Recent Trends in End-to-end Autonomous Driving

RILS LAB

Robot Intelligence Learning & System

김건우

2026. 07. 16

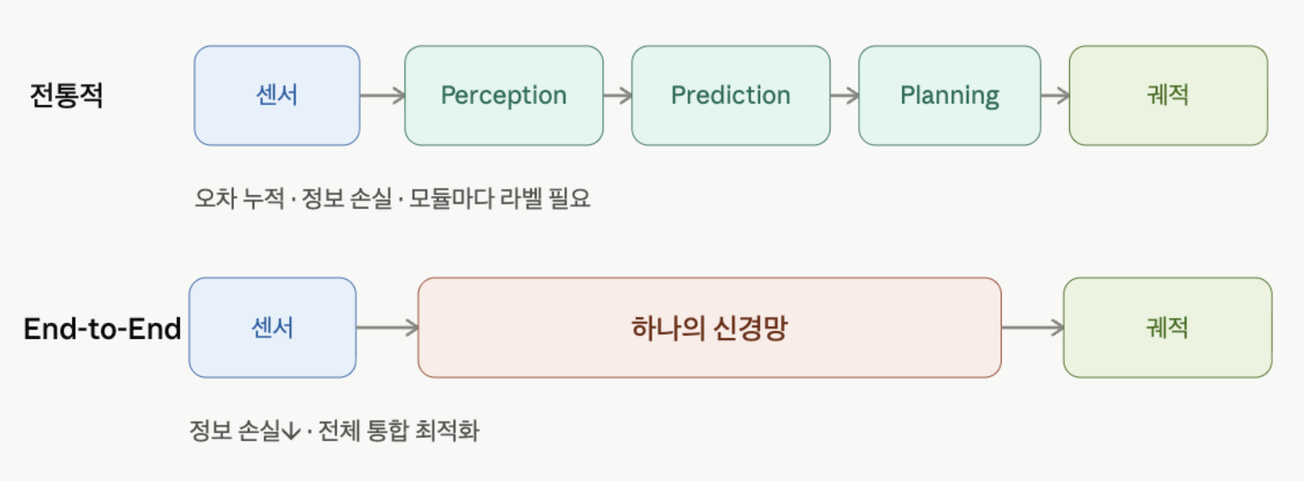
Paper List

01	Perception in Plan: Coupled Perception and Planning for End-to-End Autonomous Driving	AAAI	2026	E2E AD architecture
02	Driving on Registers	CVPR	2026	E2E AD architecture
03	LAW: Enhancing End-to-End Autonomous Driving with Latent World Model	ICLR	2025	E2E AD + World model

End-to-end 자율주행이란?

개념 : Raw sensor input에서 곧바로 주행 궤적(trajectory)을 출력하는 방식
- 즉, 중간 단계(인지→예측→계획)을 따로 두지 않고 하나의 신경망으로 통합하여 최적화

핵심 키워드
- **Perception(인지)**: 카메라 영상에서 차선·차량·보행자 등을 알아내는 단계.
- **Planning(계획)**: "어디로 어떻게 갈지" 미래 궤적(trajectory)을 정하는 단계.
- **BEV (Bird's-Eye-View)**: 카메라 여러 대 영상을 위에서 내려다본 격자 지도로 변환한 표현.
- **trajectory / waypoint**: 차가 앞으로 지나갈 점들의 좌표 시퀀스.
- **open-loop vs closed-loop**: 정답 궤적과 비교만 하는 평가(open) vs 시뮬레이터에서 실제 주행시키는 평가(closed)



01

Perception in Plan: Coupled Perception and Planning for End-to-End Autonomous Driving

- 저자 : Bozhou Zhang, Jingyu Li, Nan Song, Li Zhang
- 연구팀 / 소속 : 1. School of Data Science, Fudan University
  2. Shanghai Innovation Institute
- 학회 정보 : AAAI / 2026
- 인용 수 : 8회
- 깃허브 : https://github.com/LogosRoboticsGroup/VeteranAD

문제 설정

기존 E2E 자율주행은 perception이 planning보다 선행되는 순차적 패러다임이 대부분

즉, perception이 "지금 어디로 가려는지(좌회전? 우회전?)"를 모른 채 장면 전체를 균일하게 인지한다.

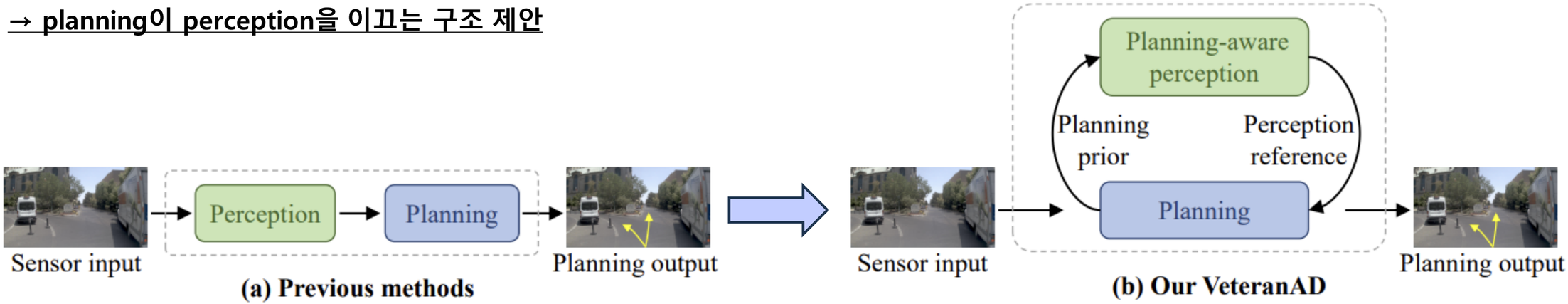
→ 계획에 정작 중요한 영역에 집중하지 못하고, 변화하는 주행 의도에 perception이 반응하지 못함

