

Recent Trends in E2E



RILS LAB

Robot Intelligence Learning & System

김채은

2026.08.07

01

DiffusionDrive: Truncated Diffusion Model for End-to-End Autonomous Driving

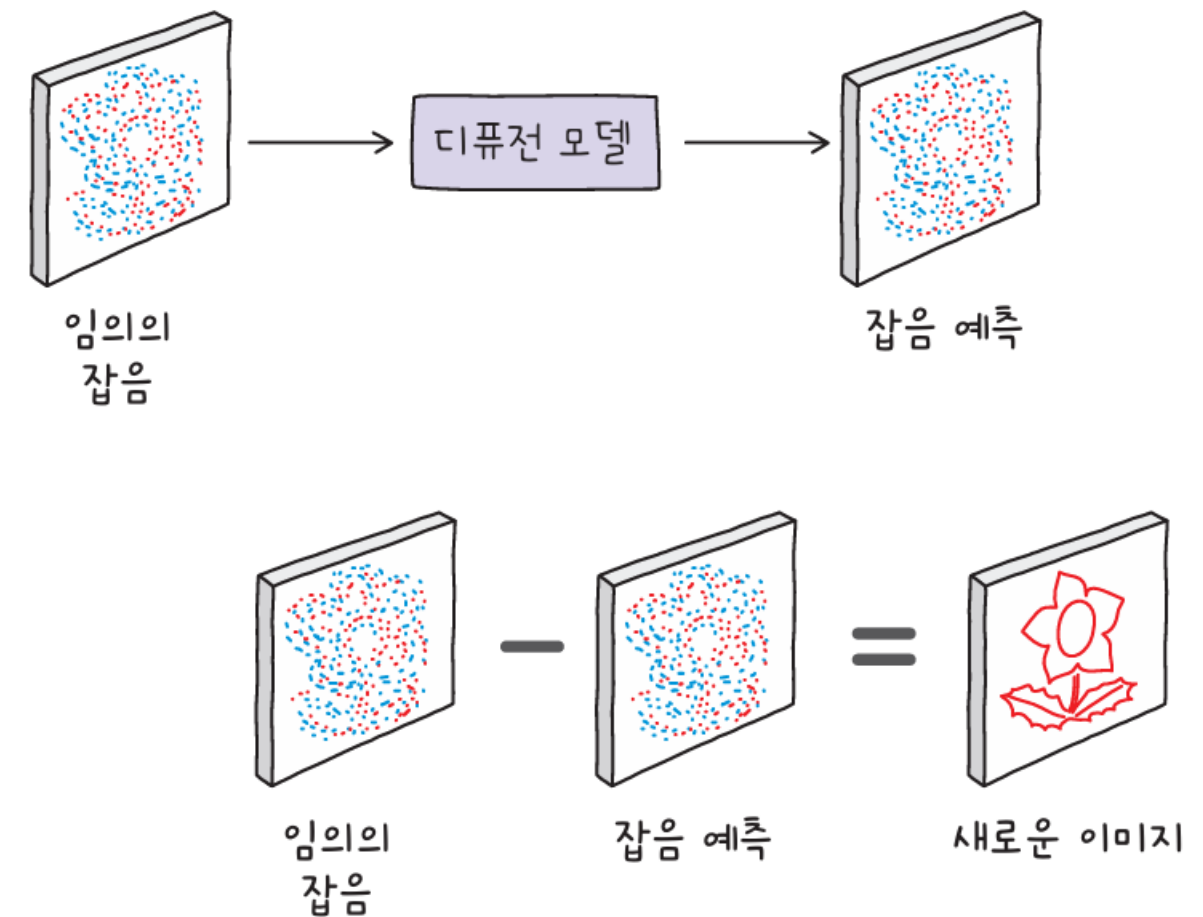
encheng Liao, Shaoyu Chen, Haoran Yin, Bo Jiang, Cheng Wang, Sixu Yan, Xinbang Zhang, Xiangyu Li, Ying Zhang, Qian Zhang, Xinggang Wang

CVPR / 2025

1. Problem Setting

Diffusion model이란?

