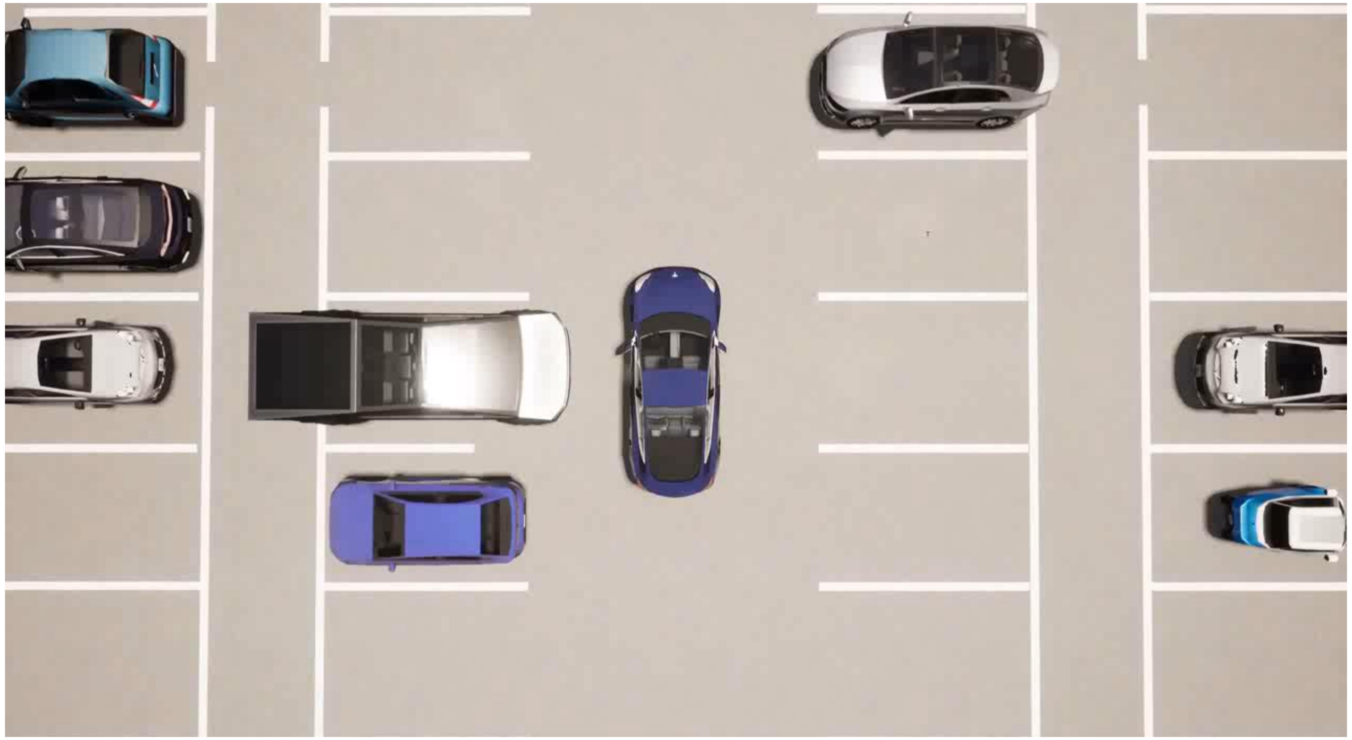
E2E Parking: Autonomous Parking by the End-to-End Neural Network on the CARLA Simulator