→ **E2E 자율주행에서 계획 지향적 최적화의 이점을 완전히 활용하기에 충분하지 않는 것이라고 지적.**

Idea

"인지를 먼저 다 하지 말고, 계획하면서 필요한 것만 그때그때 인지하자."

→ planning이 perception을 이끄는 구조 제안



#1 Perception in Plan: Coupled Perception and Planning for End-to-End Autonomous Driving

파이프라인

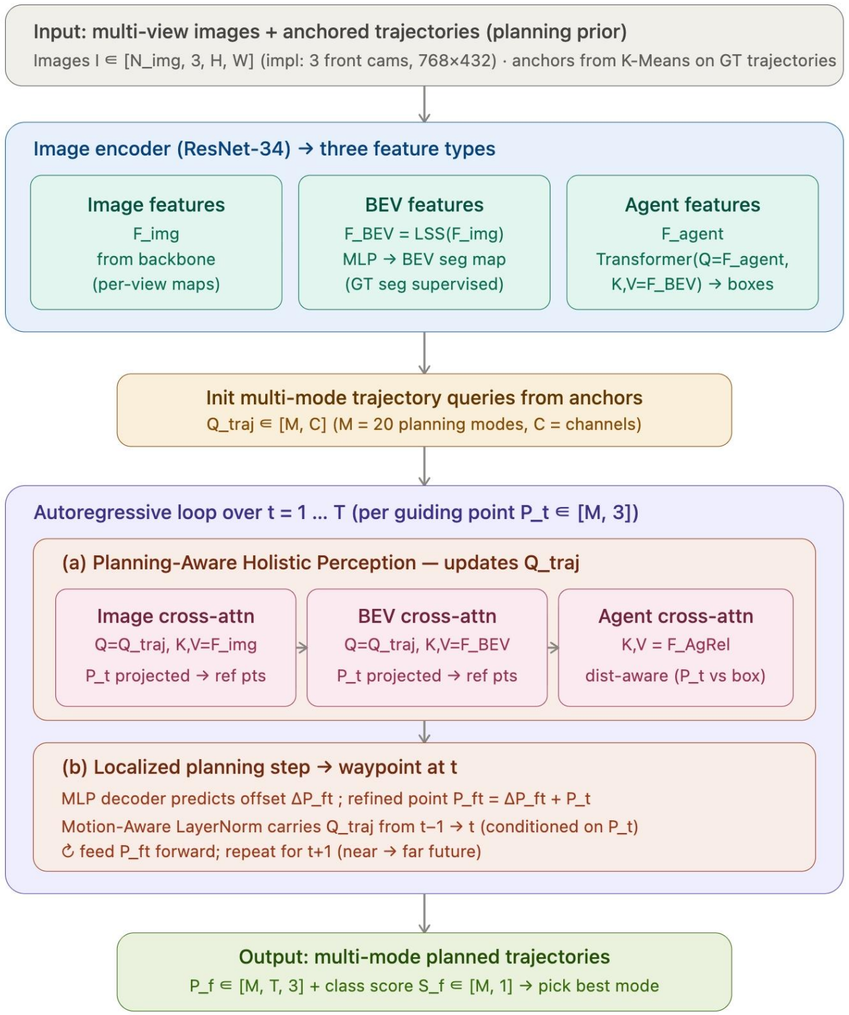
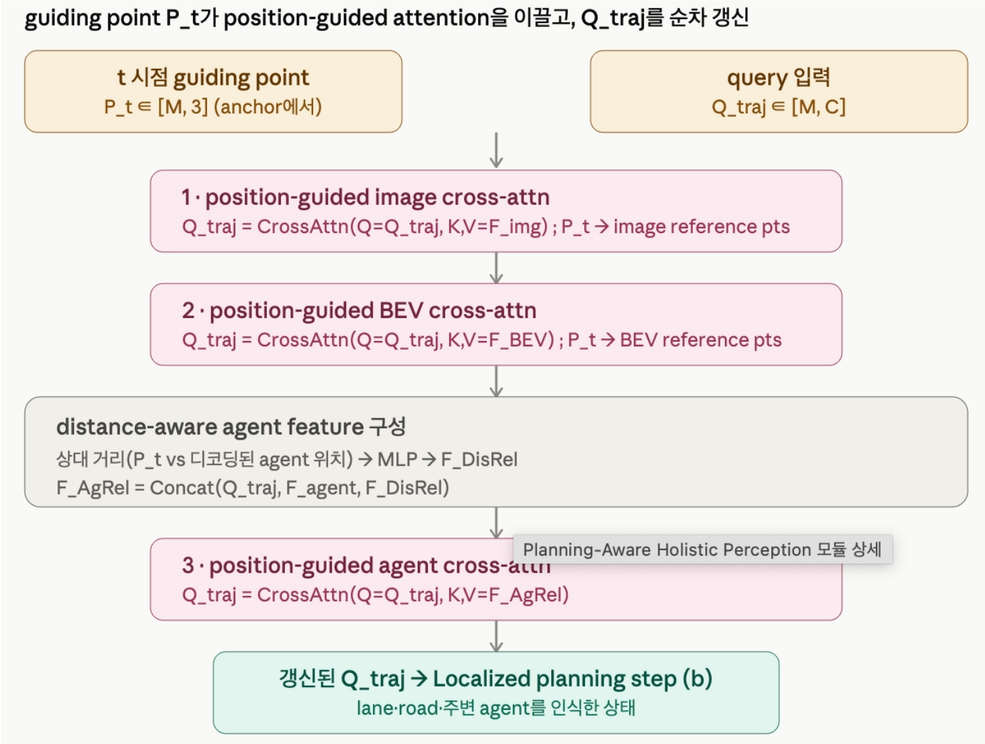


