediate representations.

NYU



Cruising around urban areas

Obstacles avoidance visualization

Weakness

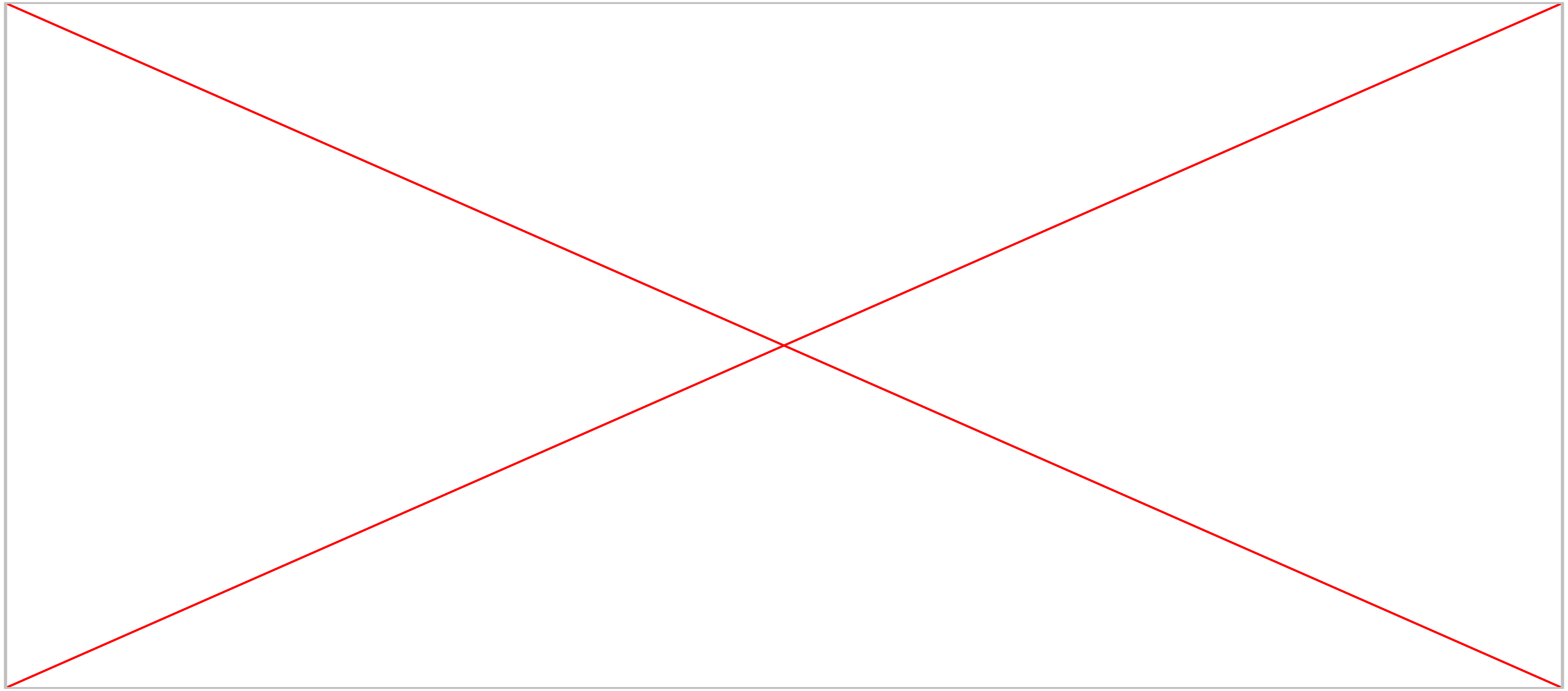
- **High computational complexity** limits deployment on resource-constrained platforms.
- Struggles with **long-tail scenarios** like large trailers or poorly lit environments.
- Adding more tasks may increase **system complexity and training difficulty**.

NYU

Applications in Embodied environment

1. **Urban Autonomous Driving:** Real-time navigation in dense traffic, handling tasks like obstacle avoidance and pedestrian yielding.
2. **Simulated Driving (CARLA):** Testing UniAD's performance in diverse traffic scenarios such as intersections and roundabouts.
3. **Warehouse Robots:** Guiding autonomous robots for dynamic obstacle avoidance and route planning in warehouses.
4. **Collaborative Driving:** Coordinating vehicle-to-vehicle communication for safe and efficient traffic flow.

Result



Future Scope, and Extensions:

Conclusion:

- UniAD introduces a novel planning-oriented framework that unifies perception, prediction, and planning tasks, achieving state-of-the-art performance across multiple benchmarks.
- The query-based design ensures effective task coordination and interpretability, paving the way for safer and more robust autonomous driving systems.

Future Scope:

- Optimize the framework for lightweight deployment in real-time applications.
- Extend UniAD to include additional tasks like depth estimation and behavior prediction.
- Explore vehicle-to-vehicle communication for collaborative driving scenarios.

NYU


Anybody got any questions?