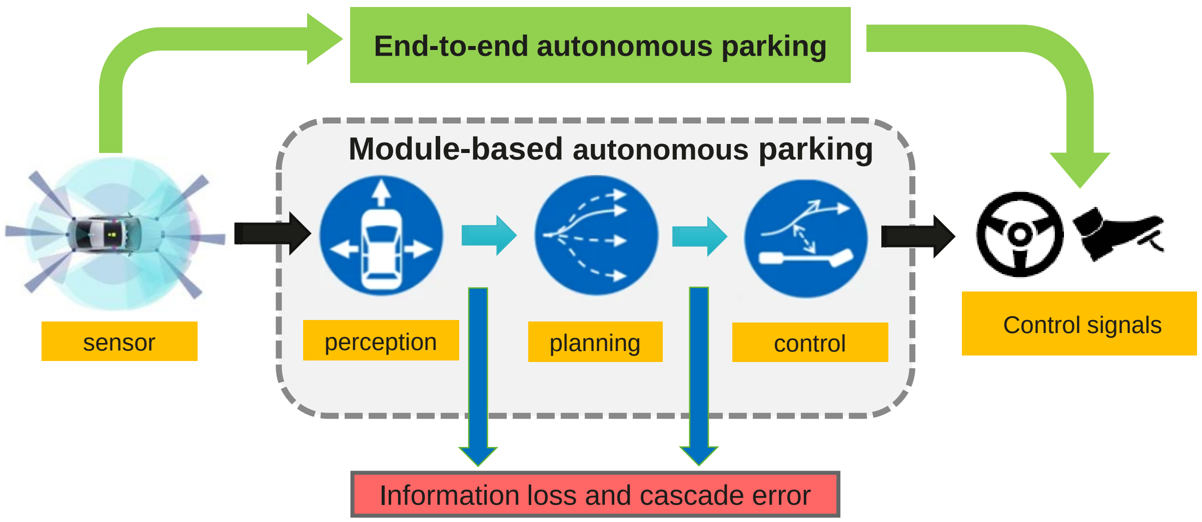
Yunfan Yang, Denglong Chen, Tong Qin, Xiangru Mu, Chunjing Xu, and Ming Yang

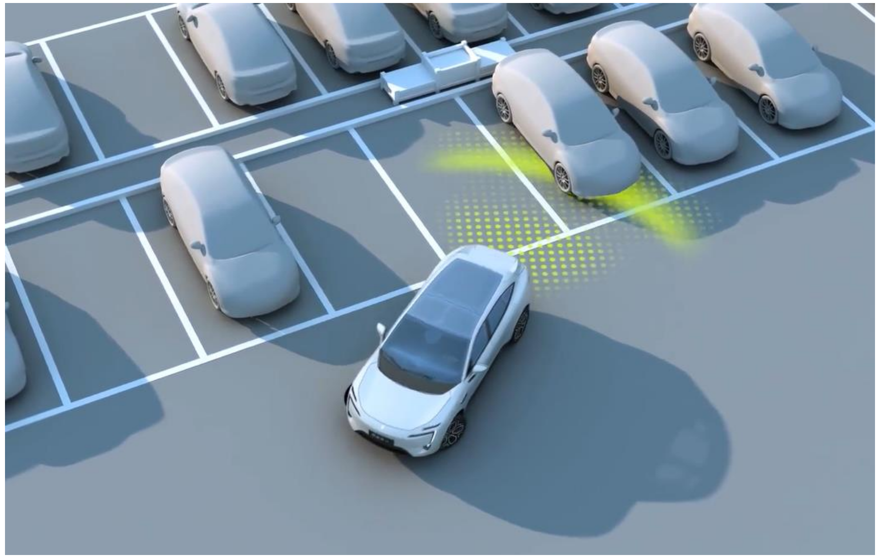


BACKGROUND

- Limited flexibility and robustness in traditional Automated Parking Assist (APA) due to accumulated uncertainty from the rule-based multi-stage pipeline