Image Encoder가 이미지를 세 표현(F_{img} / F_{BEV} / F_{agent})으로 나눠 놓고, Q_{traj}가 셋 모두를 참조하며 Autoregressive loop 안에서 매 스텝 perception을 다시 수행한다.

Planning-Aware Holistic Perception 내부

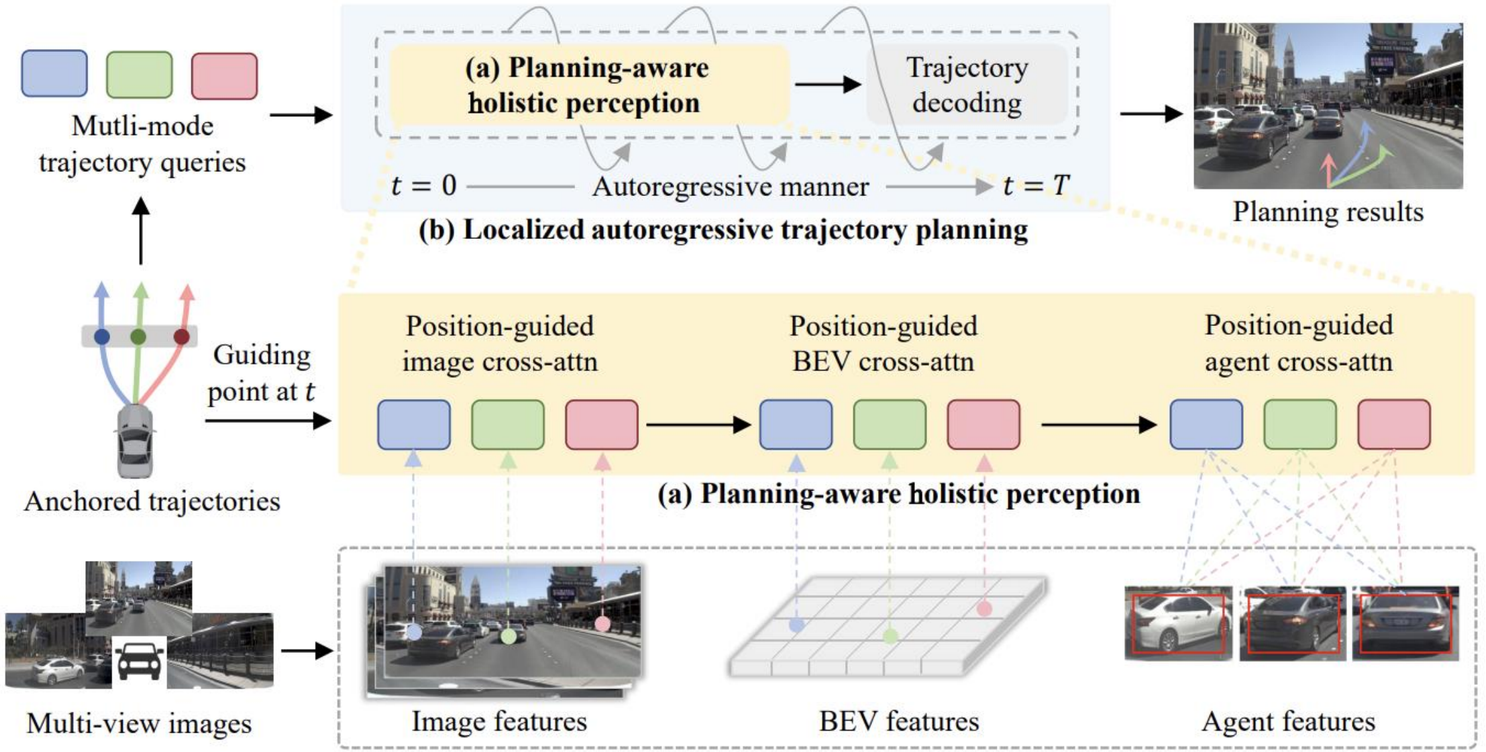


Contribution

(i) Perception을 planning 과정에 통합하는 **"perception-in-plan" 패러다임**을 따르는 새로운 프레임워크 제안

(ii) Planning-Aware Holistic Perception 모듈과 Localized Autoregressive Trajectory Planning 모듈의 2가지 핵심 모듈을 설계

→ 인지과 계획을 공동으로 결합 + E2E 자율 주행이 가능하게 하는 계획 지향 최적화의 이점을 완전히 발휘하게 함.



