Planning-oriented Autonomous Driving

By Yihan Hu* Jiazhi Yang* Li Chen*+ Keyu Li*

CVPR 2023

Seminar AI4AD

Advances in AI for Autonomous Driving

Presented by Camilo Martínez

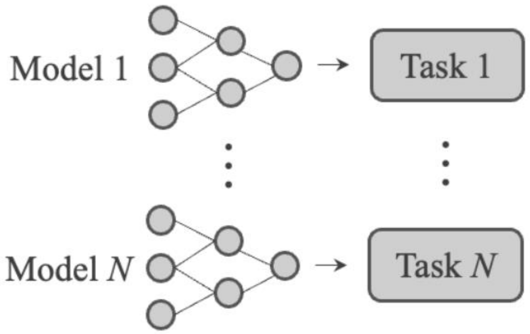
UNIVERSITÄT DES SAARLANDES

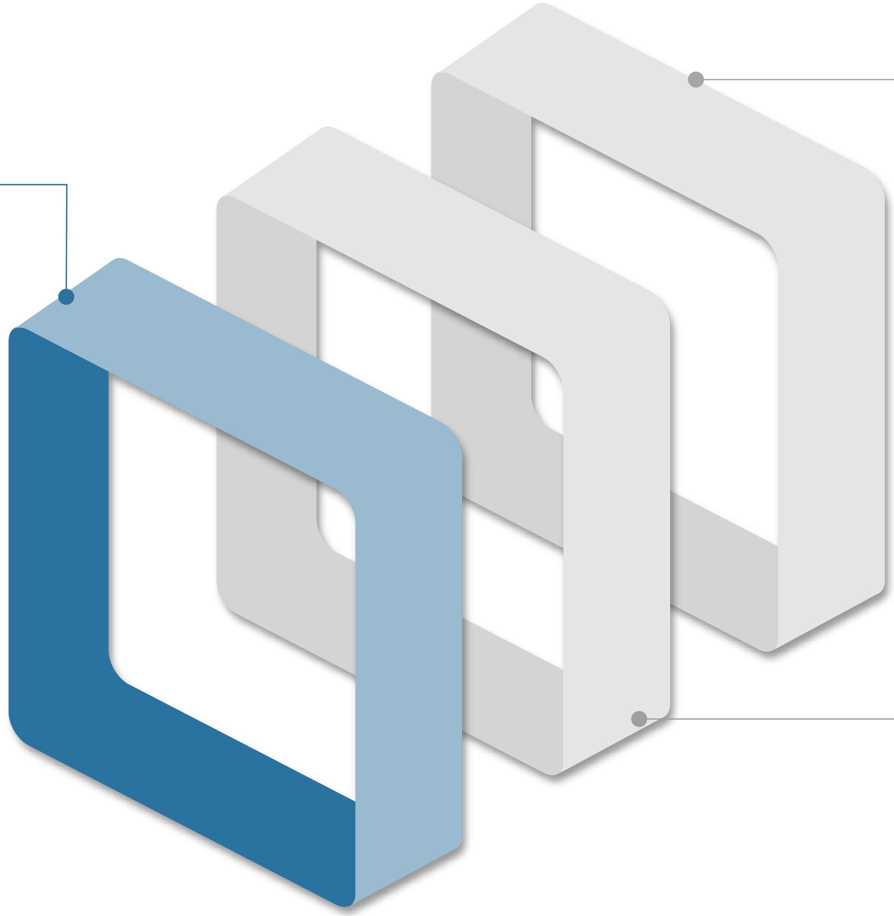
Table of Contents

Motivation

1 Standalone Models

They bare the risk of
- information loss
- error accumulation
- feature misalignment



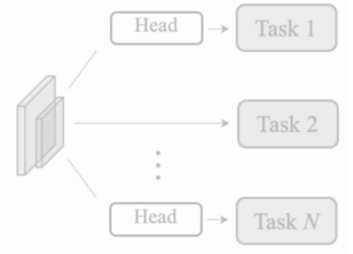


3 End-to-End Systems

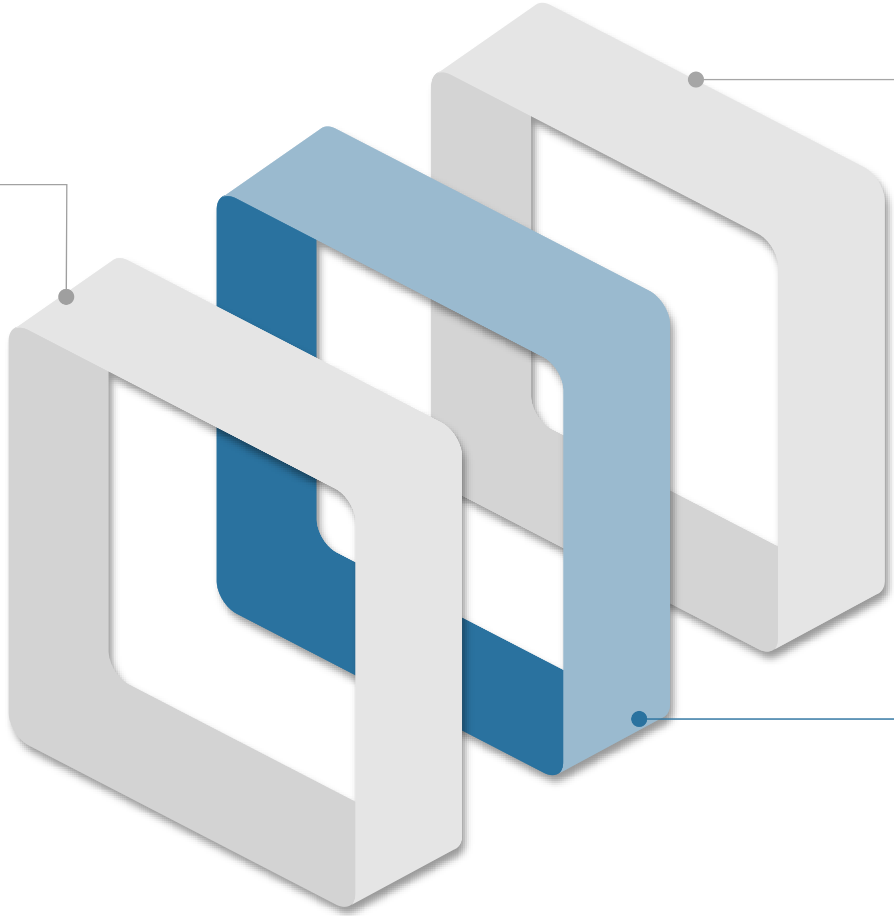
They unite **perception**, **prediction** and **planning** as a whole.

2 Multi-Task Learning

May cause undesirable "**negative transfer**".

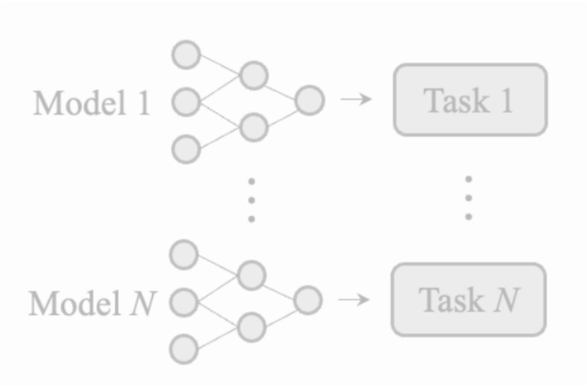


Motivation



1 **Standalone Models**

They bare the risk of
- information loss
- error accumulation
- feature misalignment

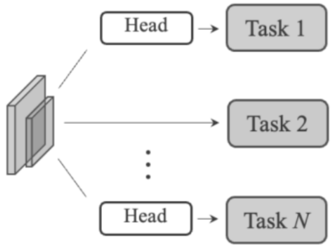


3 **End-to-End Systems**

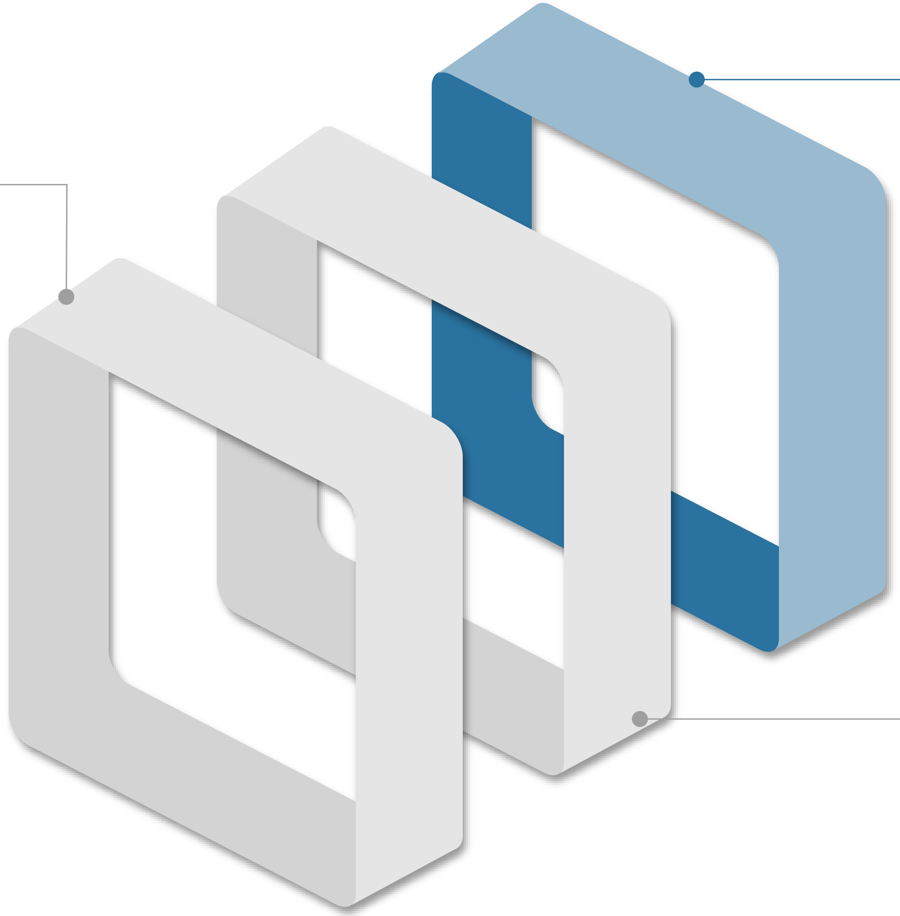
They unite **perception**, **prediction** and **planning** as a whole.

2 **Multi-Task Learning**

May cause undesirable "**negative transfer**".

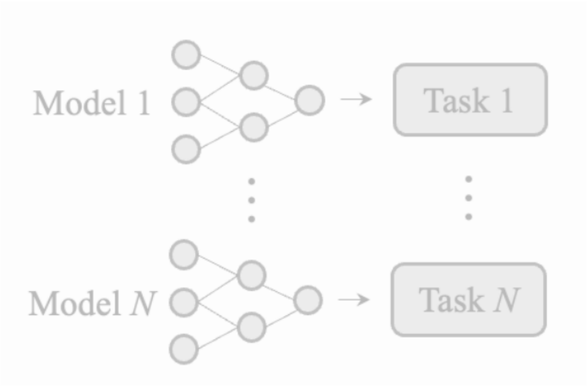


Motivation



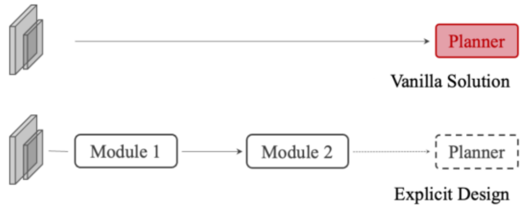
1 **Standalone Models**

They bare the risk of
- information loss
- error accumulation
- feature misalignment



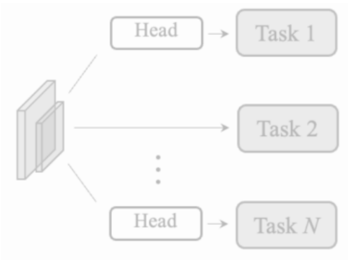
3 **End-to-End Systems**

They unite **perception**, **prediction** and **planning** as a whole.

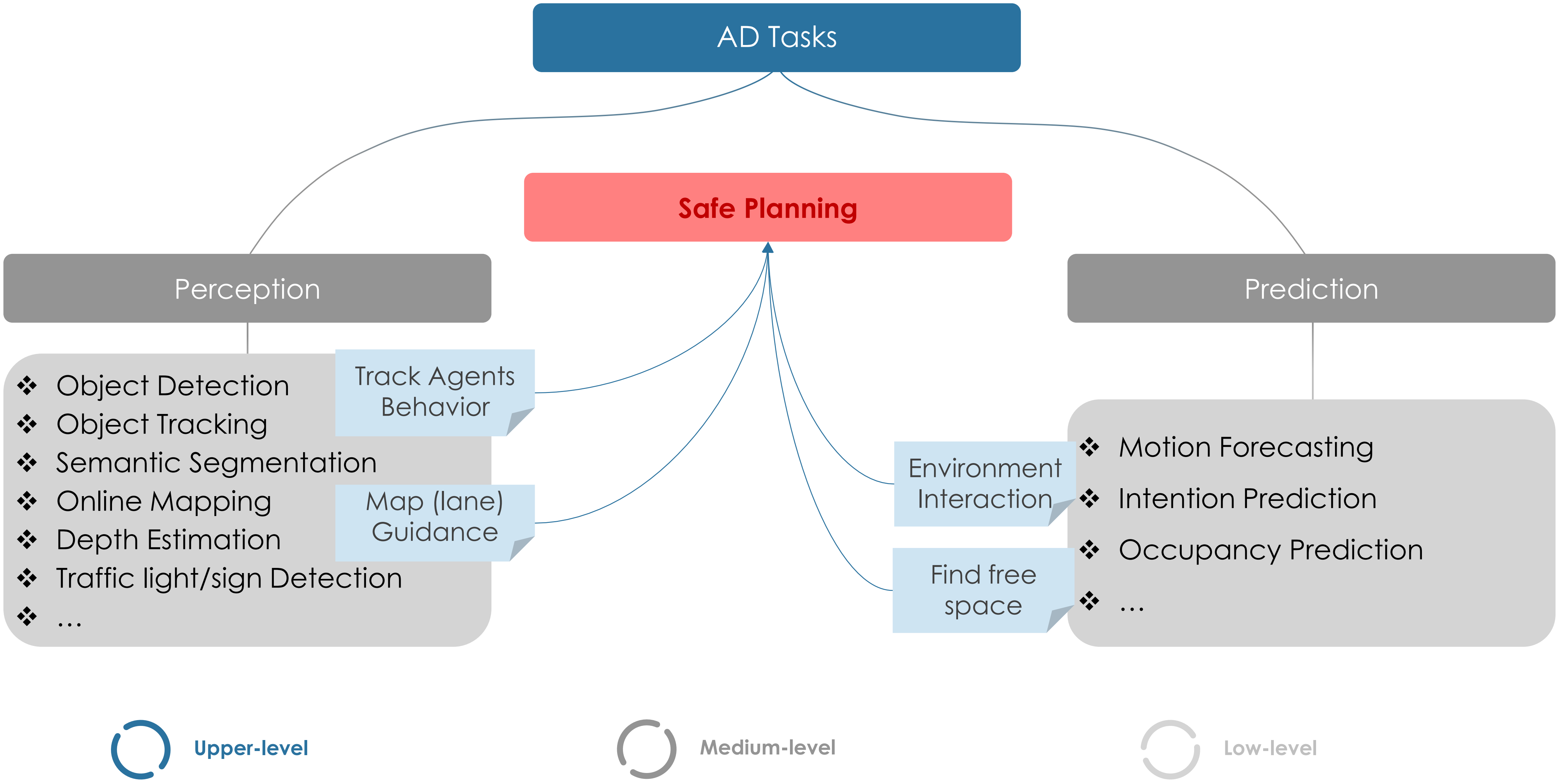


2 **Multi-Task Learning**

May cause undesirable "**negative transfer**".