원래 이미지 생성에서 나온 생성모델 방식으로
 노이즈가 낀 데이터에서 노이즈를 제거해 원하는 데이터를 만들어내는 방식
 GT에 가우시안 노이즈를 섞고 여기 낀 노이즈가 뭘지 예측하는 식으로 학습
 추론과정에서는 완전 랜덤한 노이즈에서 점점 노이즈를 제거하며 그럴듯한 영상을 만듦
 시작하는 노이즈가 매번 다르기 때문에 같은 조건에도 여러 개의 결과를 내 다양성 확보



1. Problem Setting

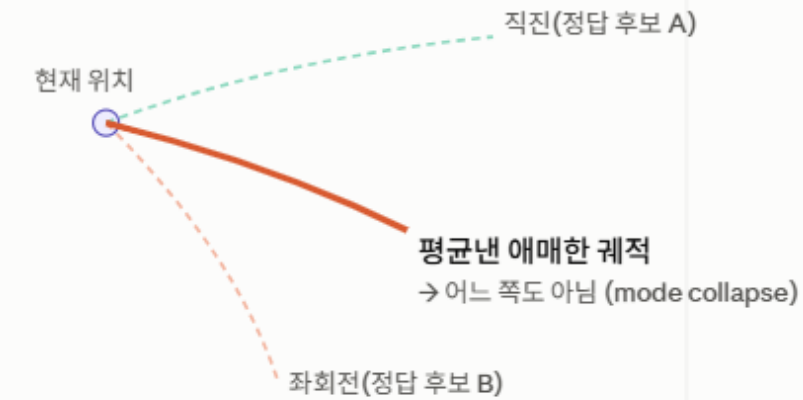
기존 자율주행의 regression 방식은 딱 한개의 궤적을 예측하는 방식이라 multi-modal 상황에서 여러 가능성을 평균내버리기 때문에 애매한 궤적이 나오는 mode collapse 문제가 발생한다.

이를 diffusion 모델을 사용하여 노이즈에서 여러번 다른 시작 점을 주어 다양한 궤적 후보를 생성하게 할 수 있다.

그러나 완전 랜덤 노이즈를 그럴듯한 모양으로 되돌리려면 수백 ~ 수천 스텝을 거쳐야하고 이는 real-time 시스템인 자율주행에 적용하기 힘들다.

Regression 방식

교차로: 직진 or 좌회전 둘 다 가능한 상황

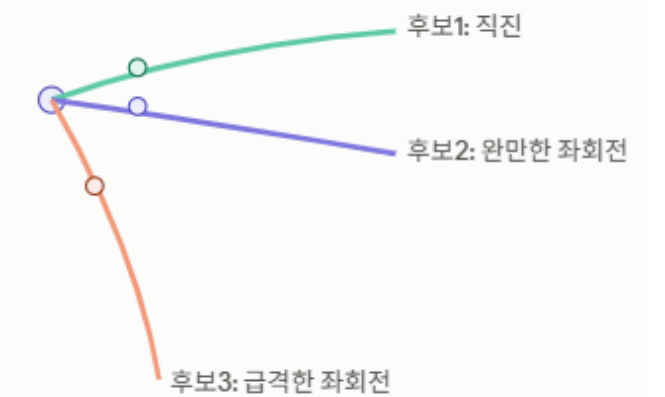


문제점

궤적 하나만 예측하는 구조라서
다중 가능성(multi-modal)을
수학적으로 평균 내버림
→ 실제로는 존재하지 않는
이도저도 아닌 경로가 출력됨