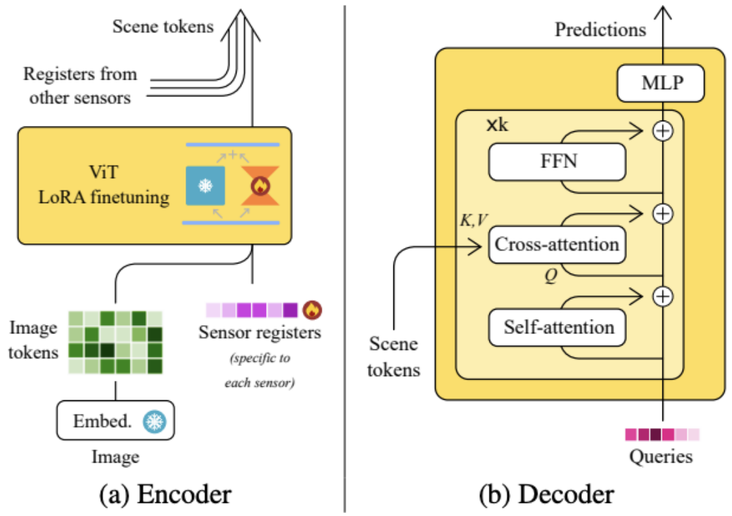
02

Driving on Registers

- 저자 : Ellington Kirby, Alexandre Boulch, Yihong Xu, Yuan Yin, Gilles Puy, Éloi Zablocki, Andrei Bursuc, Spyros Gidaris, Renaud Marlet, Florent Bartoccioni, Anh-Quan Cao, Nermin Samet, Tuan-Hung VU, Matthieu Cord
- 연구팀 / 소속 : 1. valeo.ai, Paris, France,
  2. LIGM, ENPC, IP Paris, UGE, CNRS, France
  3. Sorbonne Université, CNRS, ISIR, France
- 학회 정보 : CVPR / 2026
- 인용 수 : 14회
- 깃허브 : https://github.com/valeoai/DrivoR

문제 설정

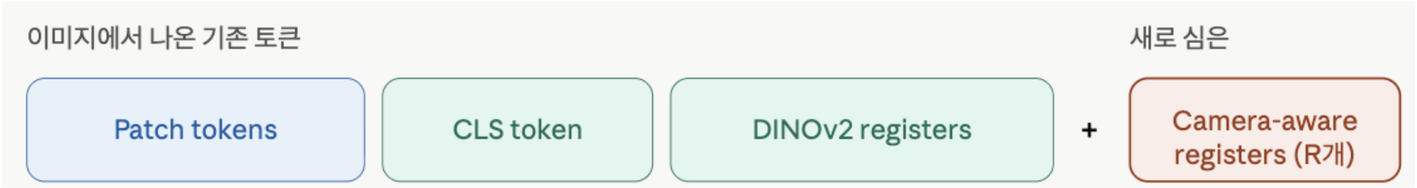
E2E 자율주행은 카메라 영상의 수많은 feature를 planning으로 넘긴다.
BEV 같은 무거운 중간 표현을 쓰거나, 모든 patch feature를 그대로 쓰면 계산량이 크다.
→ perception 정보와 planning(궤적 생성·평가)이 얽혀 서로 방해할 수 있음.



Idea

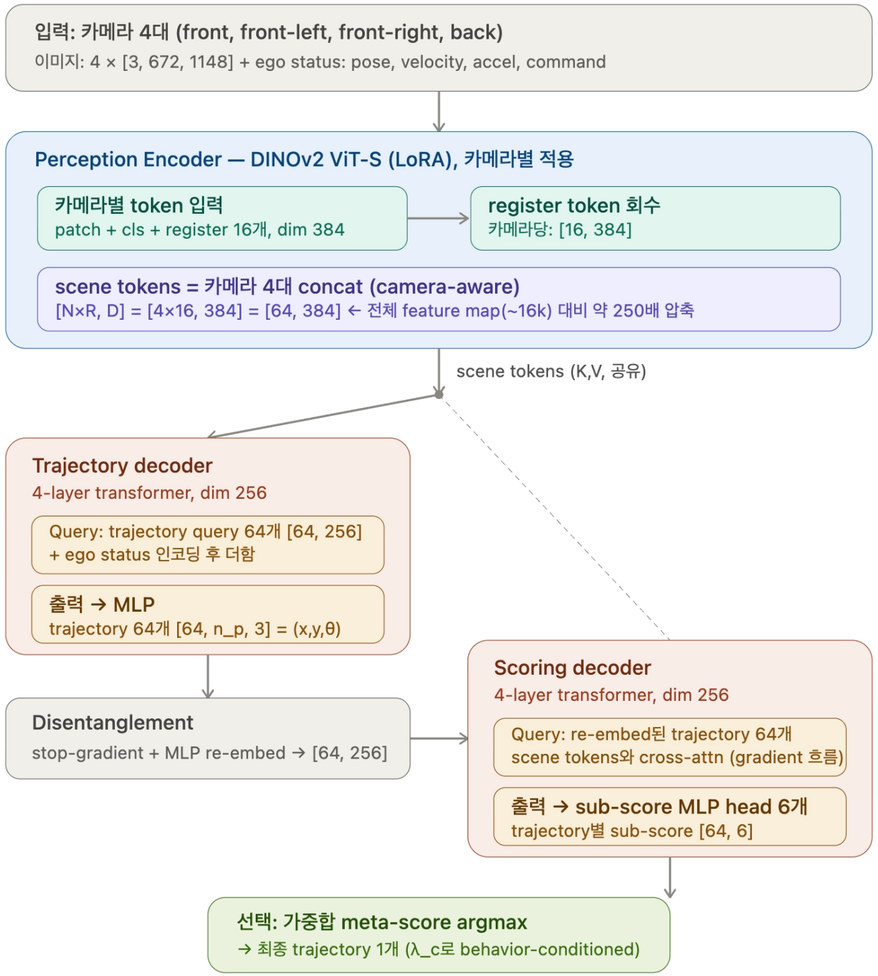
" 이미지 전체를 보지말고, "무거운 중간표현 대신 register 64개로 장면을 압축하고, 궤적 생성과 평가를 분리하자!"