Which Tasks?


AD Tasks
Safe Planning
Perception
Prediction
Object Detection
Object Tracking
Semantic Segmentation
Online Mapping
Depth Estimation
Traffic light/sign Detection
…
Track Agents Behavior
Map (lane) Guidance
Environment Interaction
Find free space
Motion Forecasting
Intention Prediction
Occupancy Prediction
…
Upper-level
Medium-level
Low-level

The UniAD

GOALS

- Unify full-stack AD **tasks**
- Coordinate all task towards **safe planning**

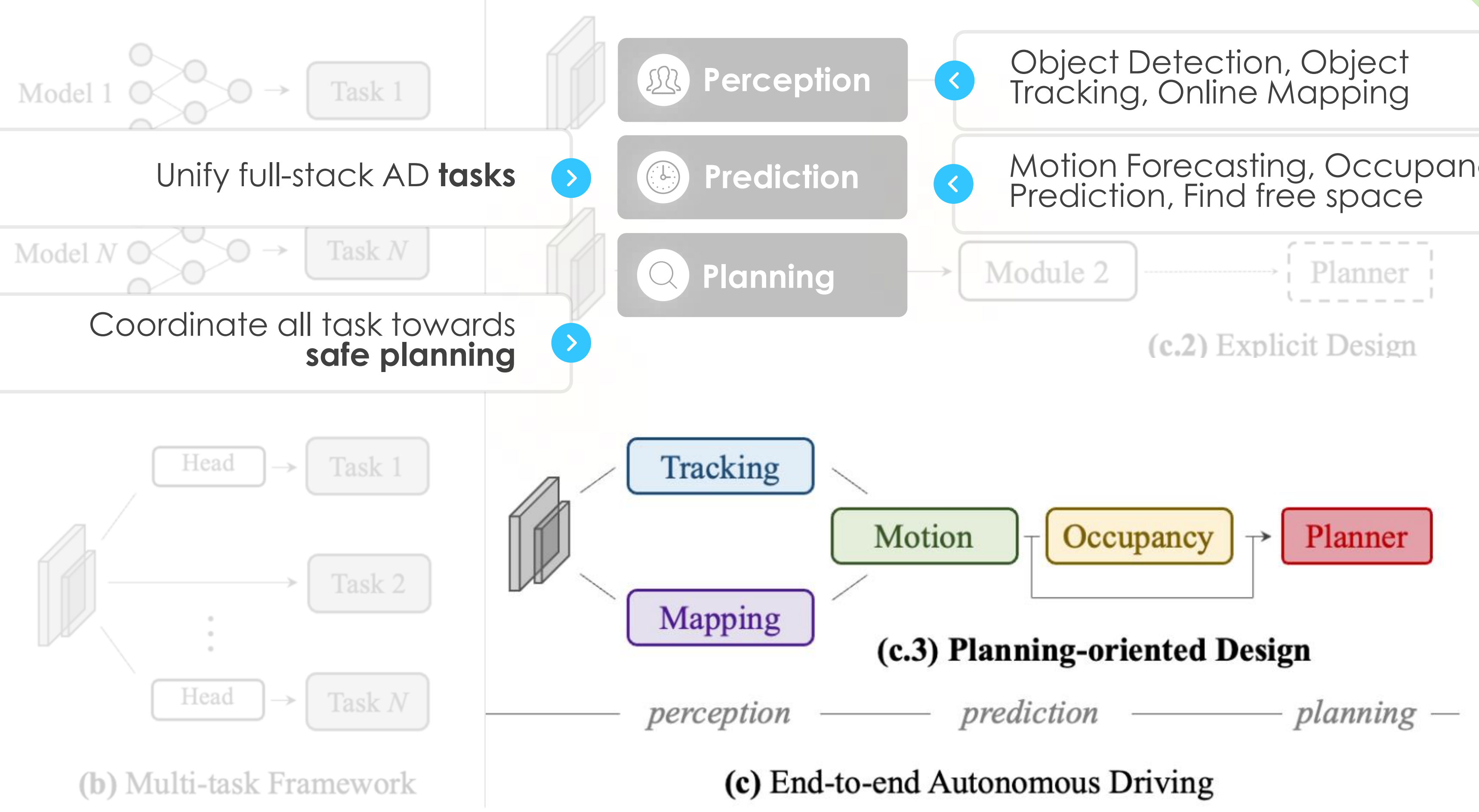
Perception — Object Detection, Object Tracking, Online Mapping

- Track Agents Behavior
- Guide with lanes

Prediction — Motion Forecasting, Occupancy Prediction, Find free space

- Environment Interaction

Planning



Tasks Comparison & Taxonomy

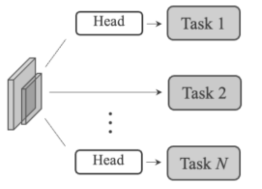
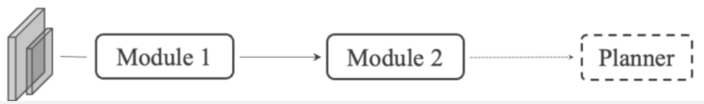
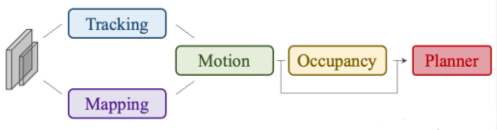
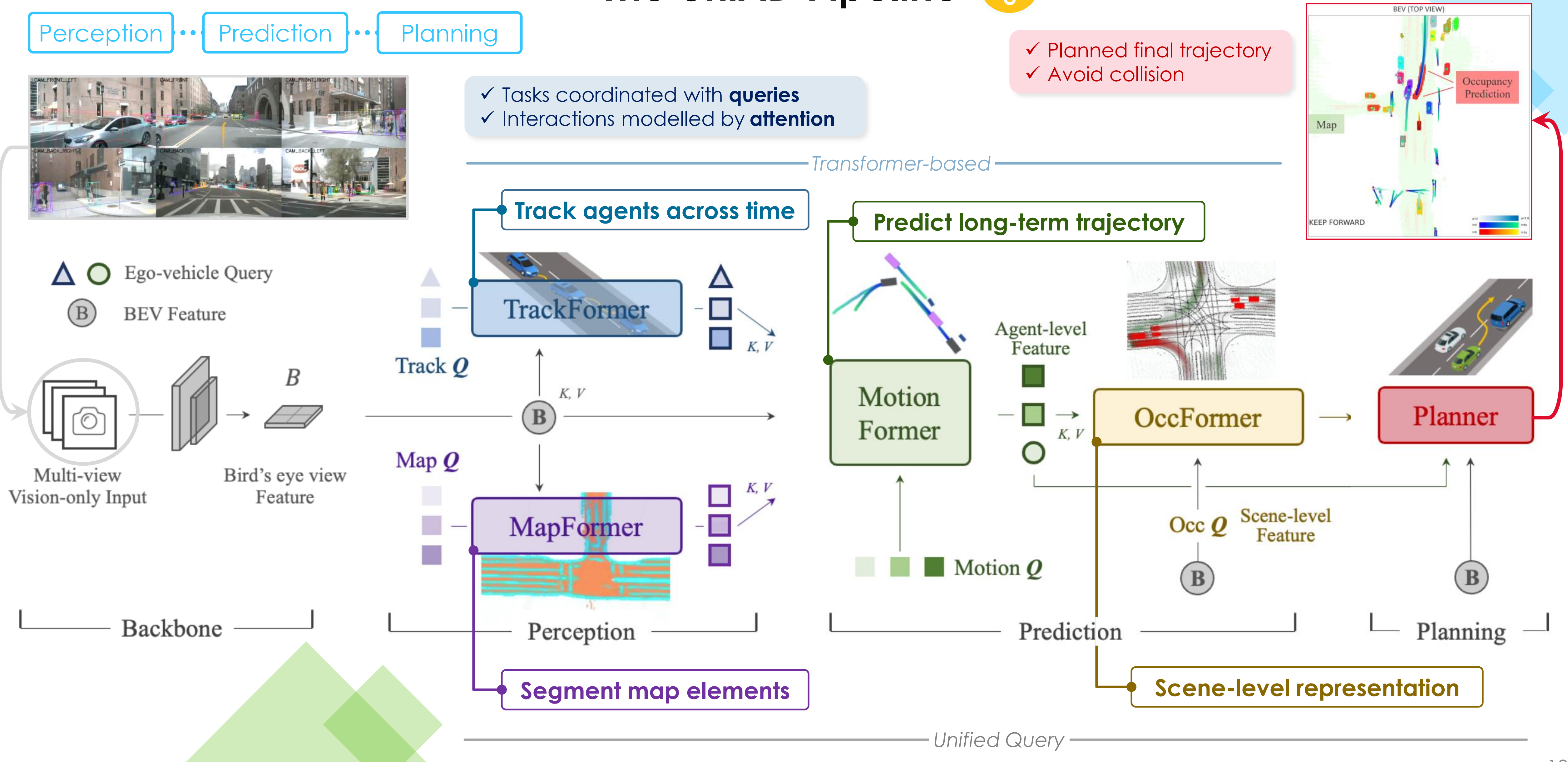
DESIGN	APPROACH	WORK	PERCEPTION			PREDICTION		PLAN
			Detect	Track	Map	Motion	Occup.	
Multi-task Framework 	NMP	CVPR 2019	✓			✓		✓
	NEAT	ICCV 2021			✓			✓
	BEVerse	2022	✓		✓		✓	
Explicit Design 	PnPNet	CVPR 2020	✓	✓		✓		
	ViP3D	CVPR 2023	✓	✓		✓		
	P3	ECCV 2020					✓	✓
	MP3	CVPR 2021			✓		✓	✓
	ST-P3	ECCV 2022			✓		✓	✓
	LAV	CVPR 2022	✓		✓	✓		✓
Vanilla Solution 	[7, 9, 45, 54]	2020-2022						✓
Planning Oriented Design 	**UniAD**	**CVPR 2023**	✓	✓	✓	✓	✓	✓

Table of Contents

The UniAD Pipeline

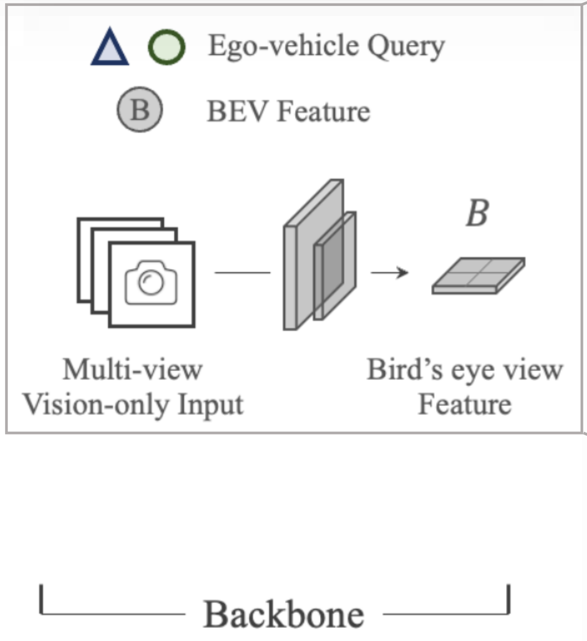


Perception ··· Prediction ··· Planning

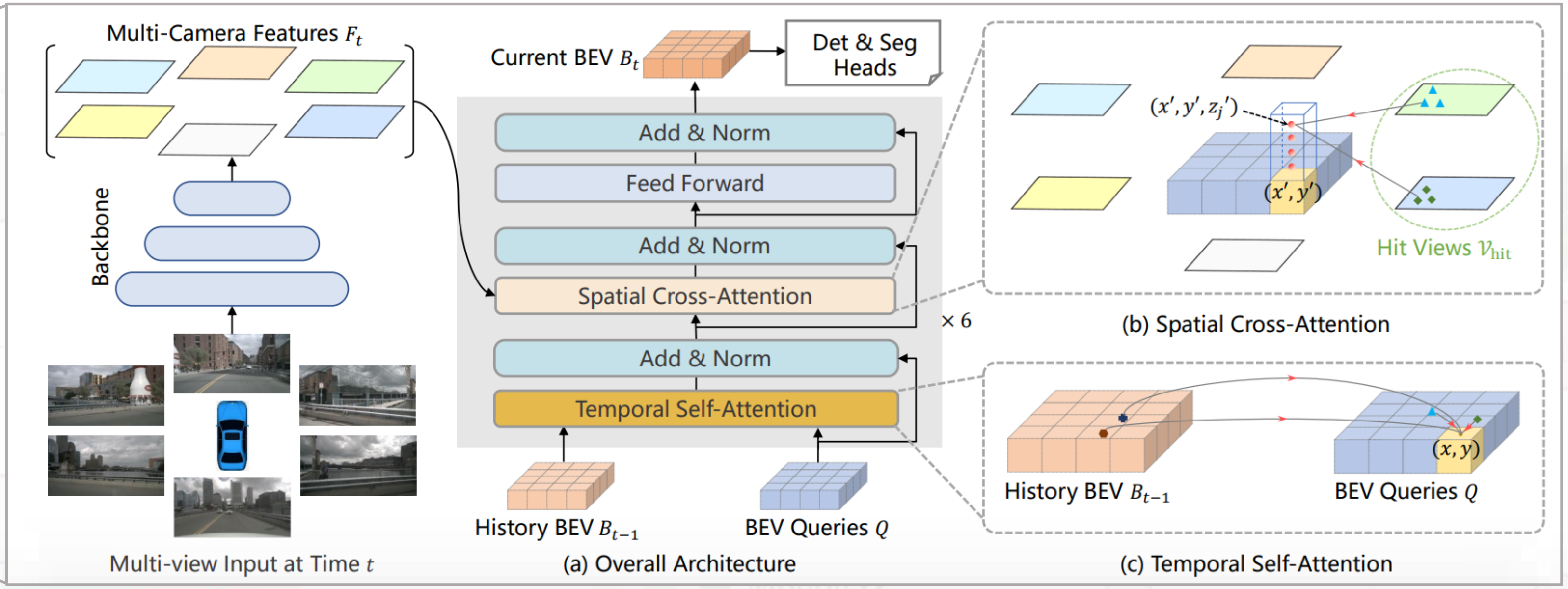
The UniAD Pipeline

INPUT

A sequence of multi-camera images are transformed into a unified bird's-eye-view (BEV) feature by BEVFormer.



BEV Encoder – BEVFormer (ECCV 2022)

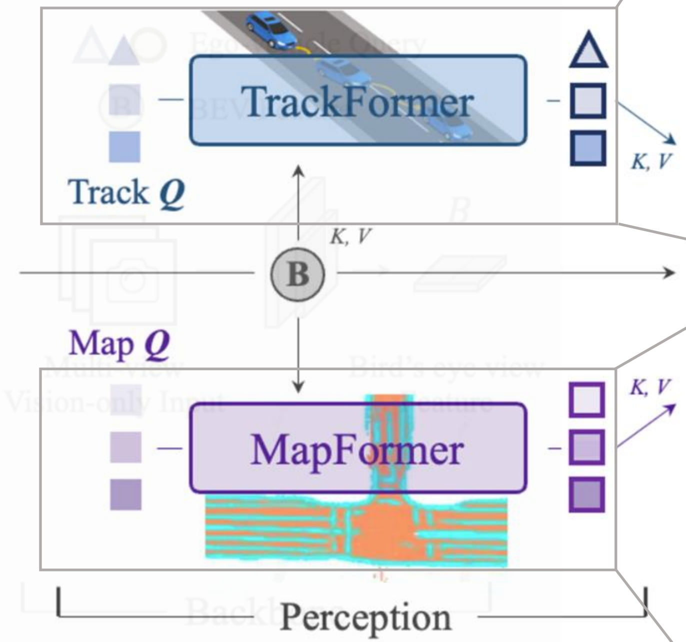


ADVANTAGES

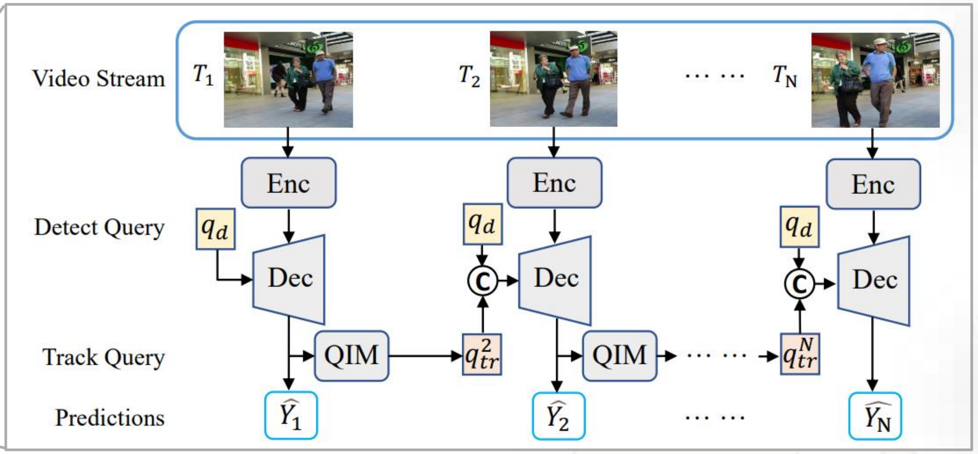
- Strong BEV encoder with **effective spatial** and **temporal** feature extraction.
- Not *really* part of the main pipeline of UniAD, that is, it's not confined to this one specifically.