Diffusion 방식

같은 상황: 여러 노이즈 시작점 사용



각 궤적은 서로 다른
노이즈 시작점(●)에서 출발해
denoise를 거쳐 만들어짐

장점

여러 그럴듯한 궤적이
각자 뚜렷한 모양으로 공존

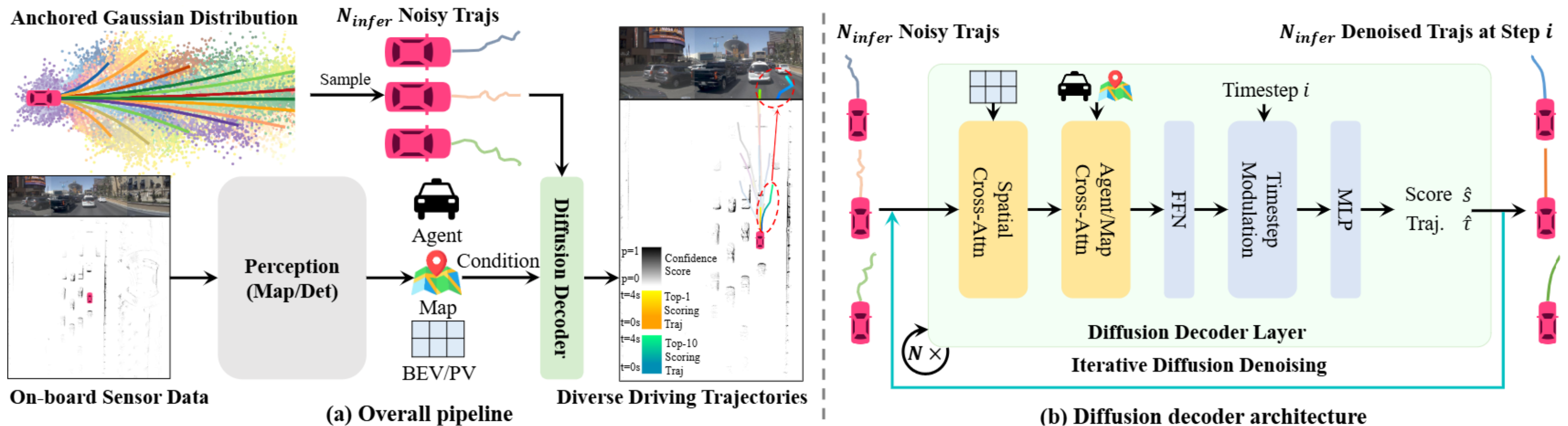
2. Solution

학습 데이터 = 현재시점의 센서 입력 + 그 이후 몇초동안의 waypoint의 나열

여기서 모든 차량의 미래 궤적만 모아서 이를 k-means클러스터링으로 묶는다. → 그렇게 나오는 대표적인 그룹이 나오는데 이 경로들의 중심을 앵커라고 말한다.

이 고정된 앵커를 중심으로 여러 개의 소량의 가우시안 노이즈를 추가하며 학습한다.

이 방법을 통해 20단계의 denoising단계를 2단계로 줄여 실시간 성능 구현



3. Contribution

- **Truncated diffusion policy 제안:**
- **효율적인 cascade diffusion decoder 설계:** spatial/agent cross-attention을 계층적으로 쌓은 2-layer 디코더로, 기존 방식 대비 파라미터를 39% 줄이면서도 계획 품질(PDMS)을 개선
- **실시간성 입증:** 앵커 20개, 2 denoising step만으로 NVIDIA 4090에서 45 FPS급 추론 속도를 달성해, diffusion 기반 방법의 실시간성 입증
- **광범위한 벤치마크 검증:** NAVSIM 등에서 기존 회귀 기반 방식(VADv2 등, 앵커 8192개 사용)보다 앵커 개수는 훨씬 적으면서도 더 나은 성능을 보임을 실험적으로 증명