→ 기존 ViT 토큰처럼 카메라별 레지스터 토큰을 함께 넣고 인코더를 통과시켜서 이 레지스터 토큰만 모아서 디코더로 전달



#2 Driving on Registers

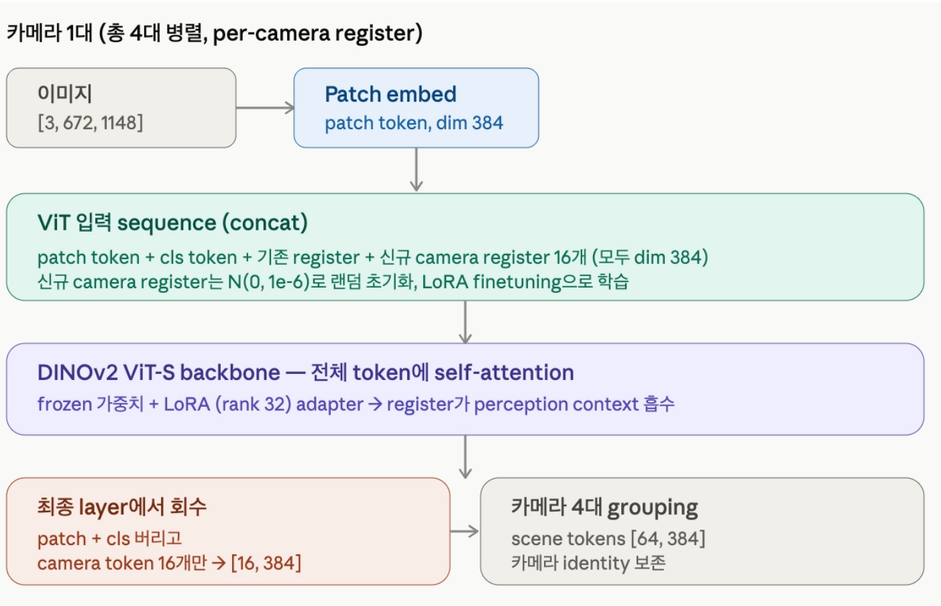
전체 파이프라인



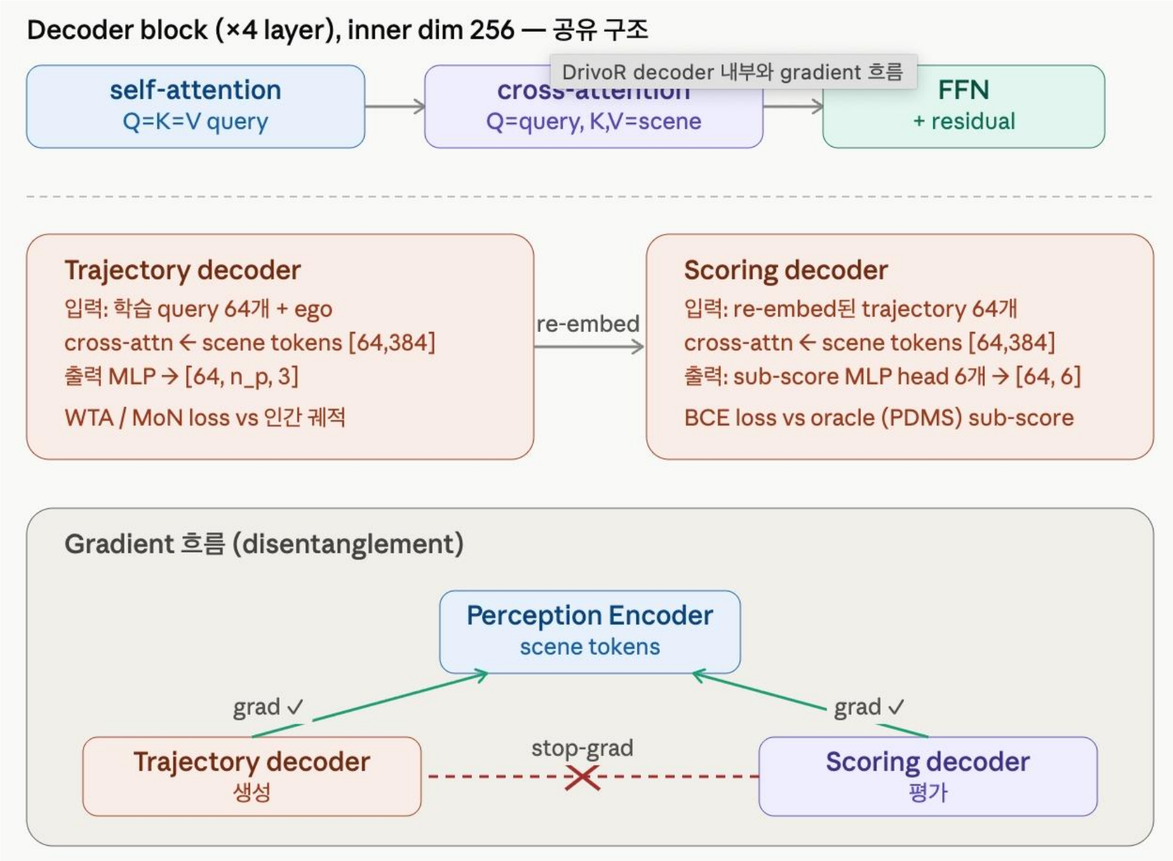
인코더가 카메라당 16개 register만 뽑아
전체 4096개(64×64) 수준의 패치 토큰 대신 [64, 384] scene token으로 압축
-> 64개 토큰이 두 디코더의 K,V로 공유된다.