Perception ··· Prediction ··· Planning

The UniAD Pipeline

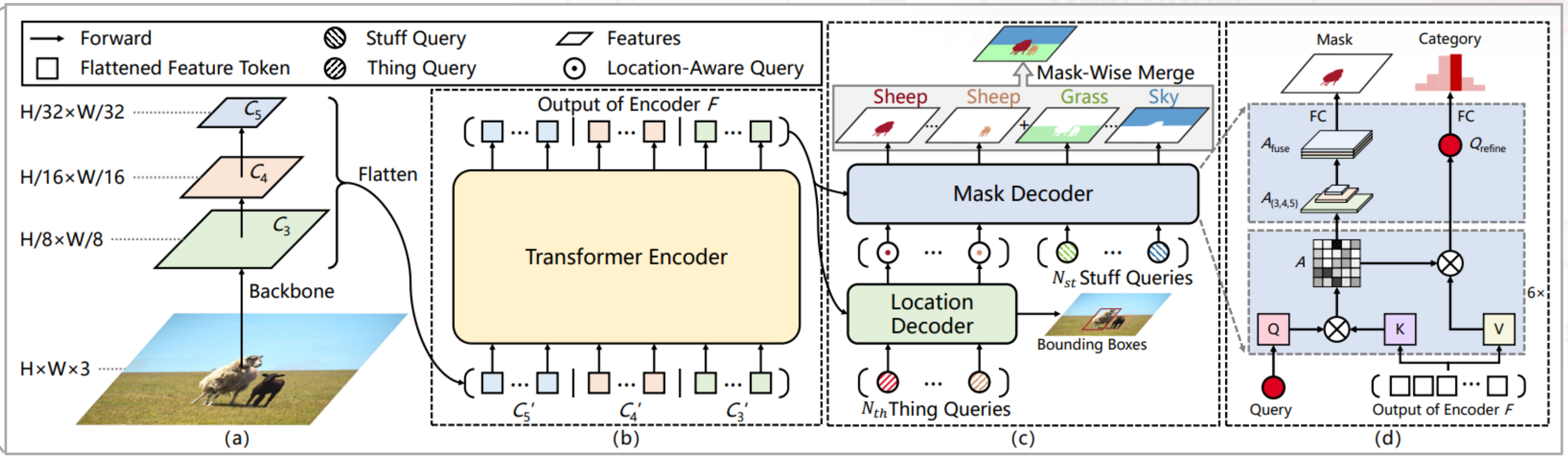


MOTR (ECCV 2022)



❖ End-to-end Multiple Object Tracking without post-association.

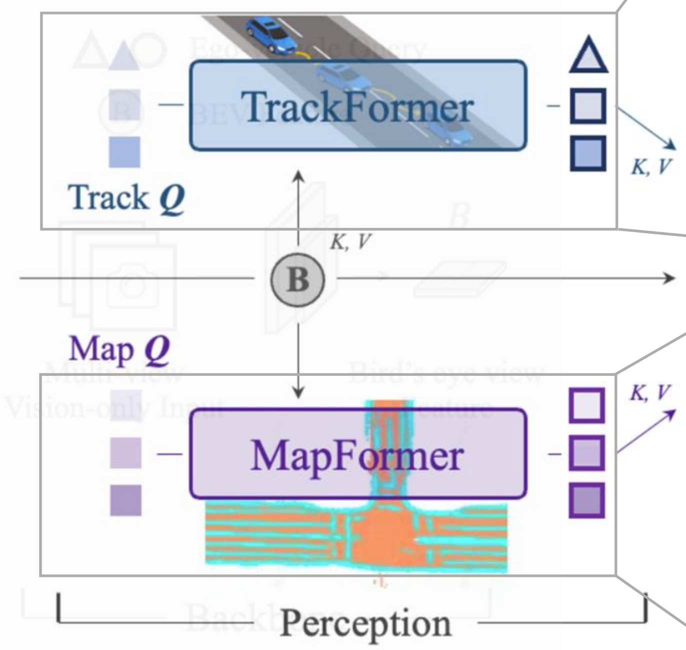
❖ Road elements are represented as map queries.



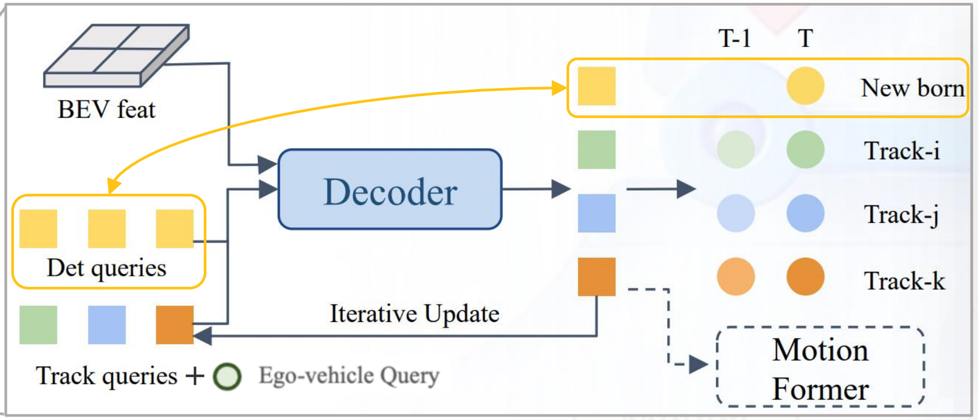
Panoptic SegFormer (CVPR 2022)

Perception ··· Prediction ··· Planning

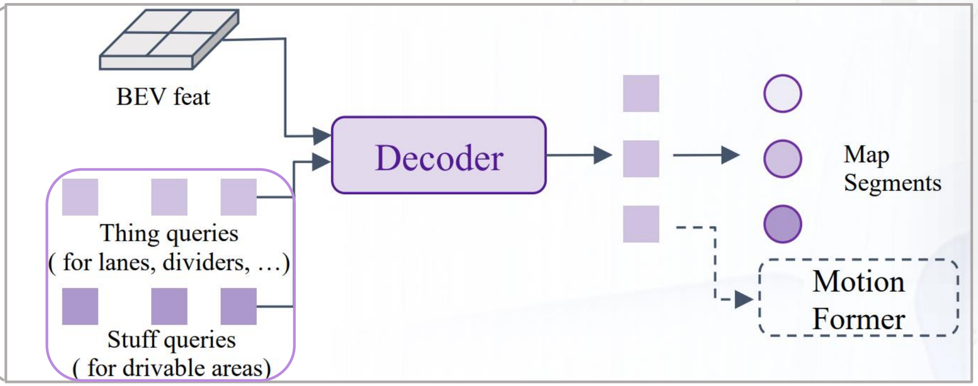
The UniAD Pipeline



MOTR (ECCV 2022)



- End-to-end Multiple Object Tracking without post-association
- Track queries ↔ agents' information
- **Detects** and **track** agents
- **Detection** queries are responsible for detecting **newborn agents**
- N layers → Q_A → N_a valid agents

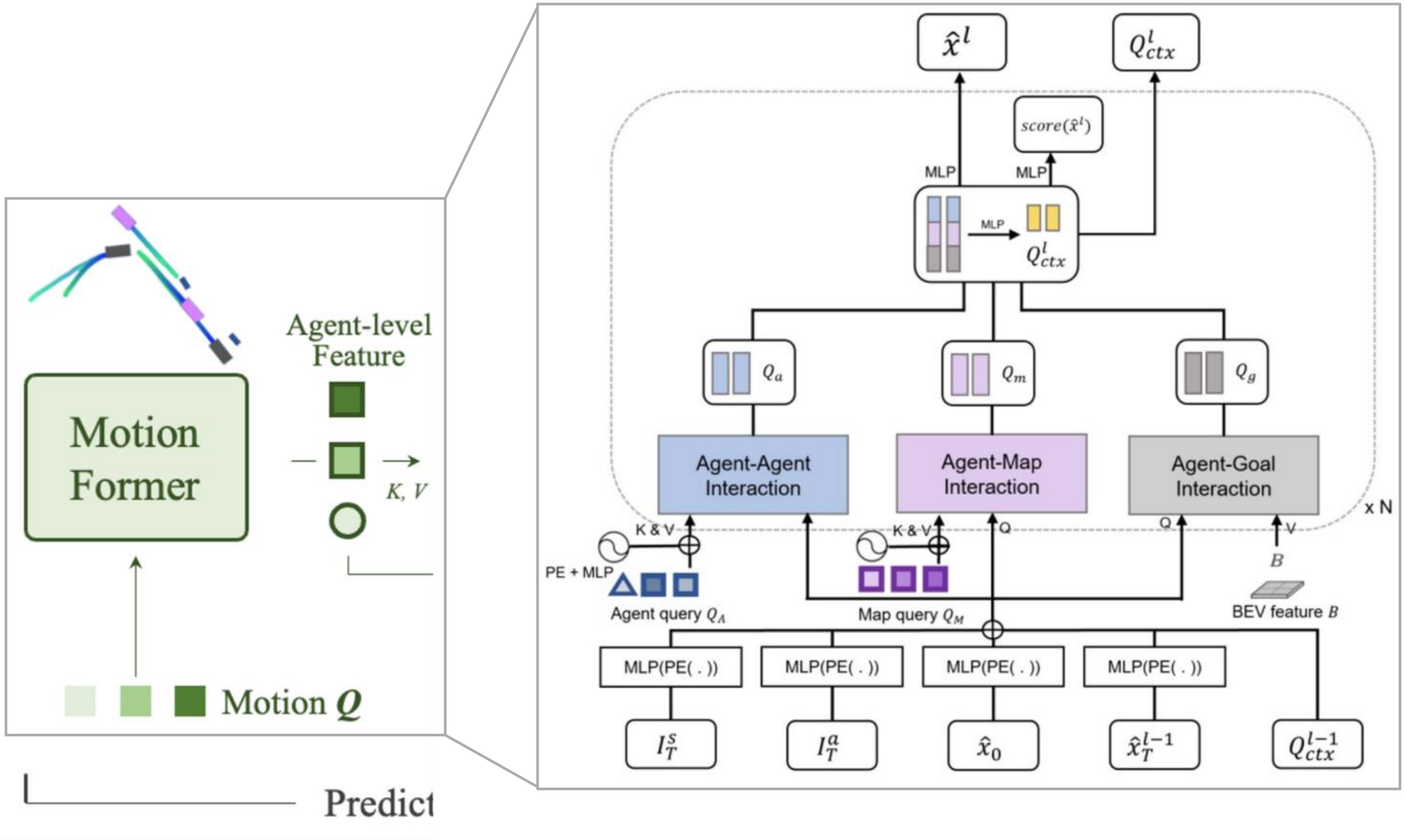


Panoptic SegFormer (CVPR 2022)

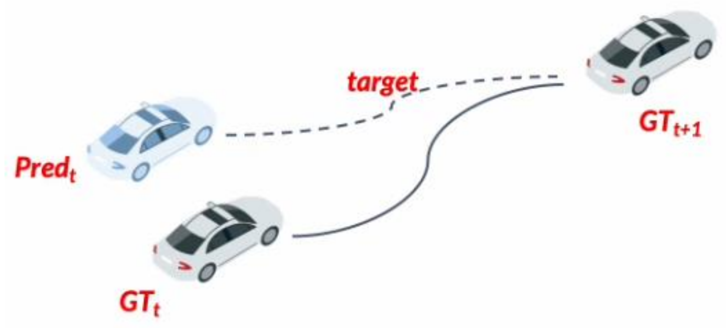
- Road elements are represented as map queries.
- Lanes, dividers, crossings as **things**
- Drivable area as **stuff**
- N layers → updated last layer Q_M

The UniAD Pipeline

Proposed in **UniAD**



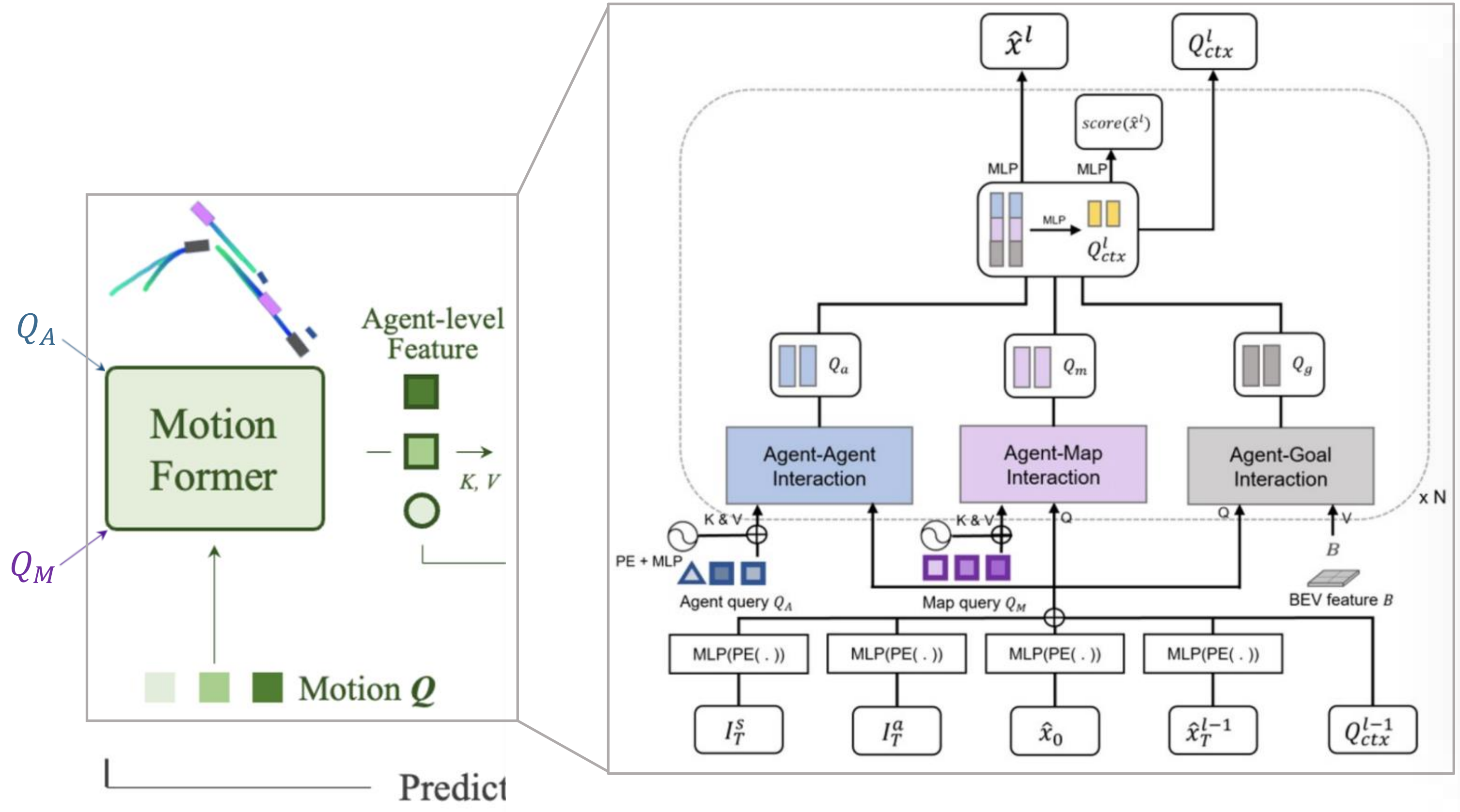
NON-LINEAR OPTIMIZATION



- Adjust ground-truth trajectory based on upstream prediction.
- Diverse relation modelings via attention mechanism:
  - agent-agent, agent-map, agent-goal