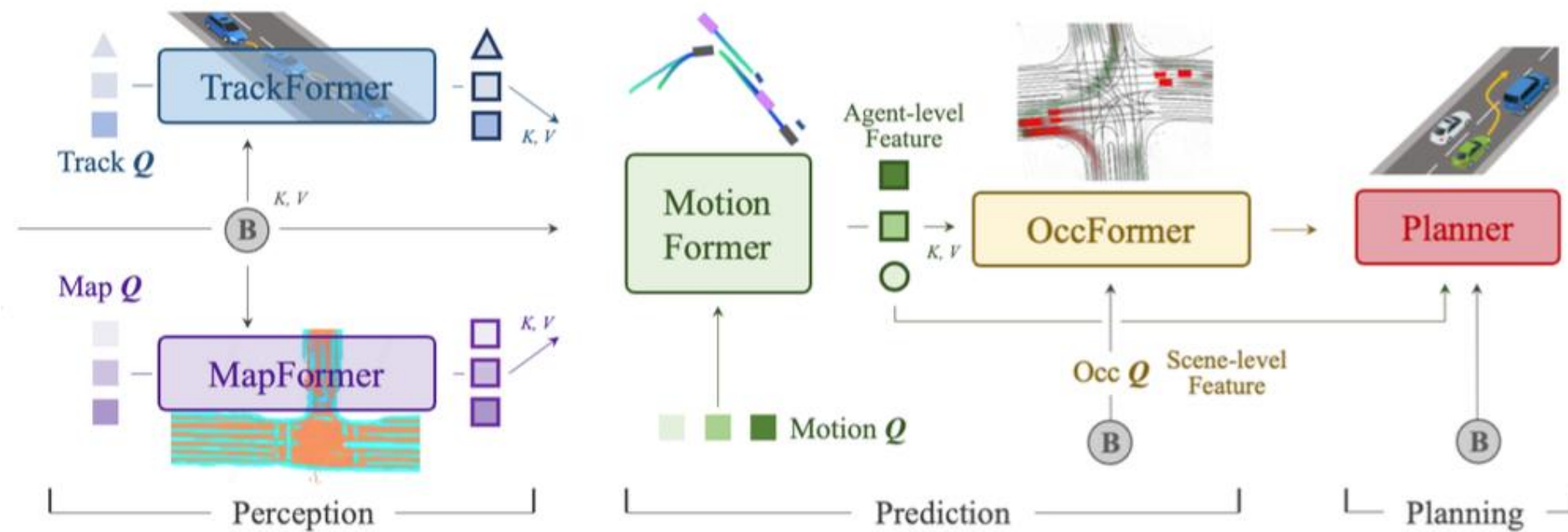
02

DriveTransformer: Unified Transformer for Scalable End-to-End Autonomous Driving

Xiaosong Jia, Junqi You, Zhiyuan Zhang, Junchi Yan

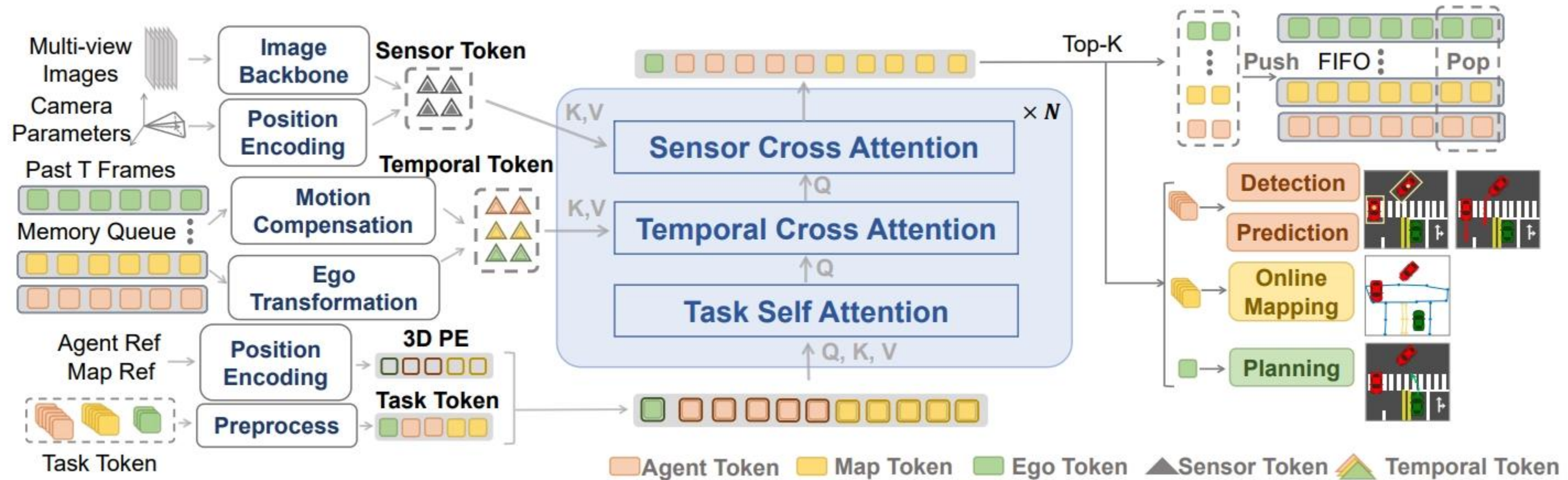
ICLR / 2025

1. Problem Setting