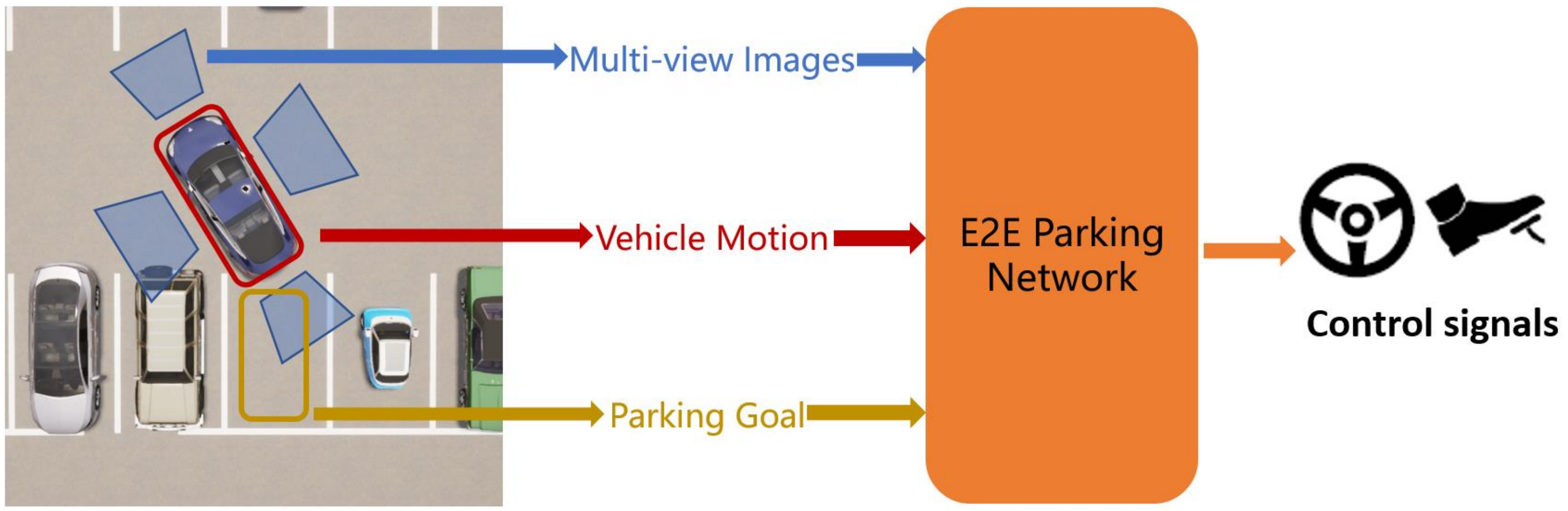
- End-to-end systems offer the potential to simplify the overall pipeline, enhance adaptability, and reduce reliance on handcrafted features and rules





MOTIVATION

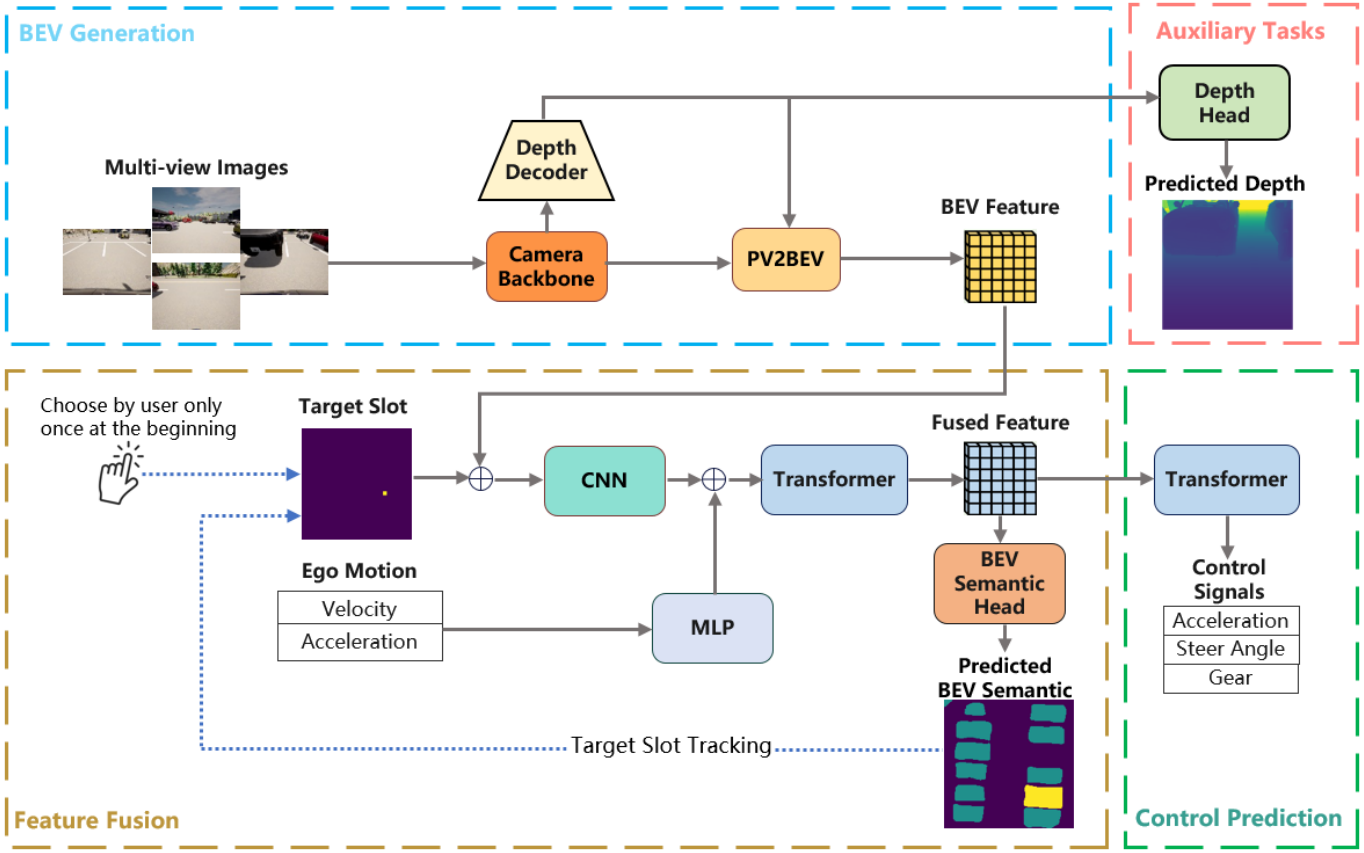
- To design an end-to-end APA framework that converts sensor data directly to the chassis control signals
- To make full use of the attention mechanism inspired by the exciting performance of transformer applied in the field of NLP
- To build a set of quantitative metrics and establish the benchmark in autonomous parking