The UniAD Pipeline

Proposed in **UniAD**



TrackFormer Q_A
MapFormer Q_M
MotionFormer
N layers
Top-k possible trajectories

- Multi-agent trajectories in the frame with a single forward pass
- Ego-vehicle query interacts with other agents (**future dynamics**)

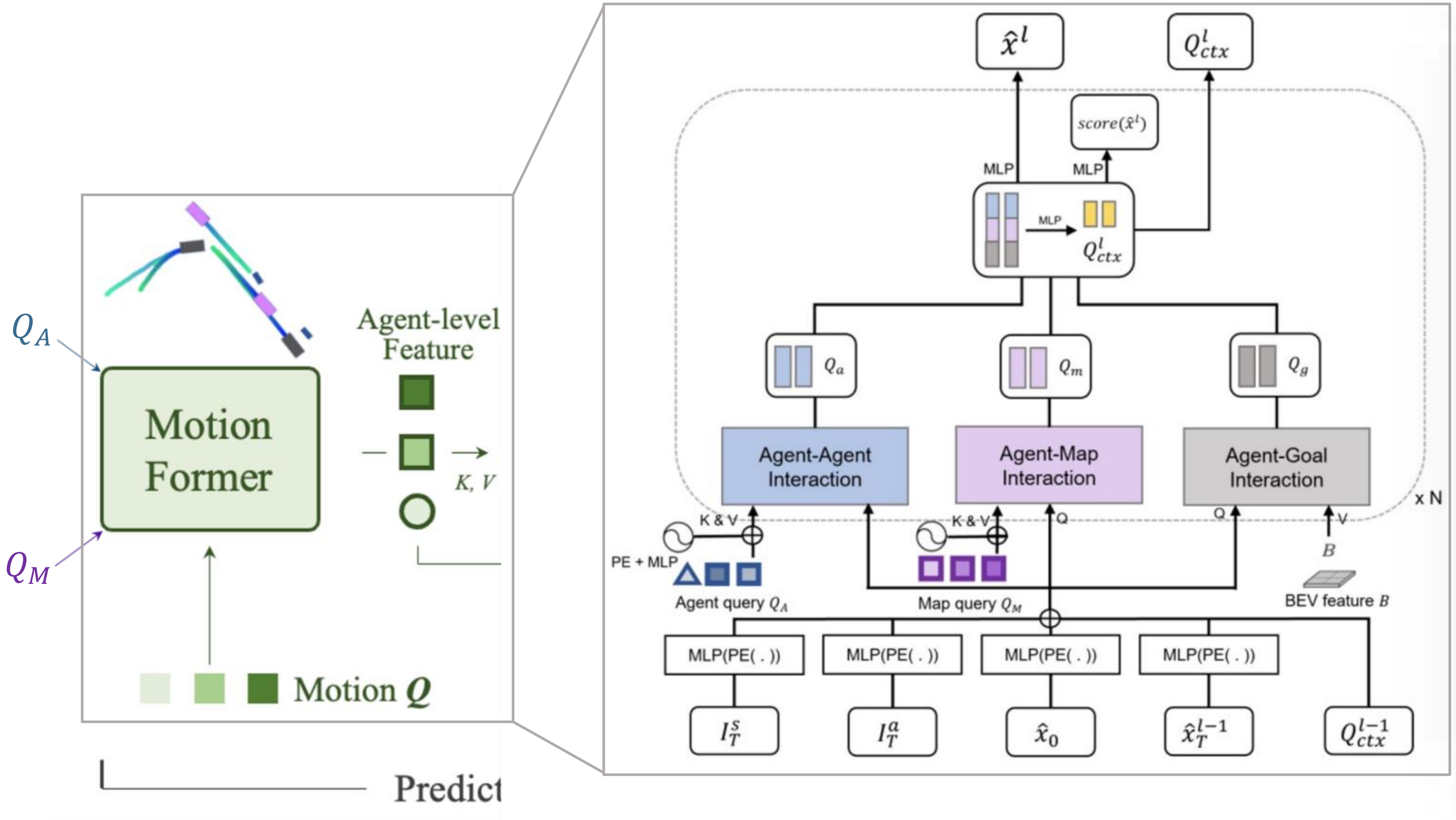
More formally,

$$\text{Output} := \{\hat{\mathbf{x}}_{i,k} \in \mathbb{R}^{T\times 2} \mid i = 1, \dots, N_a; k = 1, \dots, \mathcal{K}\}$$

- i indexes the agent
- k indexes the modality of trajectories (6)
- T is the length of **prediction horizon** (12)

The UniAD Pipeline

Proposed in **UniAD**



For a motion query $Q_{i,k}$,

agent-agent, agent-map

$$Q_{a/m} = \mathrm{MHCA}(\mathrm{MHSA}(Q), Q_A/Q_M)$$

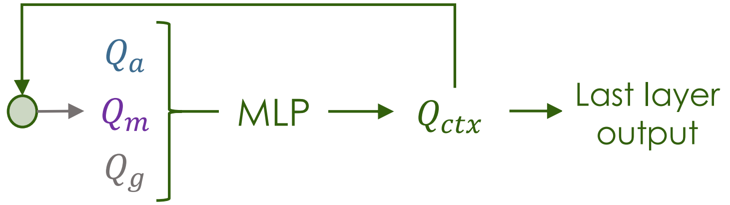
- MHCA: multi-head **cross attention**.
- MHSA: multi-head **self-attention**.

agent-goal

To focus on the intended position/goal:

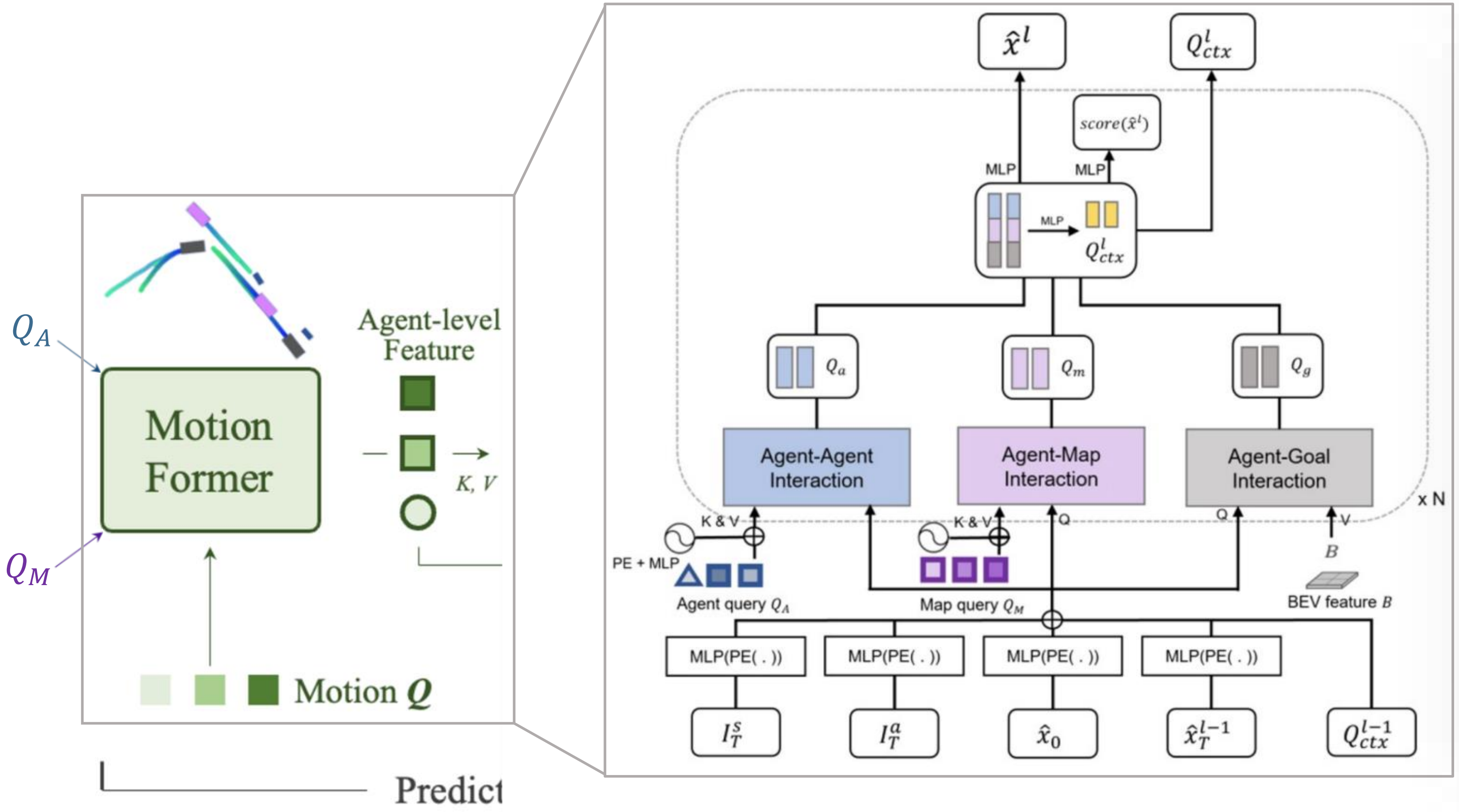
$$Q_g = \mathrm{DeformAttn}(Q, \hat{\mathbf{x}}_T^{l-1}, B)$$

- $\hat{\mathbf{x}}_T^{l-1}$: predicted trajectory of previous layer.



The UniAD Pipeline

Proposed in **UniAD**



Motion Queries: Q_{ctx} (previous layer) + Q_{pos}

$$Q_{\text{pos}} = \underbrace{\text{MLP}\big(\text{PE}(I^s)\big)}_{\text{Scene-level anchors}} + \underbrace{\text{MLP}\big(\text{PE}(I^a)\big)}_{\text{Agent-level anchors}}$$

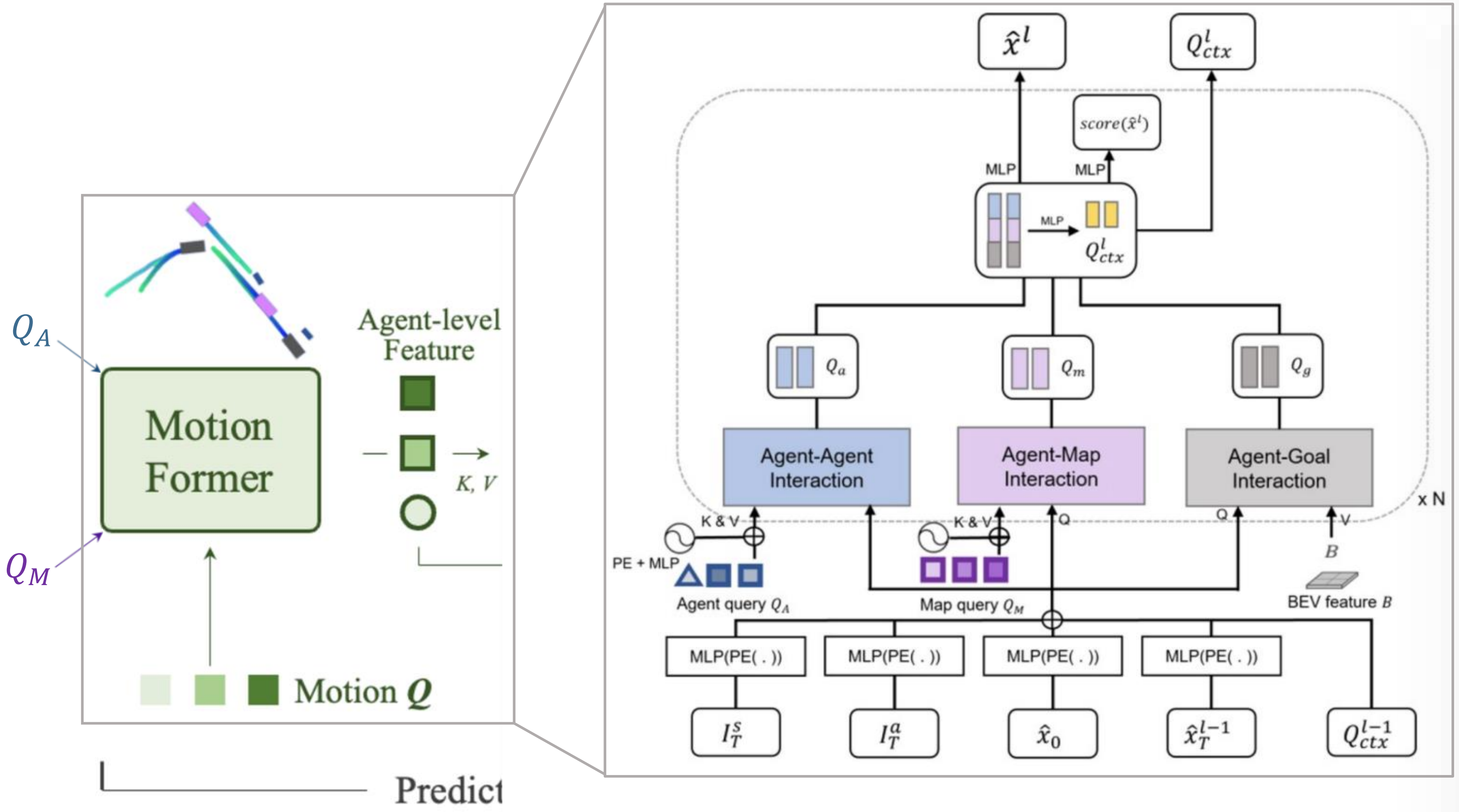
$$+ \underbrace{\text{MLP}\big(\text{PE}(\hat{\mathbf{x}}_0)\big)}_{\text{Current location}} + \underbrace{\text{MLP}\left(\text{PE}(\hat{\mathbf{x}}_T^{l-1})\right)}_{\text{Predicted goal point}}$$

- $\text{PE}(\cdot)$: sinusoidal positional encoding

❖ Predicted endpoint optimized layer-by-layer in a **coarse-to-fine** fashion.