기존 E2E 모델을 detection-prediction- planning의 순차적 파이프라인
이러한 구조는 누적된 오류와 훈련 불안정성을 가지고 있어
이러한 문제를 해결하기 위해 하나의 스텝으로 이루어진 Drive Transformer을 도입

2. Solution

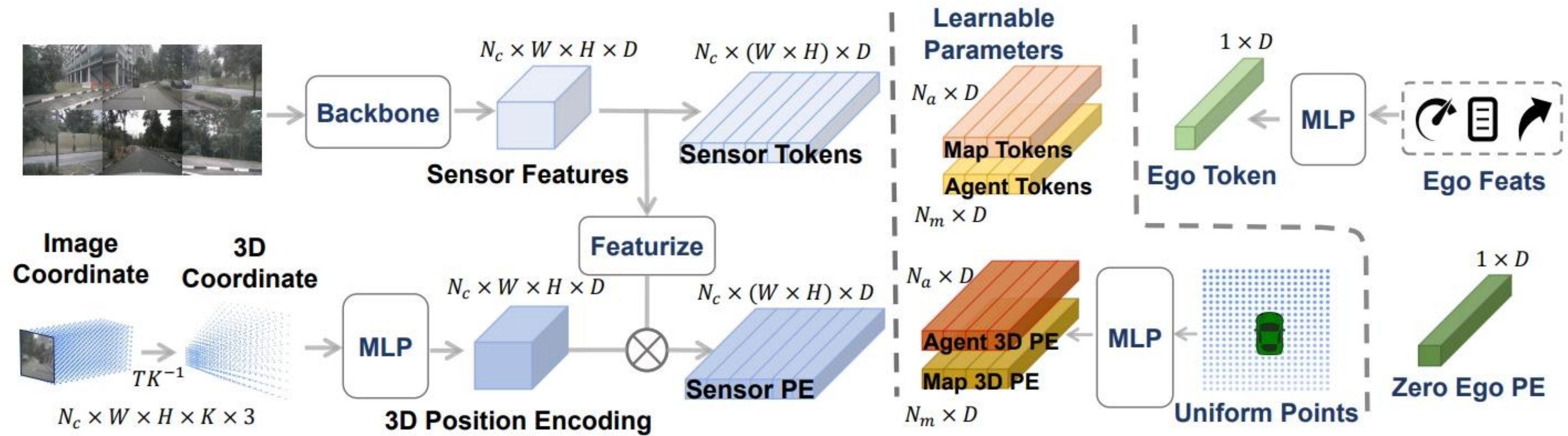


Task Parallelism : 모든 작업 쿼리가 각 블록에서 직접 상호작용하여 지식 이전이 가능하며, 시스템의 안정성을 유지합니다.