Perception Encoder 내부(카메라 1대 기준)

핵심은 ViT에 patch·cls·register token을 함께 넣되,
최종 layer에서 register token 16개만 회수하는 것



두 **decoder** 내부와 **gradient** 흐름(disentanglement)



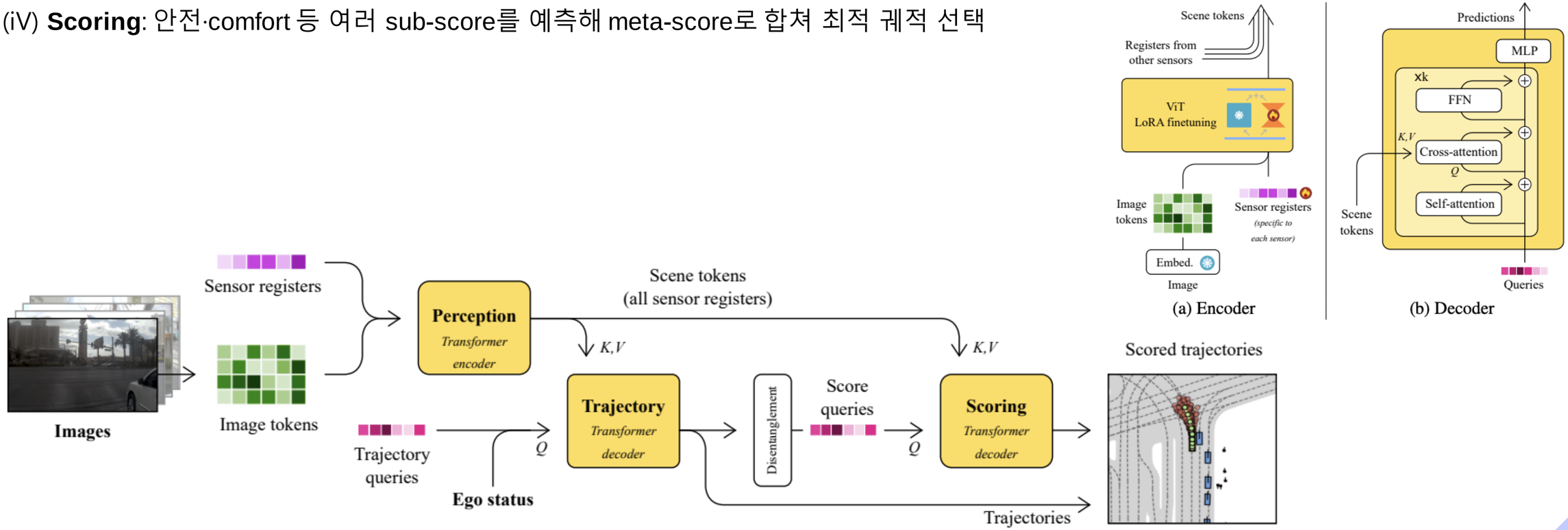
Contribution

(i) 중간 BEV 표현이나 대규모 trajectory dictionary 없는 simple transformer 아키텍처

(ii) E2E Planning을 위해 register-based 토큰 압축을 활용하여 ViT 기반 이미지 백본의 구조적인 이점을 탐구한 첫 번째 연구

(iii) **Disentanglement**: 궤적 생성(trajectory decoder)과 궤적 평가(scoring decoder)를 stop-gradient로 분리 → 서로 간섭 방지