METHOD (Overview)

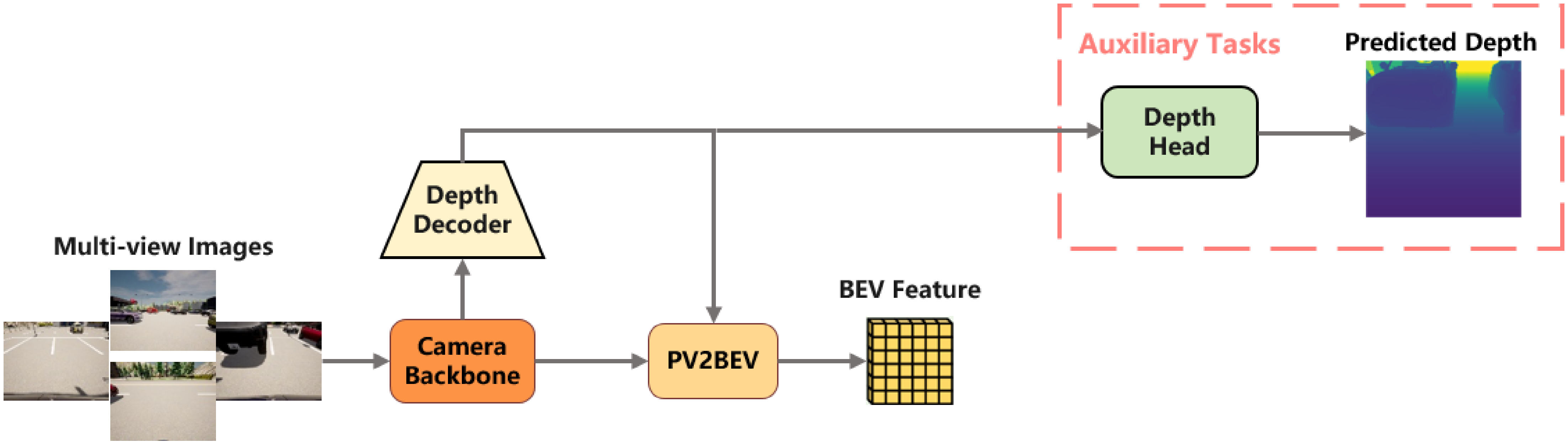
- The framework of the proposed approach comprises 4 main parts:
  > BEV Generation, Feature Fusion, Control Prediction, and Auxiliary Tasks.

- Inspired by transformer in translation task, we use the cross-attention mechanism to translate the fused feature to vehicle control signals.