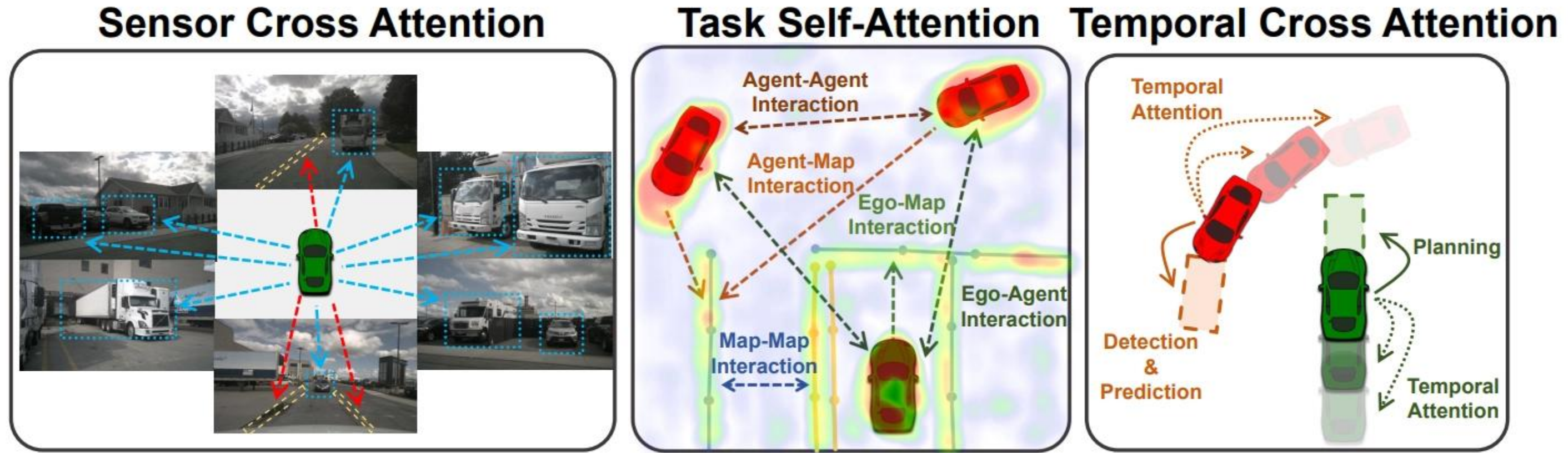
Sparse Representation : 각 작업 쿼리가 원시 센서 특징과 직접 연관되어 정보를 효율적이고 직접적으로 추출할 수 있도록 합니다.

Streaming Processing : 과거의 작업 쿼리가 FIFO 대기 열에 저장되어 시간적 정보를 통합합니다.