The UniAD Pipeline

Proposed in **UniAD**



Tackle prediction uncertainty

Imperfect detection

regression

Ground-truth waypoints

- ❌ Unrealistic trajectories
- ❌ Large curvature
- ❌ Large acceleration

Instead, make them **physically feasible** given an imprecise starting point predicted by the upstream module:

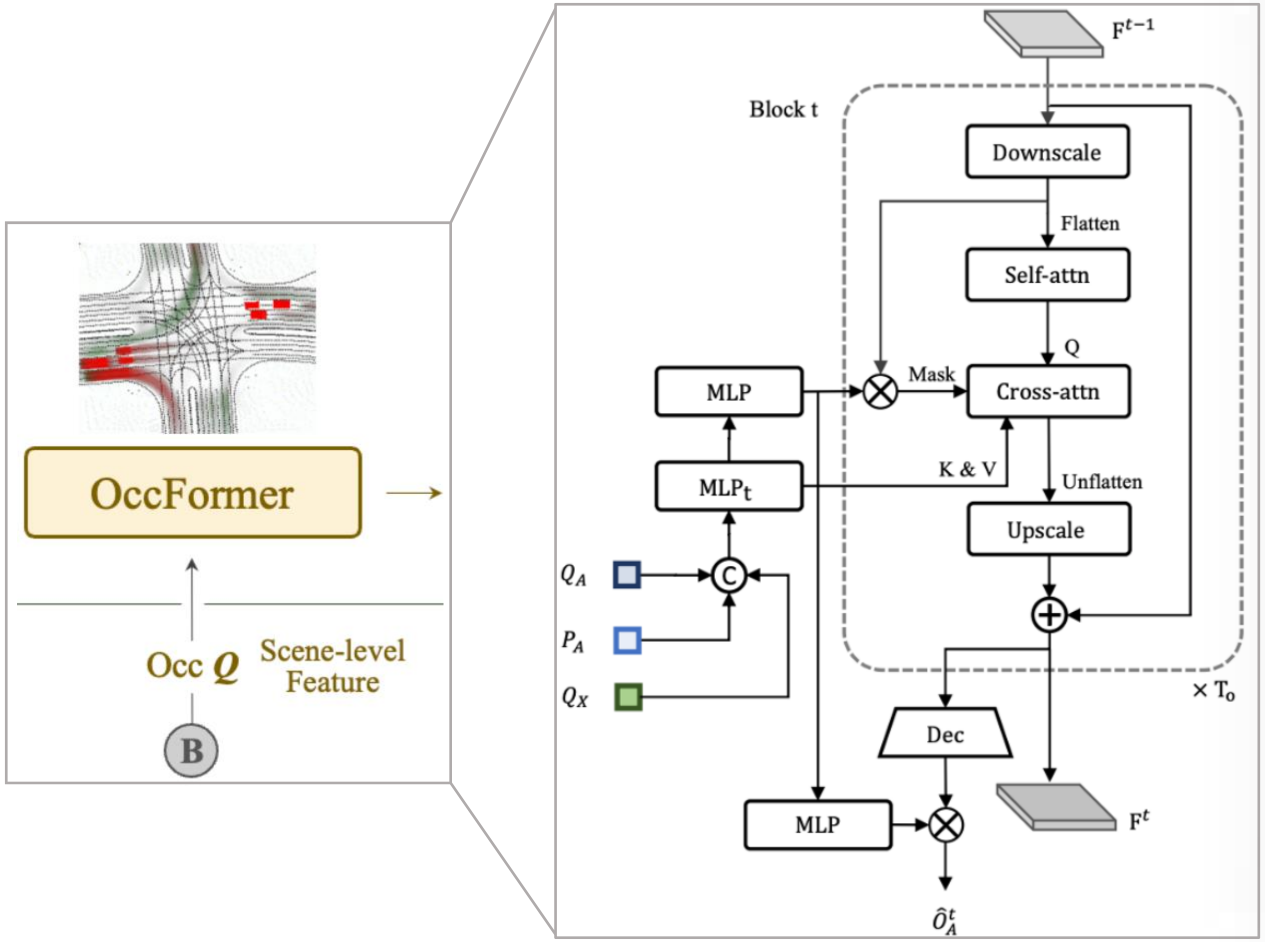
$$\tilde{\mathbf{x}}^* = \arg\min_{\mathbf{x}} c(\mathbf{x}, \tilde{\mathbf{x}})$$

$$c(\mathbf{x}, \tilde{\mathbf{x}}) = \lambda_{\mathrm{xy}} \|\mathbf{x}, \tilde{\mathbf{x}}\|_2 + \lambda_{\mathrm{goal}} \|\mathbf{x}_T, \tilde{\mathbf{x}}_T\|_2 + \sum_{\phi \in \Phi} \phi(\mathbf{x})$$

- $\tilde{\mathbf{x}}$ denotes the ground-truth
- $\tilde{\mathbf{x}}^*$ denotes the smoothed trajectory
- $\mathbf{x}$ is generated by **multiple shooting**
- Φ is the kinematic function set: jerk, curvature, curvature rate, acceleration and lateral acceleration

The UniAD Pipeline

Proposed in **UniAD**



Occupancy prediction

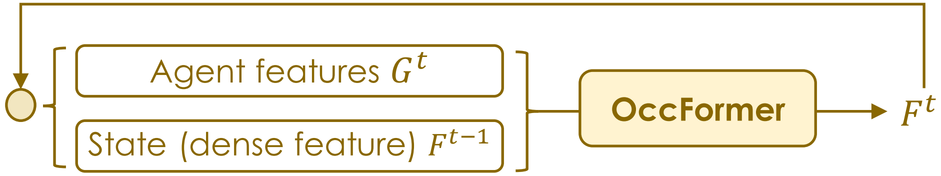
- Predict **occupancy** as attention mask
- Encode **agent-wise knowledge** into the scene representation

Scene-level semantics | **Agent-level semantics**

- Dense scene features ↔ agent-level (**attention module**)
- **Instance-wise occupancy** via matrix multiplication:

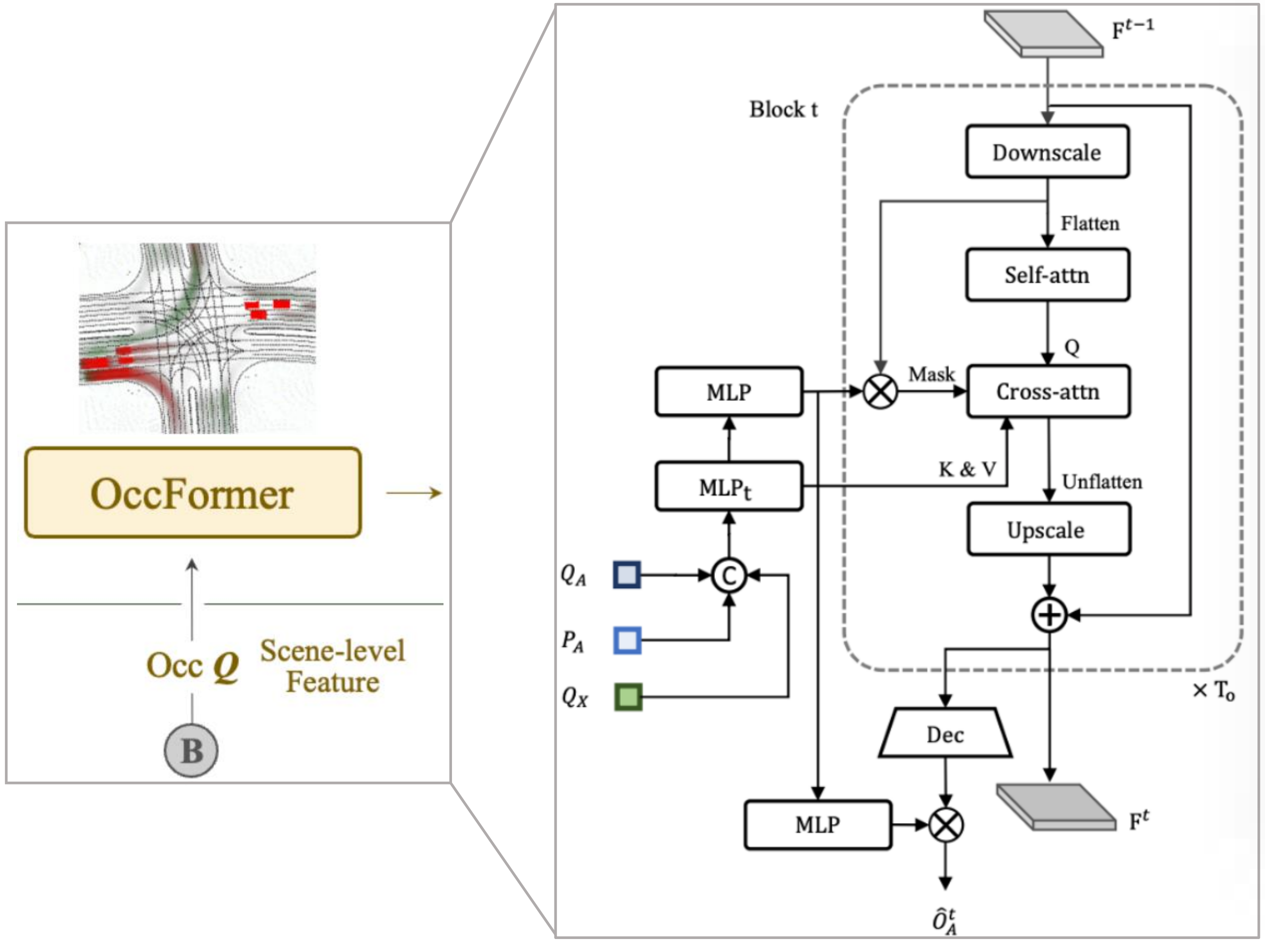
  agent-level features × dense scene-features

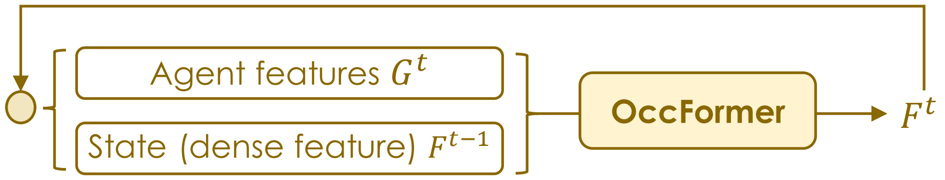
- T_o sequential blocks, with $T_o < T$ from MotionFormer
- High-computational cost



The UniAD Pipeline

Proposed in **UniAD**





Agent features contain **dynamics** and **spatial priors** via max-pooling motion queries from **MotionFormer**:

$$Q_X \in \mathbb{R}^{N_a \times D} \text{ from } Q_{ctx} \in \mathbb{R}^{N_a \times \mathcal{K} \times D}$$

Then, early-fusion via a temporal-specific MLP:

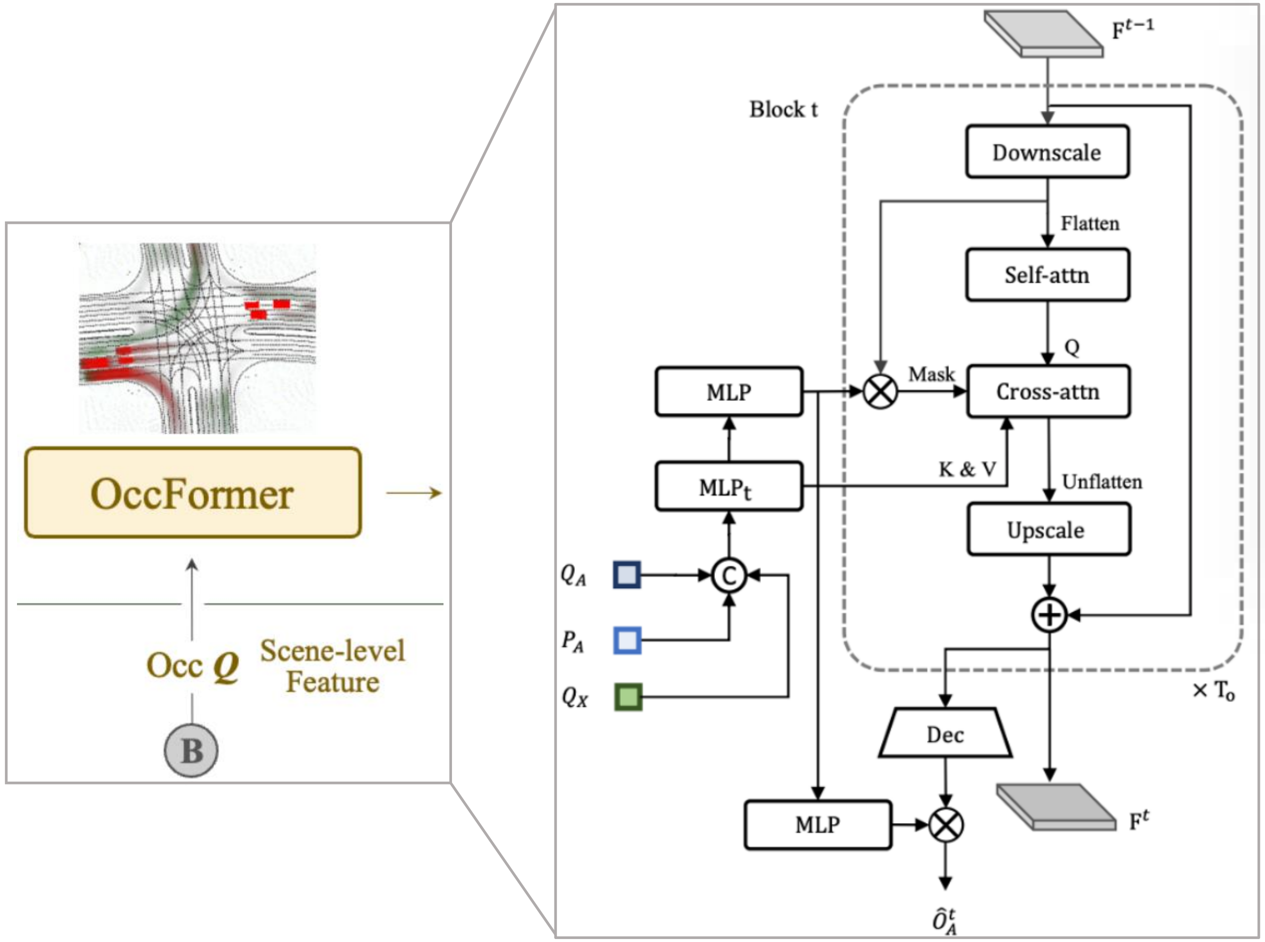
$$G^t = \mathrm{MLP_t}([Q_A, P_A, Q_X]), \qquad t = 1, \dots, T_o$$

- Q_A: upstream track query
- P_A: current position embedding
- $[\cdot]$: concatenation

State (dense feature) F^0 is the BEV feature ¼ downscaled

The UniAD Pipeline

Proposed in **UniAD**



Pixel-level interaction Scene ↔ Agent-level understanding

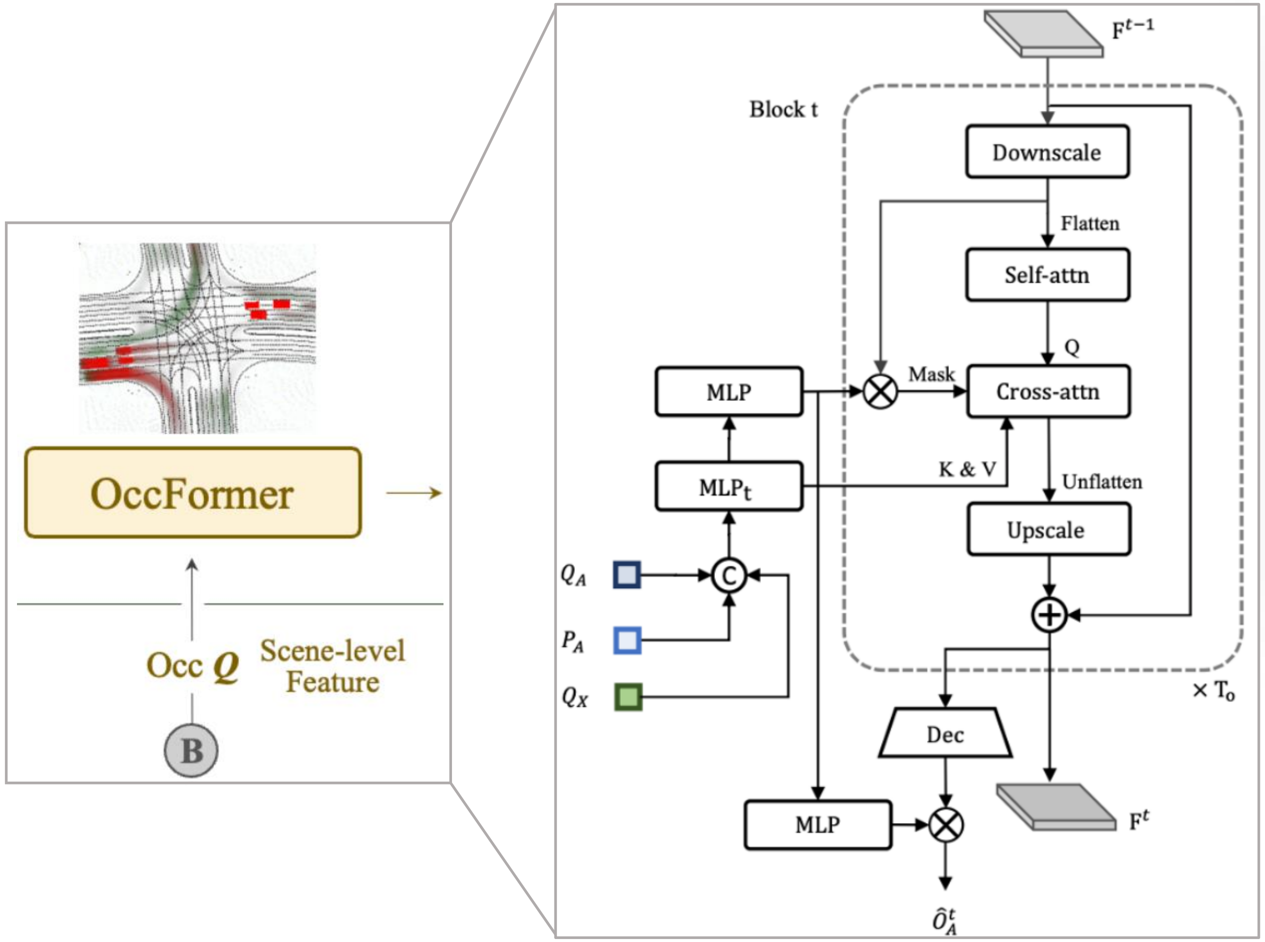
- Dense feature F_{ds}^{t} as queries
- Instance-level features as keys and values to update the dense feature over time

$$D_{\mathrm{ds}}^{t} = \mathrm{MHCA}\big(\mathrm{MHSA}(F_{\mathrm{ds}}^{t}), G^{t}, \mathrm{att\_mask} = O_{m}^{t}\big)$$