METHOD (BEV Generation)

- We adopt LSS[1] with explicit depth supervision to obtain the BEV feature from surrounding images.
- 4 onboard cameras on front, left, right, and rear



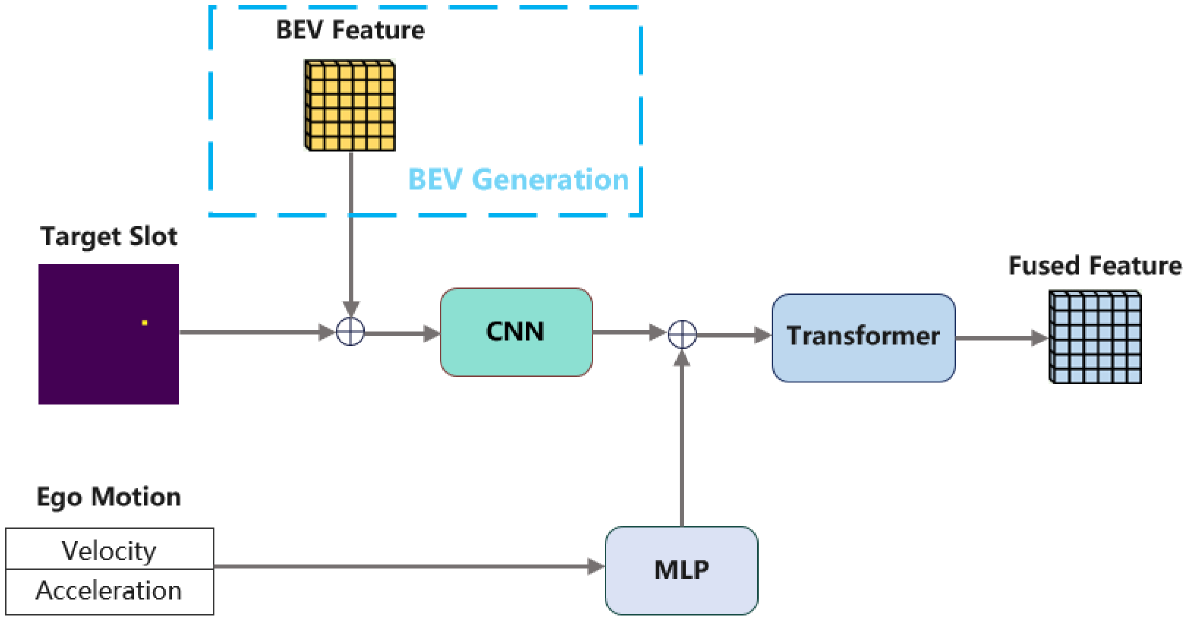
[1] Philion, J., & Fidler, S. (2020). Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *Computer Vision–ECCV 2020, Proceedings, Part XIV 16* (pp. 194-210).

METHOD (Feature Fusion)

- We add an extra channel, which draws the position of the target slot relative to the BEV grid as a point, to the BEV feature map