(iV) **Scoring**: 안전·comfort 등 여러 sub-score를 예측해 meta-score로 합쳐 최적 궤적 선택



03

LAW: Latent World Model

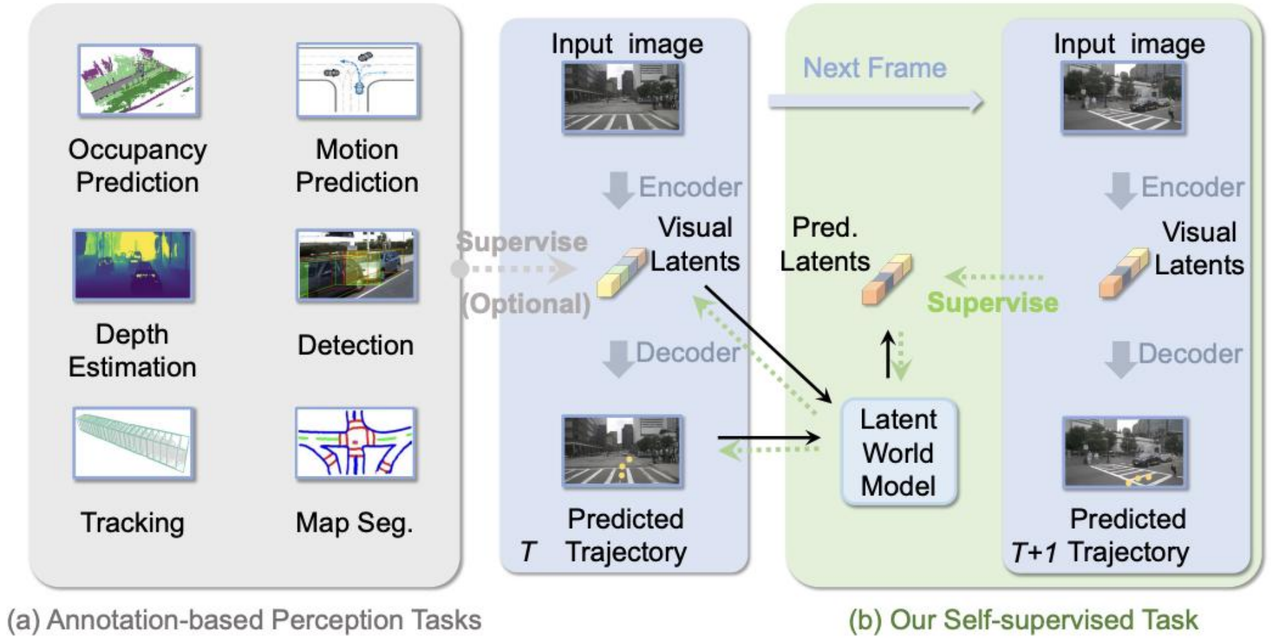
- 저자 : Yingyan Li, Lue Fan, Jiawei He, Yuqi Wang, Yuntao Chen, Zhaoxiang Zhang, Tieniu Tan
- 연구팀 / 소속 : 1. Institute of Automation, Chinese Academy of Sciences, CASIA
  2. New Laboratory of Pattern Recognition, NLPR
  3. State Key Laboratory of Multimodal Artificial Intelligence Systems, MAIS
  4. School of Future Technology, University of Chinese Academy of Sciences, UCA
- 학회 정보 : ICLR / 2025
- 인용 수 : 151회
- 깃허브 : https://github.com/BraveGroup/LAW

문제 설정

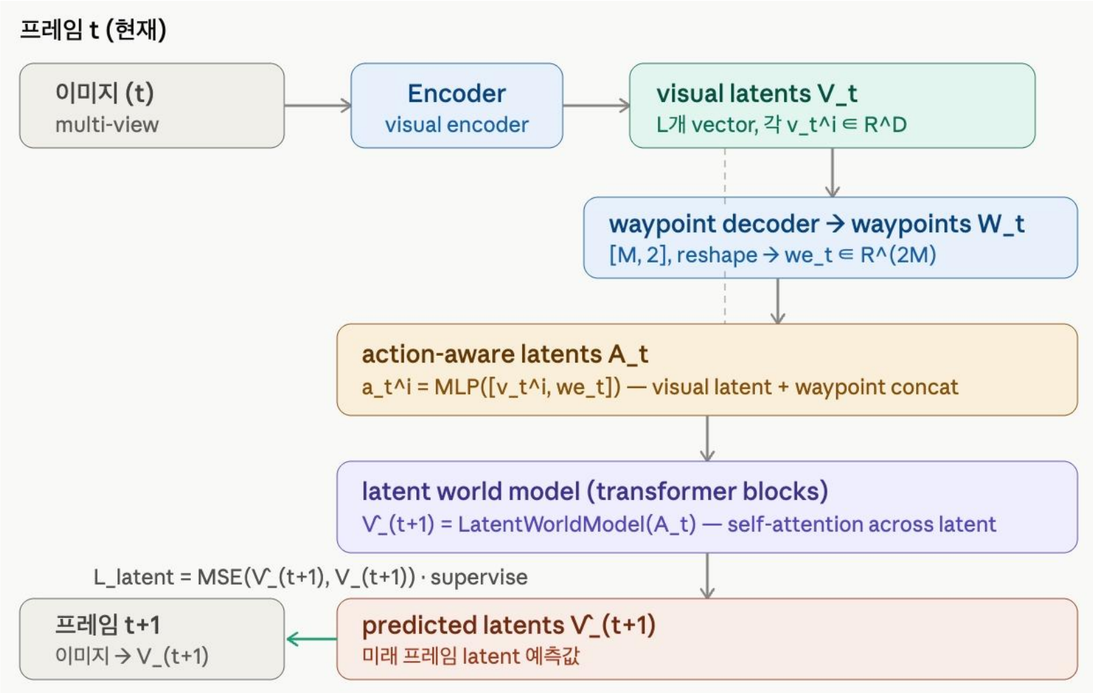
E2E 자율주행의 scene feature를 잘 학습시키려면 보통 perception annotation (detection·segmentation 등 라벨)이 많이 필요하다.
하지만 라벨은 비싸고, 주행은 연속적인 비디오인데 기존 방법들은 temporal(시간) 정보를 충분히 활용하지 못했다.