- O_{m}^{t} is obtained by multiplying $M^{t} = \mathrm{MLP}(G^{t})$ with F_{ds}^{t}
- D_{ds}^{t} is **upsampled** to ¼ the size of B and add F^{t-1} as residual connection

The UniAD Pipeline

Proposed in **UniAD**



Instance-level occupancy

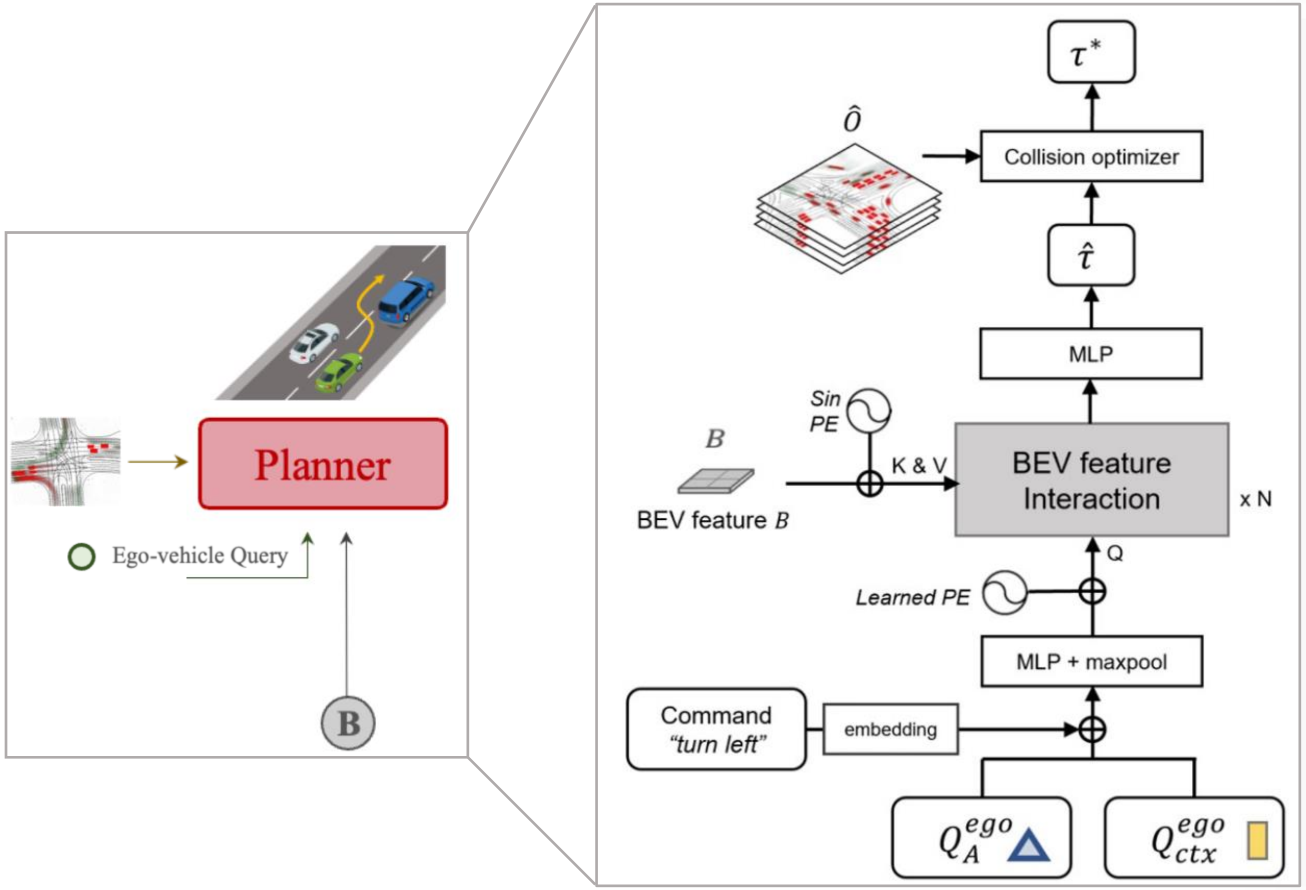
- It represents the occupancy with each agent's identity preserved
- Original size $H \times W$ of BEV feature B
- F^t is **upsampled** to $F^t_{\text{dec}} \in \mathbb{R}^{C \times H \times W}$
- Update the coarse mask again, $U^t = \text{MLP}(M^t) \in \mathbb{R}^{N_a \times C}$
- Finally, the instance-level occupancy at time t is

$$\tilde{O}_A^t = U^t \cdot F^t_{\text{dec}}$$

Perception ··· Prediction ··· Planning

The UniAD Pipeline

Proposed in **UniAD**



Planning

- **Ego-vehicle query** consistently passed through **TrackFormer**, **MotionFormer**, and now **Planner**.

turn left
turn right
forward

Ego-vehicle Query
\+
3 learnable embeddings
(command embeddings)

"plan query"
attend
BEV Features B

$\hat{\tau}$

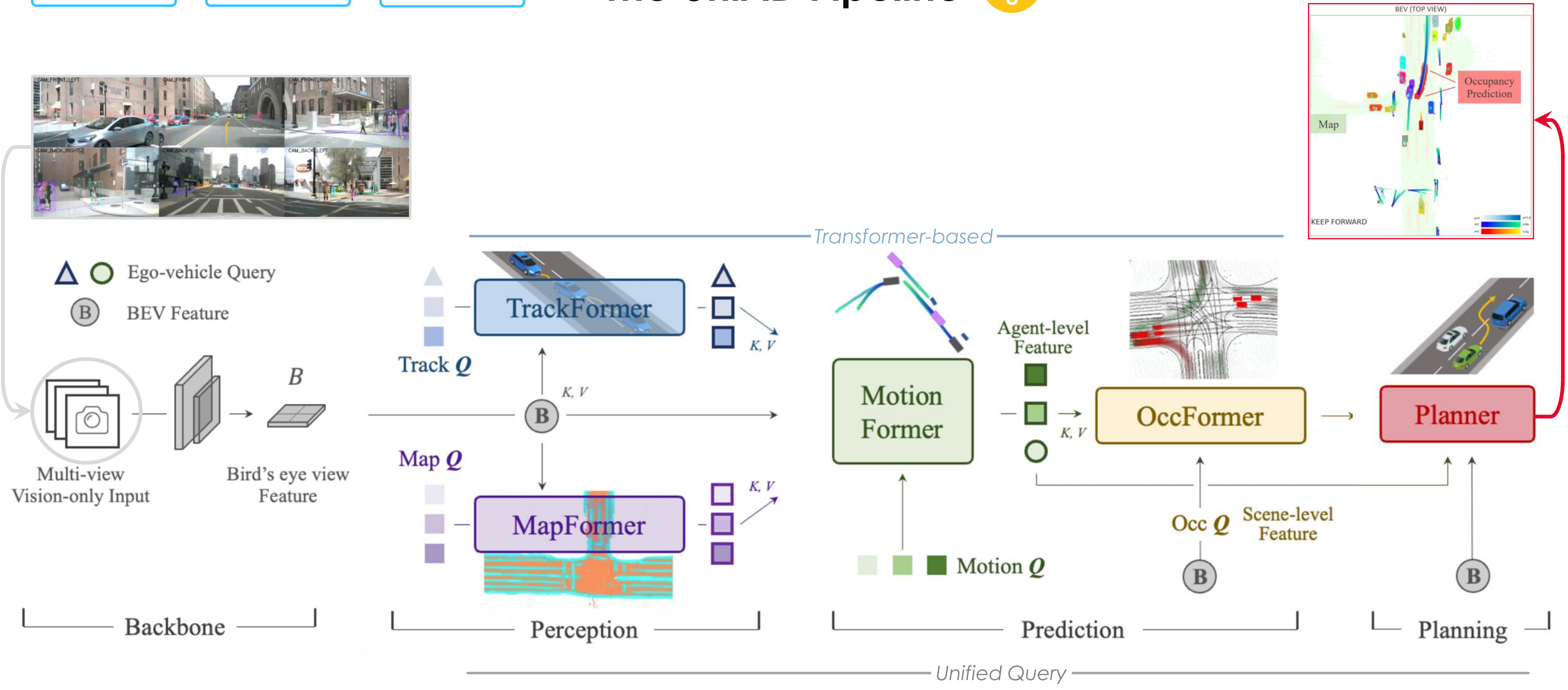
- Avoid **collisions** through **Newton's method** in inference

$$\tau^* = \arg\min_{\tau} f(\tau, \hat{\tau}, \hat{O})$$

- $\hat{\tau}$: original planning prediction
- τ^*: optimized planning
- τ: generated by **multiple shooting**
- $f(\cdot)$: cost function to minimize
- $\hat{O}$: binary occupancy map from **OccFormer**

Perception ··· Prediction ··· Planning

The UniAD Pipeline



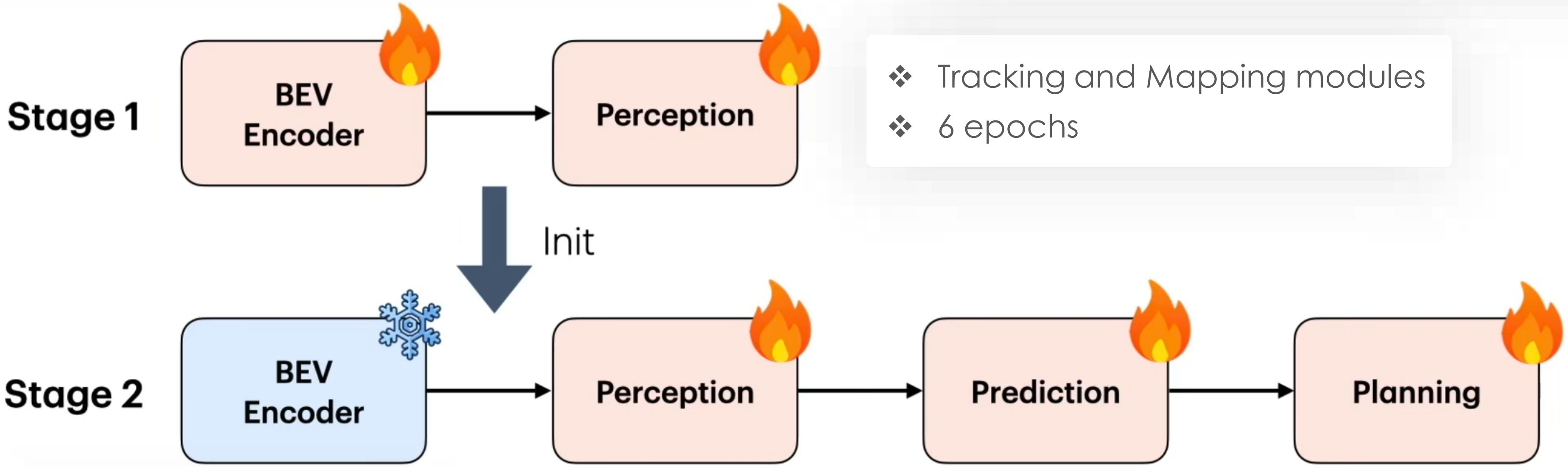
How to train it?

Two-phase training

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

Shared matching (Bipartite Matching Algorithm)

- Inspired by **DETR** (DEtection TRansformer), which matches predictions to ground truth entities minimizing a cost function
- Consistent learning of agent identities
- Converges faster