- Motion feature is also concatenated to the BEV feature map

- Concatenated feature is fused via self-attention

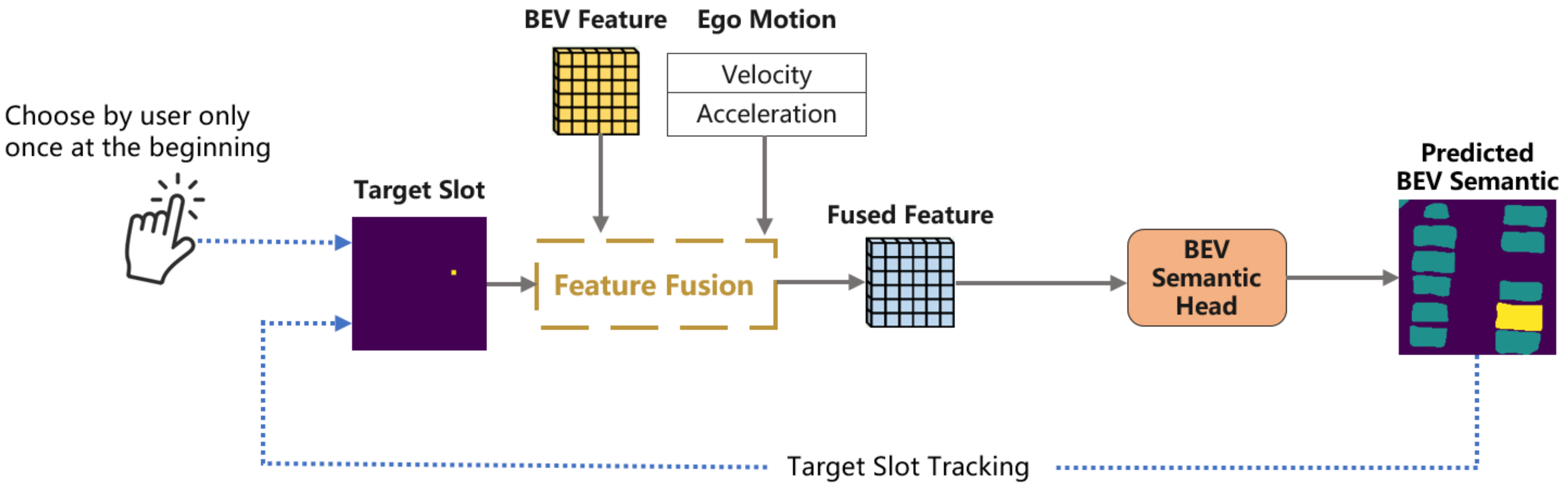


METHOD (BEV Semantic)

- BEV semantic has 3 categories: target slot, static vehicle, and background

- Target Slot Restore:
  - > Input: a point at BEV grid with noise
  - > Output: the whole parking slot

- Target Slot Tracking:
  - > The first timestamp: target position chosen by user
  - > The following timestamps: predicted target position from the previous timestamp

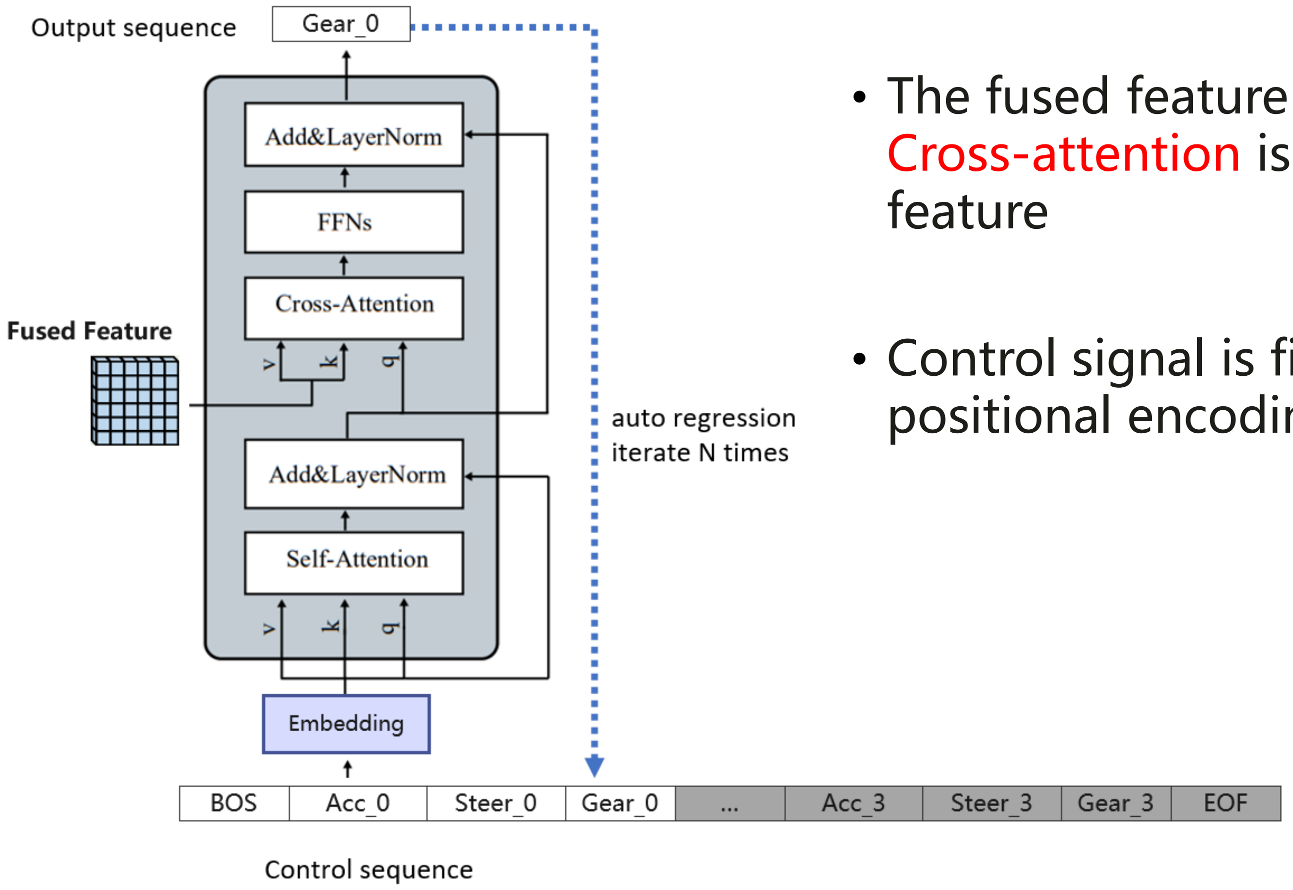


METHOD (Control Prediction)