2. Solution: initialization & Tokenization



2. Solution:

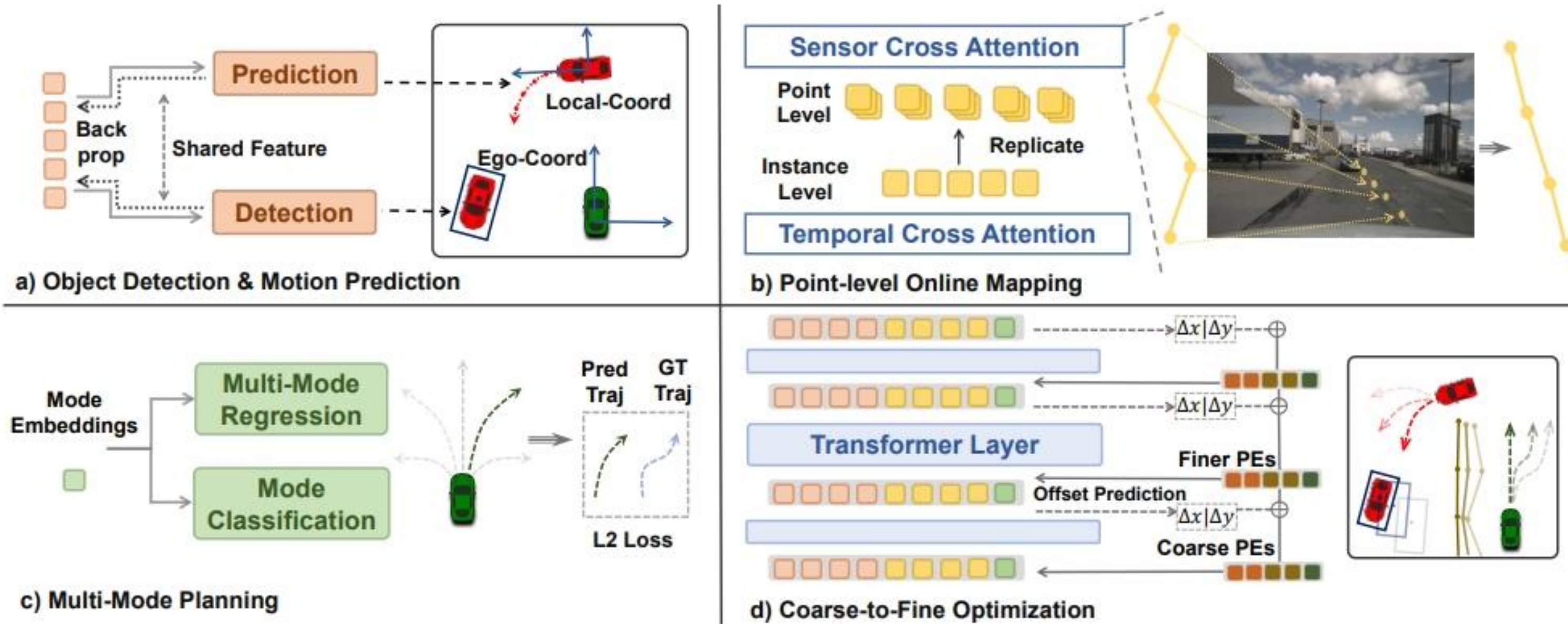


Senser Cross Attention : 쿼리가 BEV같은 중간표현을 거치지않고 센서 feature에서 직접 정보를 추출한다

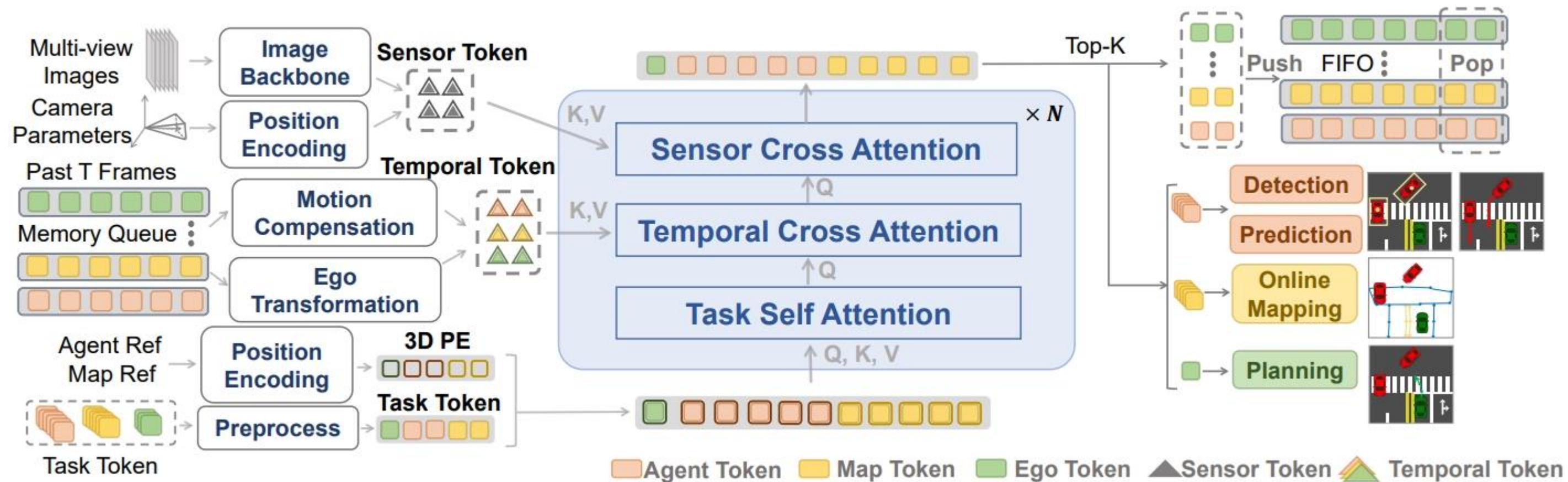
Task Self-Attention : 모든 Task token이 직접 서로 정보를 주고받아 상호의존성을 학습한다.

Temporal Cross Attention : 현재 시점의 Task토큰이 과거 시점의 Task 토큰과 상호작용하여 시간적 일관성을 유지한다.

2. Solution:



3. Contribution



- 순차 파이프라인 없이 **단일 스테이지(one-stage)** 학습만으로 순수 Transformer E2E 프레임워크
- 트래킹 모듈 없이 detection-prediction association을 attention으로 자연스럽게 해결

Thank you

Question & Discussion



RILS LAB