- Tracking and Mapping modules
- 6 epochs

- End-to-end training
- 20 epochs

From: https://www.youtube.com/watch?v=GdEVfcLYpPc

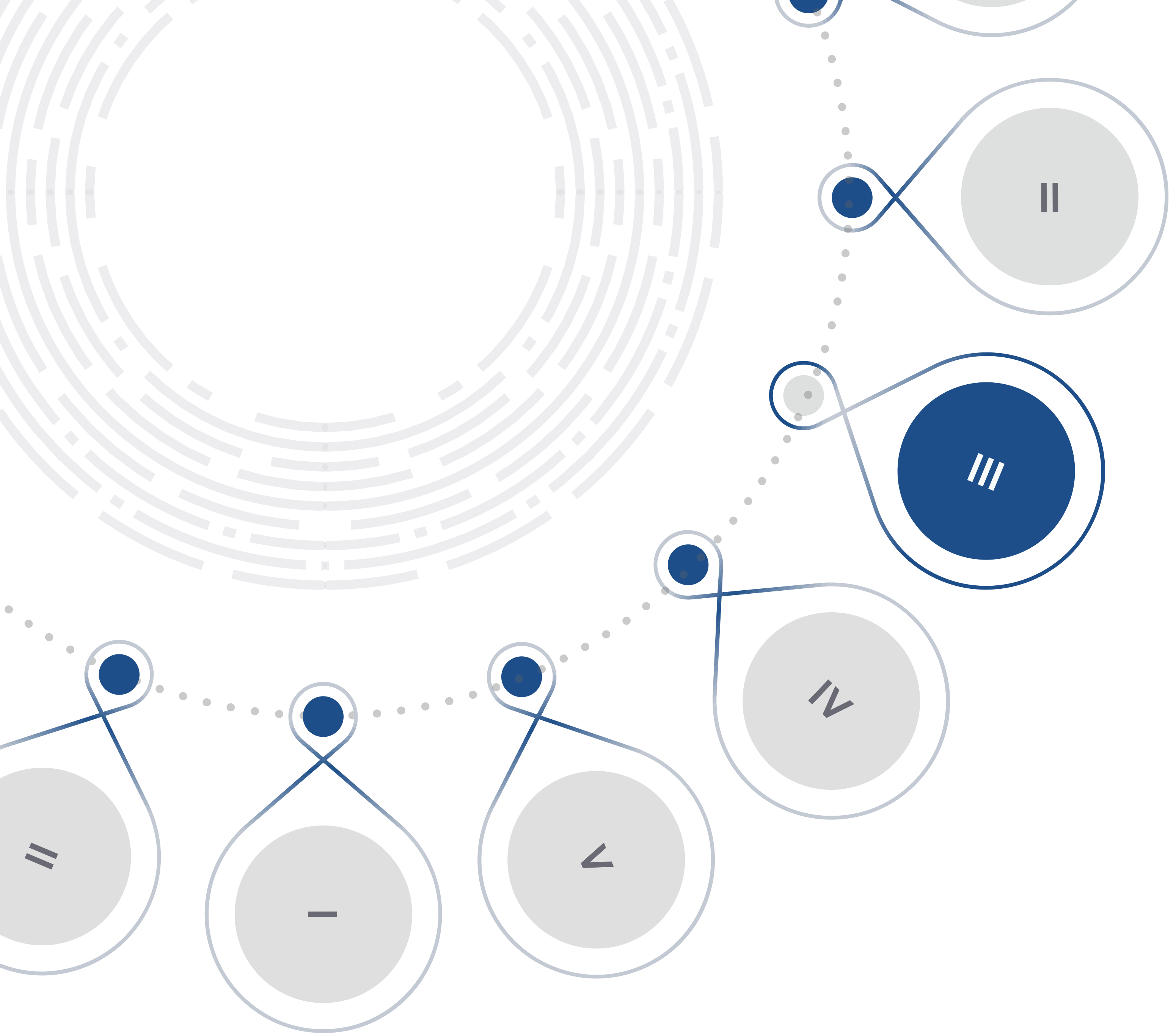


Table of Contents

Ablation Results

Evaluation on **nuScenes**

ID	Modules: Track	Map	Motion	Occ.	Plan	Tracking: AMOTA↑	AMOTP↓	IDS↓	Mapping: IoU-lane↑	IoU-road↑	Motion Forecasting: minADE↓	minFDE↓	MR↓	Occupancy Prediction: IoU-n.↑	IoU-f.↑	VPQ-n.↑	VPQ-f.↑	Planning: avg.L2↓	avg.Col.↓
0*	✓	✓	✓	✓	✓	0.356	1.328	893	0.302	0.675	0.858	1.270	0.186	55.9	34.6	47.8	26.4	1.154	0.941
1	✓					0.348	1.333	791	-	-	-	-	-	-	-	-	-	-	-
2		✓				-	-	-	**0.305**	0.674	-	-	-	-	-	-	-	-	-
3	✓	✓				0.355	1.336	785	0.301	0.671	-	-	-	-	-	-	-	-	-
4			✓			-	-	-	-	-	0.815	1.224	0.182	-	-	-	-	-	-
5	✓		✓			0.360	1.350	919	-	-	0.751	1.109	0.162	-	-	-	-	-	-
6	✓	✓	✓			0.354	1.339	820	0.303	0.672	0.736(-9.7%)	1.066(-12.9%)	0.158	-	-	-	-	-	-
7				✓		-	-	-	-	-	-	-	-	60.5	37.0	52.4	29.8	-	-
8	✓			✓		0.360	**1.322**	809	-	-	-	-	-	62.1	38.4	52.2	32.1	-	-
9	✓	✓	✓	✓		0.359	1.359	1057	0.304	**0.675**	**0.710**(-3.5%)	**1.005**(-5.8%)	**0.146**	**62.3**	39.4	**53.1**	32.2	-	-
10					✓	-	-	-	-	-	-	-	-	-	-	-	-	1.131	0.773
11	✓	✓	✓		✓	**0.366**	1.337	889	0.303	0.672	0.741	1.077	0.157	-	-	-	-	1.014	0.717
12	✓	✓	✓	✓	✓	0.358	1.334	**641**	0.302	0.672	0.728	1.054	0.154	**62.3**	**39.5**	52.8	**32.3**	**1.004**	**0.430**

01

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

02

- ✓ ID. 4-6: Track & Map → Motion
- ✓ ID. 7-9: Motion ↔ Occupancy
- ✓ ID. 10-12: Motion & Occupancy → Planning

03

The two perception sub-tasks greatly help **motion forecasting**, and **prediction performance**

04

UniAD outperforms naïve MTL solution

Ablation Results

Design of the **MotionFormer**

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**

Non-Linear Optimization

Ego-vehicle Query

Agent-goal interaction

Rotated scene-level anchors

All components contribute to the ultimate performance

Ablation Results

Design of the **OccFormer**

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**

Re-use of Mask feature (Instance-level occupancy)

Cross- and Attention mask (Pixel-level interaction)

- Cross-attention with masks and the reuse of mask feature helps improve the prediction
- Not all metrics show **best performance**…

Ablation Results

Design of the **Planner**

ID	BEV Att.	Col. Loss	Occ. Optim.	L2↓ 1s	L2↓ 2s	L2↓ 3s	Col. Rate↓ 1s	Col. Rate↓ 2s	Col. Rate↓ 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**

Optimization strategy with Occupancy

Collision Loss

Attending to BEV feature

- Results demonstrate the necessity of each preceding task
- Same as in **OccFormer**, not all metrics improve

Results: Planning

	Method	L2(m)↓ 1s	2s	3s	Avg.	Col. Rate(%)↓ 1s	2s	3s	Avg.
LiDAR-based	NMP† [57]	-	-	2.31	-	-	-	1.92	-
	SA-NMP† [57]	-	-	2.05	-	-	-	1.59	-
	FF† [22]	0.55	1.20	2.54	1.43	0.06	0.17	1.07	0.43
	EO† [26]	0.67	1.36	2.78	1.60	0.04	0.09	0.88	0.33
camera-based	ST-P3 [23]	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**

01 UniAD achieves the lowest L2 error and collision rate in all time intervals

02 UniAD outperforms **LiDAR-based** methods in all investigated metrics

Results: Other tasks

Multi-Object Tracking

Method	AMOTA↑	AMOTP↓	Recall↑	IDS↓
Immortal Tracker† [52]	0.378	1.119	0.478	936
ViP3D [18]	0.217	1.625	0.363	-
QD3DT [21]	0.242	1.518	0.399	-
MUTR3D [58]	0.294	1.498	0.427	3822
UniAD	**0.359**	**1.320**	**0.467**	**906**

Motion Forecasting

Method	minADE(m)↓	minFDE(m)↓	MR↓	EPA↑
PnPNet† [32]	1.15	1.95	0.226	0.222
ViP3D [18]	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**

Online Mapping

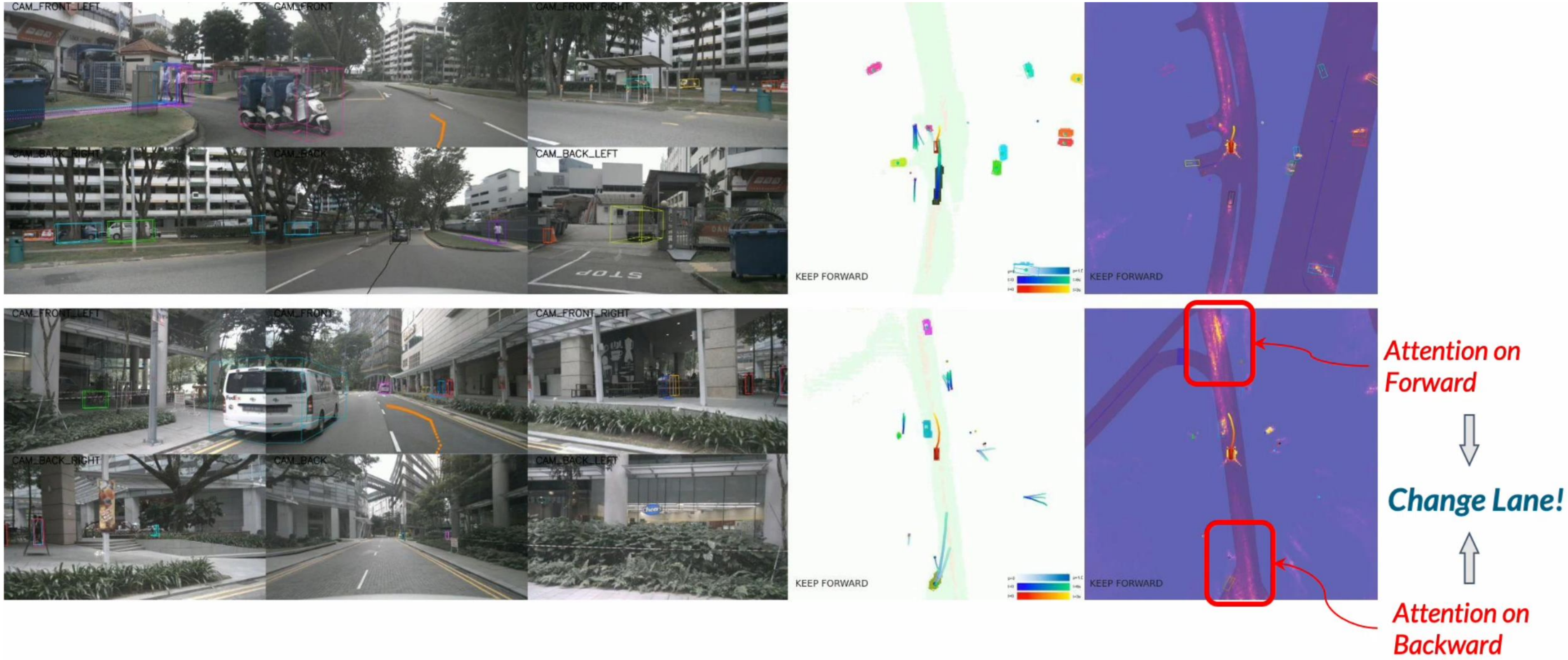
Method	Lanes↑	Drivable↑	Divider↑	Crossing↑
VPN [42]	18.0	76.0	-	-
LSS [44]	18.3	73.9	-	-
BEVFormer [30]	23.9	**77.5**	-	-
BEVerse† [59]	-	-	**30.6**	**17.2**
UniAD	**31.3**	69.1	25.7	13.8

Occupancy Prediction

Method	IoU-n.↑	IoU-f.↑	VPQ-n.↑	VPQ-f.↑
FIERY [20]	59.4	36.7	50.2	29.9
StretchBEV [1]	55.5	37.1	46.0	29.0
ST-P3 [23]	-	38.9	-	32.1
BEVerse† [59]	61.4	**40.9**	54.3	**36.1**
UniAD	**63.4**	40.2	**54.7**	33.5

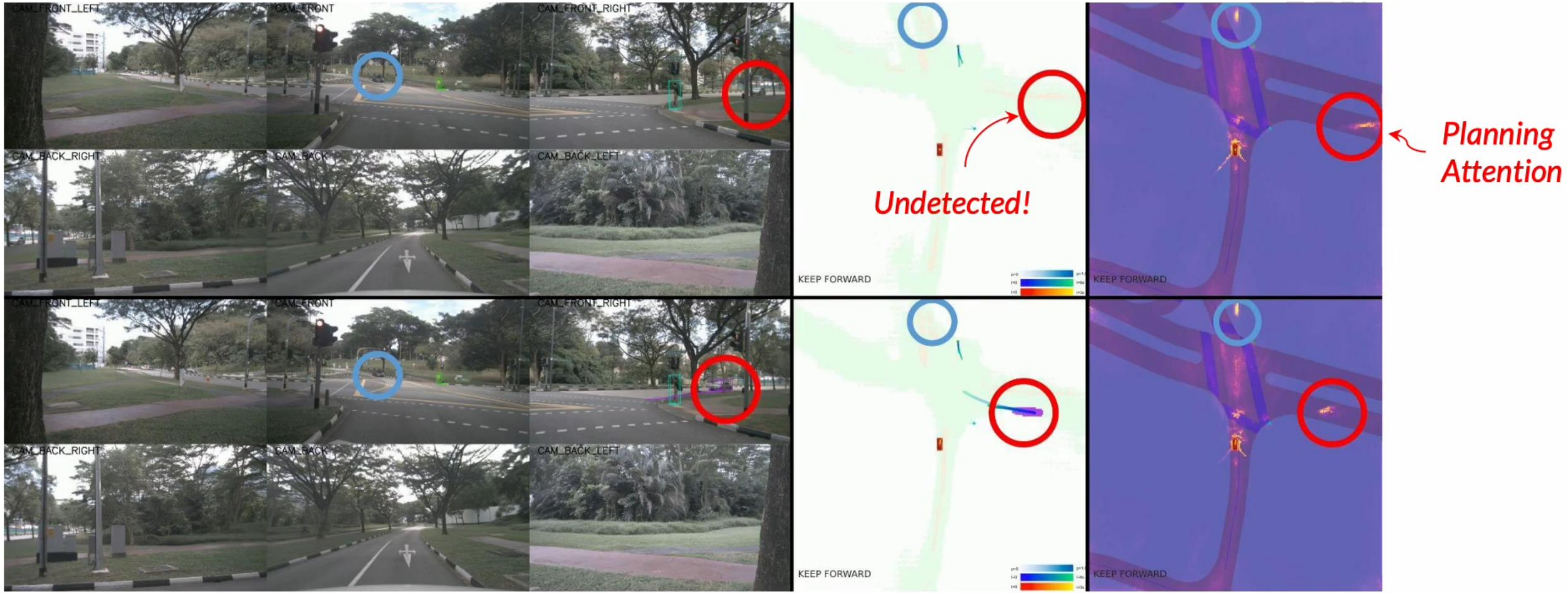
- **SOTA** performance on almost all investigated tasks

Qualitative Results



Obstacles avoidance visualization
The ego vehicle is changing lanes attentively to avoid an obstacle.

Qualitative Results



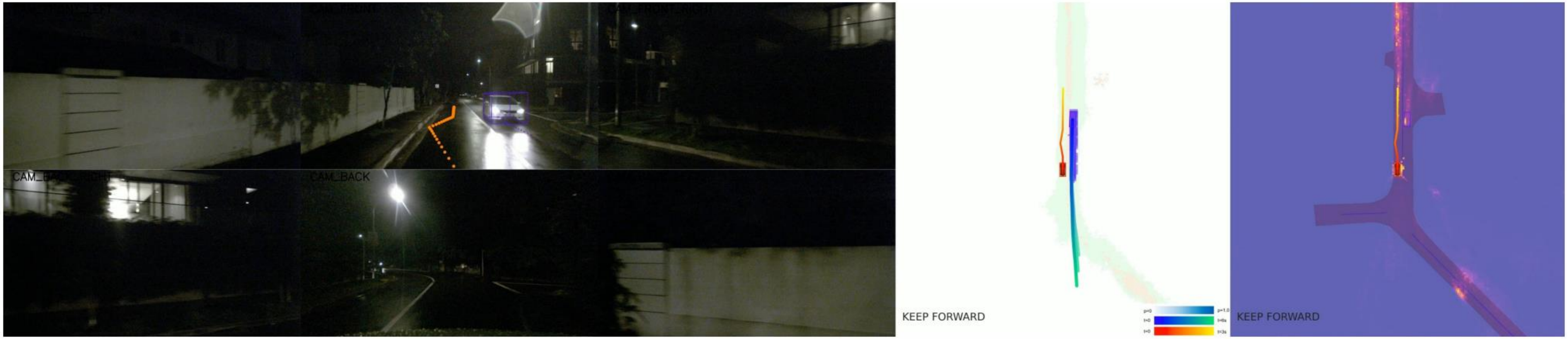
Visualization for planning recovering from perception failures
Inaccurate results occur in prior modules while the later tasks could still recover.

Table of Contents

Limitations

Long-tail scenario
A large trailer with a white container occupies the entire road.

Limitations



Over-cautious planner
Incoming vehicle in a narrow street in a dark environment.