- A language-modeling style transformer decoder is used to predict the control signal sequence in an auto-regressive manner

- The fused feature serves as the "memory" to the decoder. Cross-attention is taken between control sequence and the fused feature

- Control signal is first tokenized and then embedded with positional encoding



Experiment (Closed-loop evaluation in CARLA)

- Our method proves its accuracy and efficiency in CARLA closed-loop experiment
- With an overall success rate over 90%, our method reaches 0.3 meters for average positional error and 0.87 degrees for orientation (yaw angle) error
- We compare our method to the expert we learn from and a rookie driver. The result demonstrates that our method has surpassed rookie drivers in the parking task

Agent	TSR(%)↑	TFR(%)↓	CR(%)↓	APD(m)↓	AOD(deg)↓	APT(s)↓
Ours	91.41	2.08	2.08	0.30	0.87	15.72
Expert	100.00	0.00	0.00	0.23	0.48	14.96
Rookie	75.00	18.75	6.25	0.35	4.00	20.125

TSR: Target Success Rate
TFR: Target Fail Rate
CR: Collision Rate

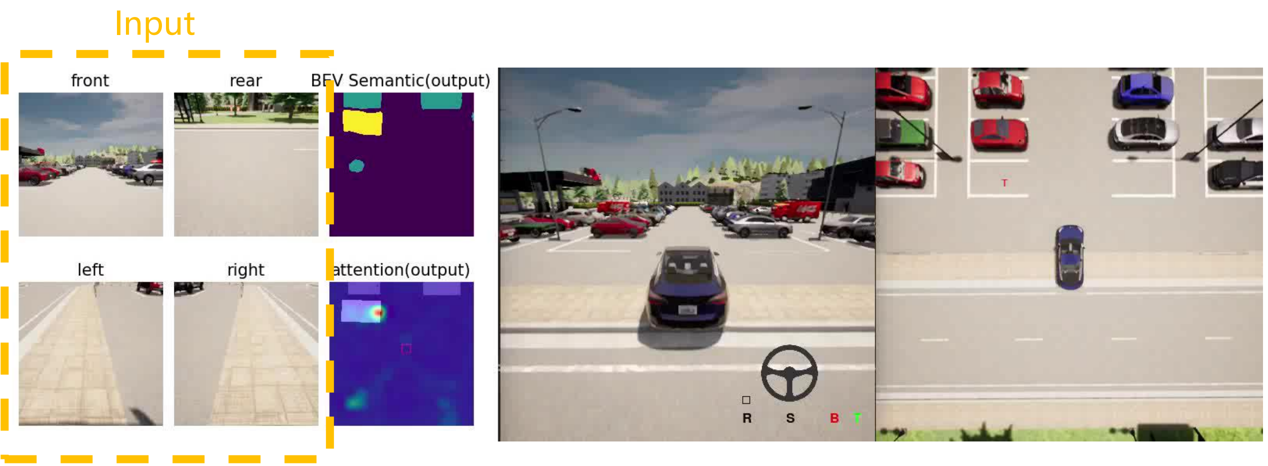
APD: Average Position Deviation
AOD: Average Orientation Deviation
APT: Average Parking Time