Idea

" 자기지도학습으로 데이터 자체에서 학습 신호를 뽑고, 현재 latent와 예측된 궤적으로 미래 latent를 예측하자!"



latent world model의 self-supervised loop

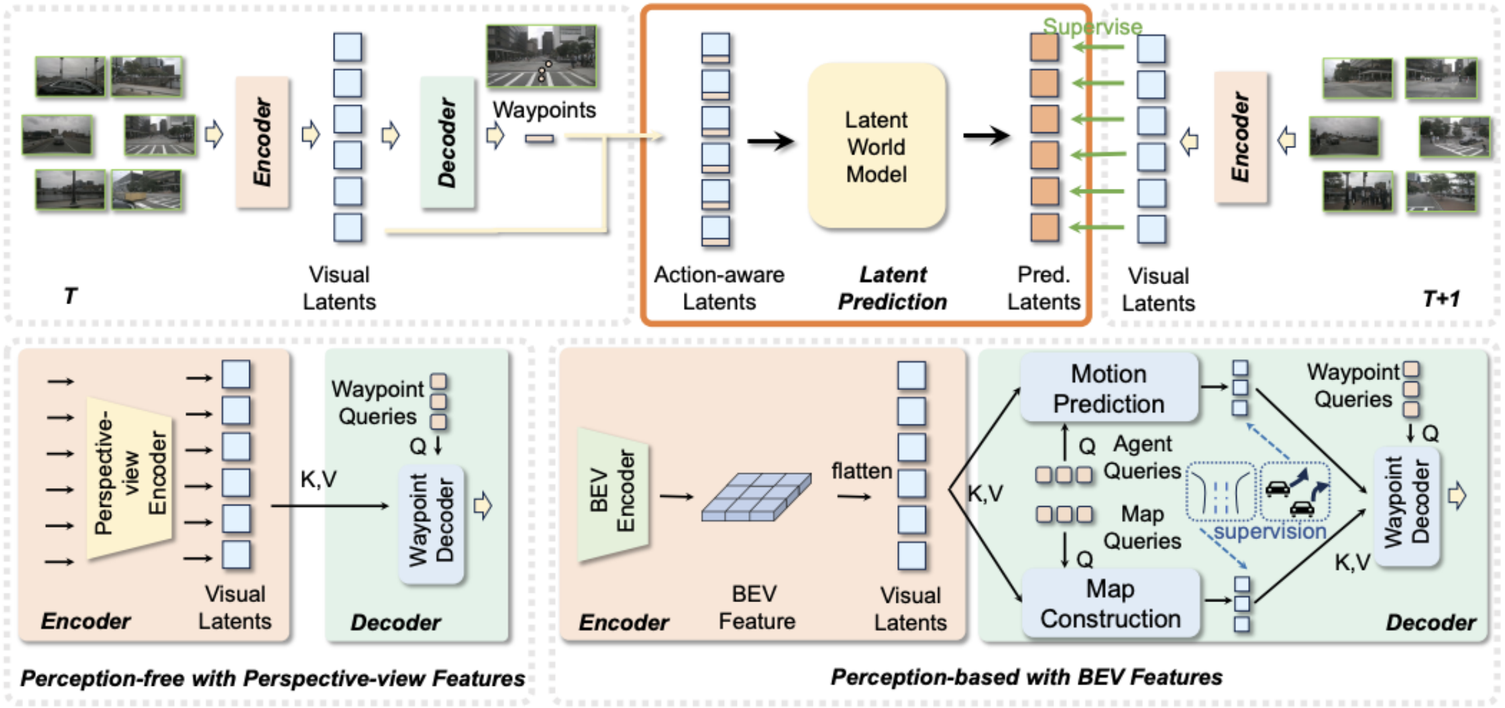


V_t(현재 latent)와 W_t(예측 waypoint)를 합쳐 action-aware latents A_t를 만들고,
latent world model이 미래 latent V̂_{t+1}을 예측한 뒤, 실제 다음 프레임에서 뽑은 V_{t+1}으로 MSE supervise한다.
이 self-supervised task가 scene feature 학습과 trajectory 예측을 동시에 개선한다. perception annotation이 필요 없다.

Contribution

(i) **Latent World Model을 통한 미래 예측 :** 현재 latent + 예측한 waypoint로 미래 latent을 예측(실제 다음 프레임 latent로 지도학습)

(ii) **Cross-framework universality** : 같은 방식이 perception-free와 perception-based 두 framework에 모두 붙는 범용성을 가진다.



정리

- **Perception in Plan**: perception을 planning 안에 결합 (planning이 인지를 이끔)
- **DrivoR**: perception을 register로 압축하고 planning과 분리
- **LAW**: perception을 self-supervised latent 예측으로 대체

Thank you

Question & Disscusion

RILS LAB