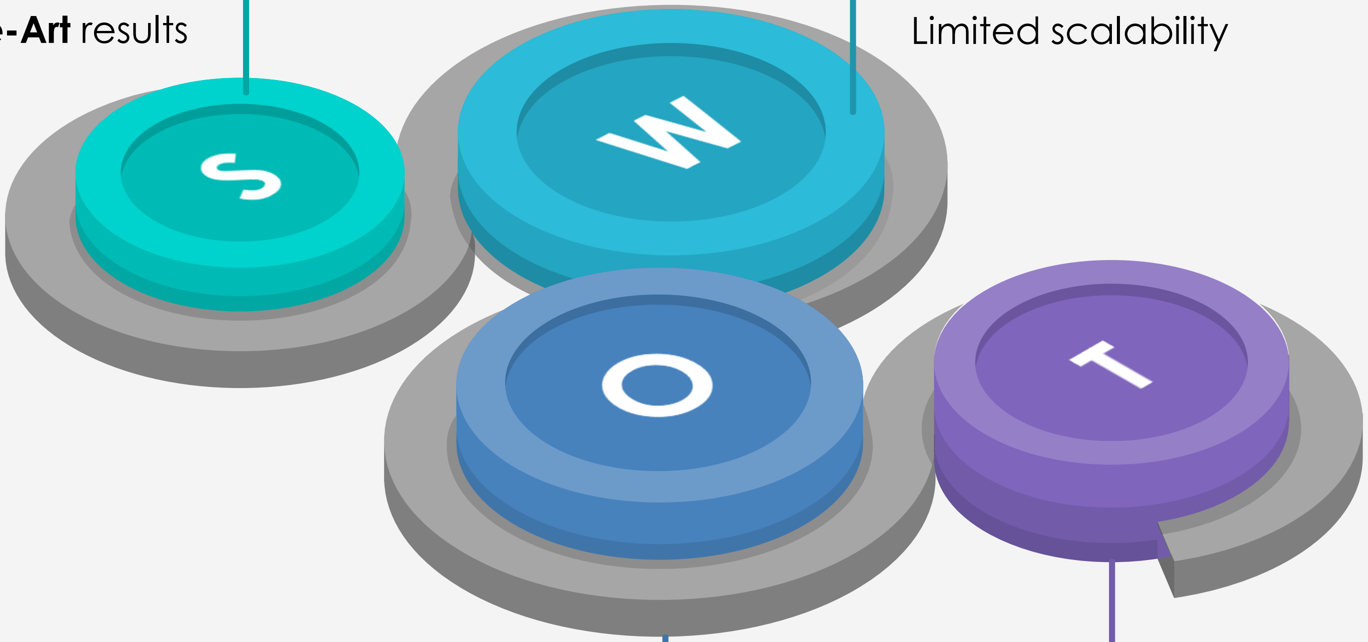
Limitations: SWOT Analysis

STRENGTH

Unified end-to-end framework
Planning-oriented design
Query-based task coordination
State-of-the-Art results

WEAKNESS

Extensive **computational power**
Temporal history training
Limited scalability



OPPORTUNITY

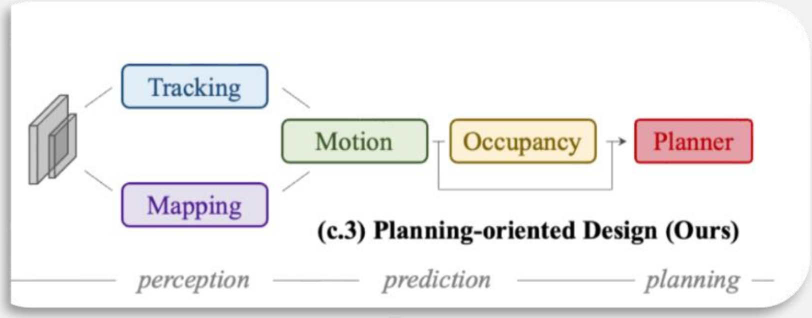
Curation for a **lightweight deployment**
Integration & embedding of new tasks
(depth estimation, behaviour prediction)

THREAT

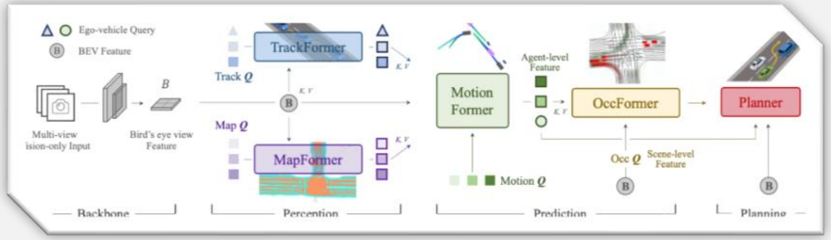
Performance in **rare/dark scenarios**
Simpler/more lightweight methods
LiDAR-based methods

Table of Contents

Conclusions – Authors’ View



UniAD



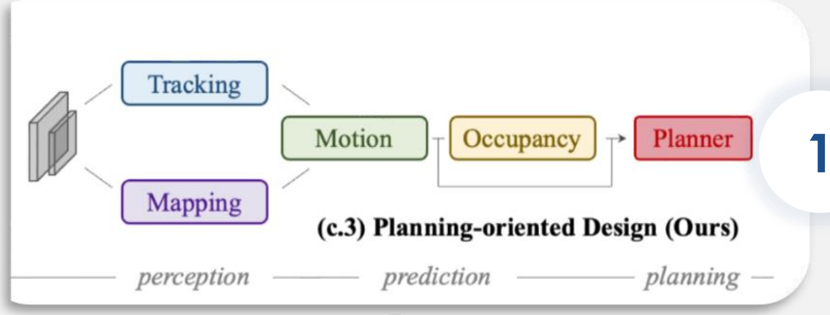
Integrates ***perception***, ***prediction***, and ***planning*** together, reducing error propagation compared to modular approaches

Unified Query design: Queries as interfaces to connect and coordinate all tasks in an interpretable fashion

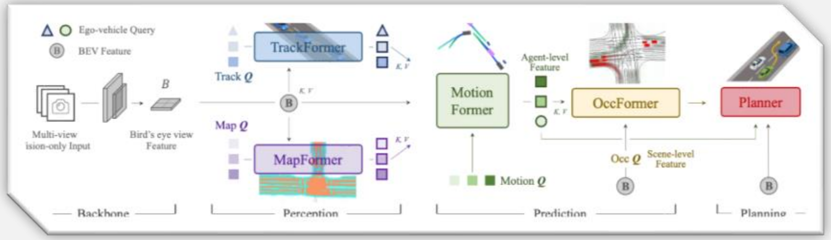
State-of-the-art (SOTA) Performance with vision-only input

First Step towards **Autonomous Driving Foundational Models**

Conclusions – Authors' View



UniAD



1

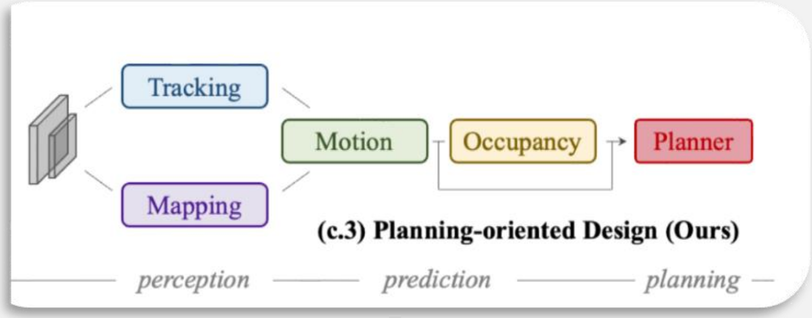
Integrates ***perception***, ***prediction***, and ***planning*** together, reducing error propagation compared to modular approaches

Unified Query design: Queries as interfaces to connect and coordinate all tasks in an interpretable fashion

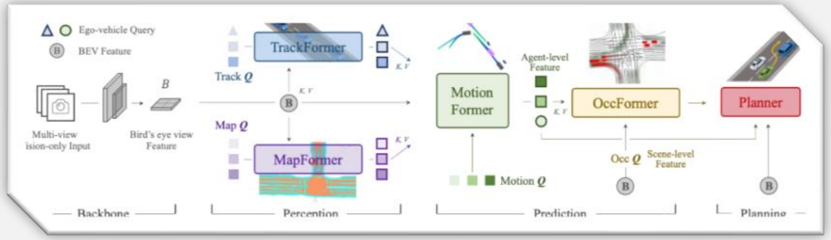
State-of-the-art (SOTA) Performance with vision-only input

First Step towards **Autonomous Driving Foundational Models**

Conclusions – Authors' View



UniAD



Integrates ***perception***, ***prediction***, and ***planning*** together, reducing error propagation compared to modular approaches

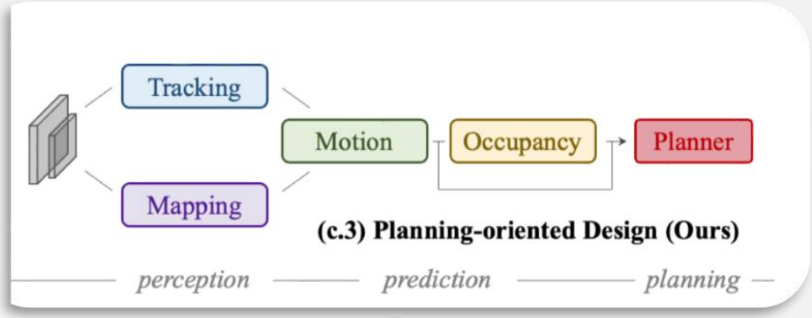
2

Unified Query design: Queries as interfaces to connect and coordinate all tasks in an interpretable fashion

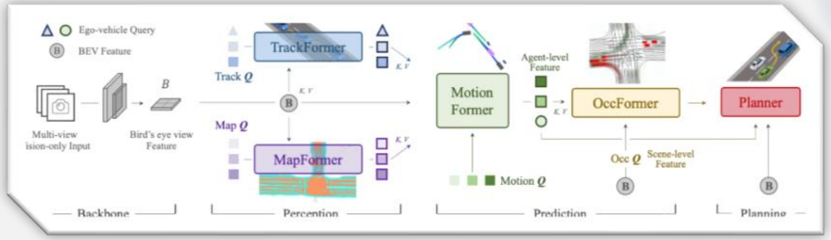
State-of-the-art (SOTA) Performance with vision-only input

First Step towards **Autonomous Driving Foundational Models**

Conclusions – Authors' View



UniAD



Integrates ***perception***, ***prediction***, and ***planning*** together, reducing error propagation compared to modular approaches

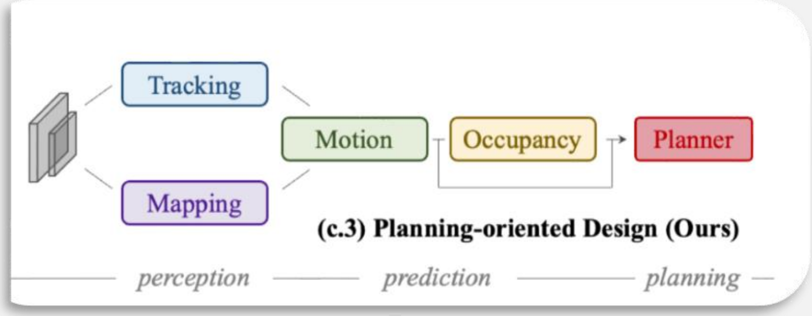
Unified Query design: Queries as interfaces to connect and coordinate all tasks in an interpretable fashion

3

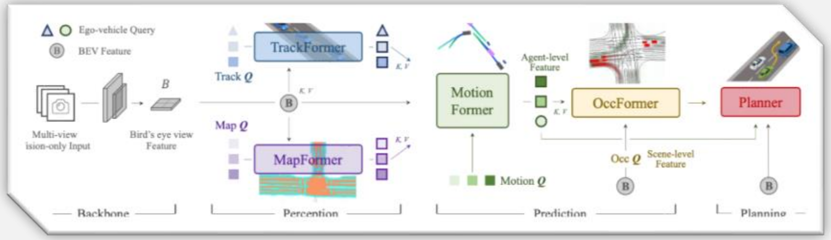
State-of-the-art (SOTA) Performance with vision-only input

First Step towards **Autonomous Driving Foundational Models**

Conclusions – Authors' View



UniAD



Integrates ***perception***, ***prediction***, and ***planning*** together, reducing error propagation compared to modular approaches

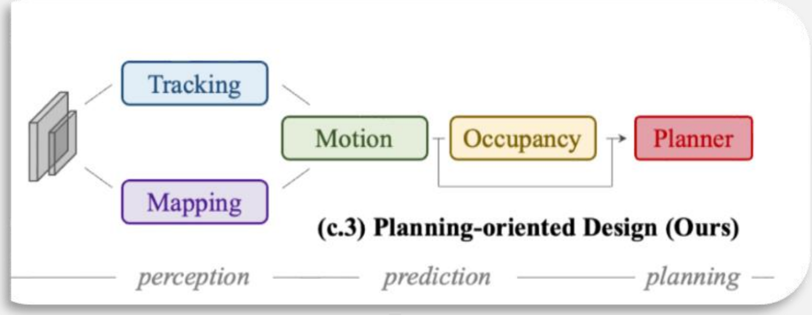
Unified Query design: Queries as interfaces to connect and coordinate all tasks in an interpretable fashion

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

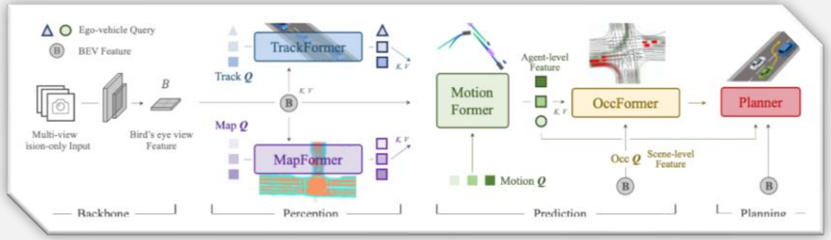
4

First Step towards **Autonomous Driving Foundational Models**

Conclusions – My View



UniAD



Some **Ablation Results** are not conclusive, i.e., not all metrics are improving w.r.t features

Unified Query design: Queries as interfaces to connect and coordinate all tasks ~~in an interpretable fashion~~

Some Design choices seem too **abstract**

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

Better comparison against **LiDAR** needed

~~First Step towards~~ ~~**Autonomous Driving Foundational Models**~~

Much to learn we still have

Thank you for your attention

Time for questions!

References

Carion, N., F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. End-to-End Object Detection with Transformers, 2020. arXiv: 2005.12872 [cs.CV]. https://arxiv.org/abs/2005.12872.

Chen, L., P. Wu, K. Chitta, B. Jaeger, A. Geiger, and H. Li. "End-to-end Autonomous Driving: Challenges and Frontiers." IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024.

Hu, Y. [CVPR2023 Award Candidate] UniAD: Planning-oriented Autonomous Driving. YouTube Video. Candidate Award Talk, available at https://www.youtube.com/watch?v=R4iuq3zDBL4, 2023.

Hu, Y. CVPR23 Plenary Talk — [Best Paper] UniAD: Planning-oriented Autonomous Driving. YouTube Video. Plenary talk at CVPR 2023, available at https://www.youtube.com/watch?v=cyrxJJ_nnaQ, 2023.

Hu, Y., J. Yang, L. Chen, K. Li, C. Sima, X. Zhu, S. Chai, et al. Planning-oriented Autonomous Driving, 2023. arXiv: 2212.10156 [cs.CV]. https://arxiv.org/abs/2212.10156.

Li, Z., W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Q. Yu, and J. Dai. BEVFormer: Learning Bird's-Eye-View Representation from Multi-Camera Images via Spatiotemporal Transformers, 2022. arXiv: 2203.17270 [cs.CV]. https://arxiv.org/abs/2203.17270.

Li, Z., W. Wang, E. Xie, Z. Yu, A. Anandkumar, J. M. Alvarez, P. Luo, and T. Lu. Panoptic SegFormer: Delving Deeper into Panoptic Segmentation with Transformers, 2022. arXiv: 2109.03814 [cs.CV]. https://arxiv.org/abs/2109.03814.

Zeng, F., B. Dong, Y. Zhang, T. Wang, X. Zhang, and Y. Wei. MOTR: End-to-End Multiple-Object Tracking with Transformer, 2022. arXiv: 2105.03247 [cs.CV]. https://arxiv.org/abs/2105.03247.

All illustrations and tables taken from the original paper: Planning-oriented Autonomous Driving (including Supplementary Material) by Hu, Y., J. Yang, L. Chen, K. Li, C. Sima, X. Zhu, S. Chai, et al., unless specified otherwise. The references listed on the tables have the enumeration from the original paper.