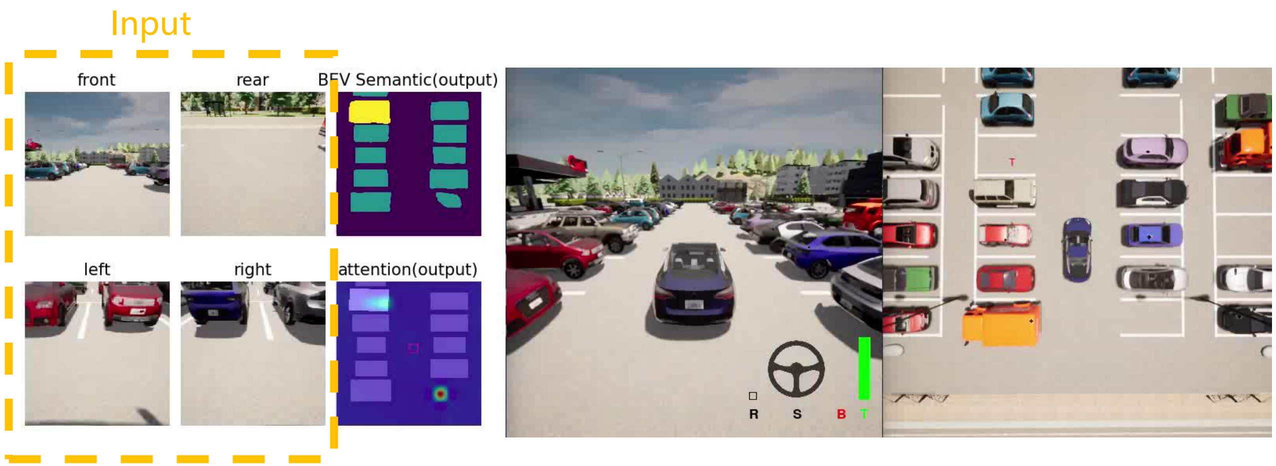
Experiment (Design choices)

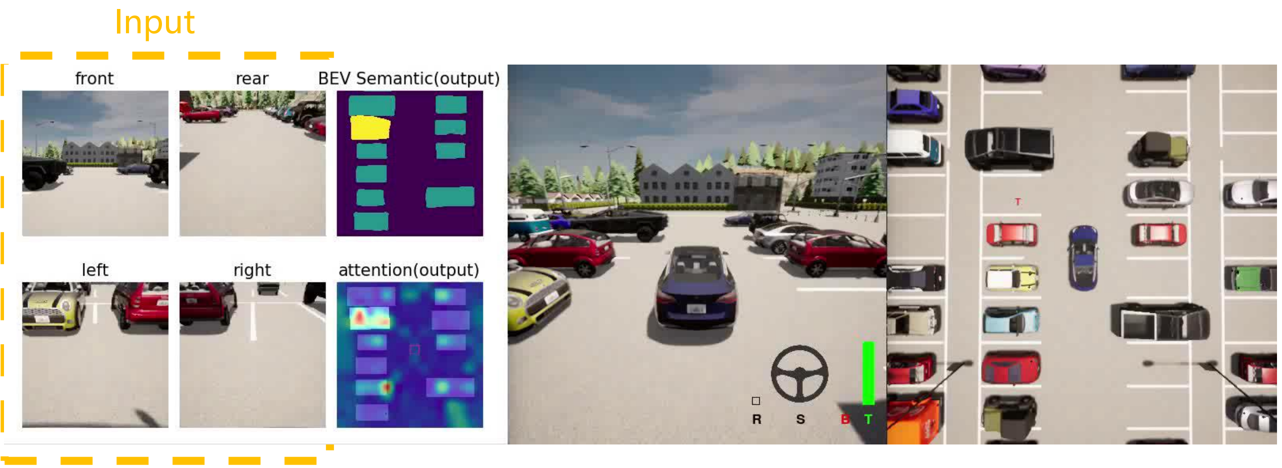
- We validate our design via extensive ablation studies
- Explicit depth supervision can boost the success rate by 14%
- Replacing the transformer decoder with an MLP structure would decrease the success rate by 8%

Agent	TSR(%)	TFR(%)	CR(%)	APD(m)	AOD(deg)	APT(s)
Baseline	91.41	2.08	2.08	0.30	0.87	15.72
w/o depth	77.08	5.20	6.25	0.29	0.80	16.37
MLP decoder	83.33	1.30	1.04	0.25	0.54	16.58

Visualization (CARLA)







Conclusion

- In this paper, we proposed a novel and feasible end-to-end visual parking solution which directly maps images and motions to the control signals.
- We designed a coordinate-free system that does not rely on explicit coordinate points and hence could track the parking goal by itself.
- To the best of our knowledge, we established the first quantitative benchmark on parking tasks in CARLA and published the parking datasets generated in CARLA for public availability
- Closed-loop experiments shows that our method achieves adequate accuracy and success rate
- In future, we will conduct experiment on real environment with more parking scenarios. We also plan to investigate the potential application of Deep Reinforce learning on parking task