

26-하계 논문 세미나(E2E)



RILS LAB

Robot Intelligence Learning & System

김건우

2026.08.21

Paper List

FastDriveVLA: Efficient End-to-End Driving via Plug-and-Play Reconstruction-based Token Pruning	AAAI	2026	 E2E AD
DriveSuprim: Towards Precise Trajectory Selection for End-to-End Planning	AAAI	2026	 E2E AD
ResAD: Normalized Residual Trajectory Modeling for End-to-End Autonomous Driving	CVPR	2026	 E2E AD
SafeDrive: Fine-Grained Safety Reasoning for End-to-End Driving in a Sparse World	CVPR	2026	 E2E AD
SimScale: Learning to Drive via Real-World Simulation at Scale	CVPR	2026	 E2E AD

01

FastDriveVLA: Efficient End-to-End Driving via Plug-and-Play Reconstruction-based Token Pruning

**Jiajun Cao^{1,2}, Qizhe Zhang¹, Peidong Jia¹, Xuhui Zhao^{1,2}, Bo Lan^{1,2}, Xiaoan Zhang^{1,2},
Zhuo Li², Xiaobao Wei¹, Sixiang Chen¹, Liyun Li², Xianming Liu², Ming Lu¹,
Yang Wang^{2†}, Shanghang Zhang^{1†}**

¹State Key Laboratory of Multimedia Information Processing, School of Computer Science,
Peking University ²XPeng Motors

AAAI 2026 (인용 34회)

<https://arxiv.org/abs/2507.23318>

문제 상황

기존 Token Pruning 방법

VLA 모델은 E2E 자율주행에 점점 더 채택되고 있지만, 긴 visual token이 계산 비용을 크게 증가시킨다. 기존 VLM의 token pruning은 두 갈래인데, 둘 다 자율주행에서 성능이 좋지 않았다.

- ① Attention 기반 — visual-text attention을 최대화하는 토큰 선택
- ② Similarity 기반 — 토큰 간 유사도를 최소화하는 선택

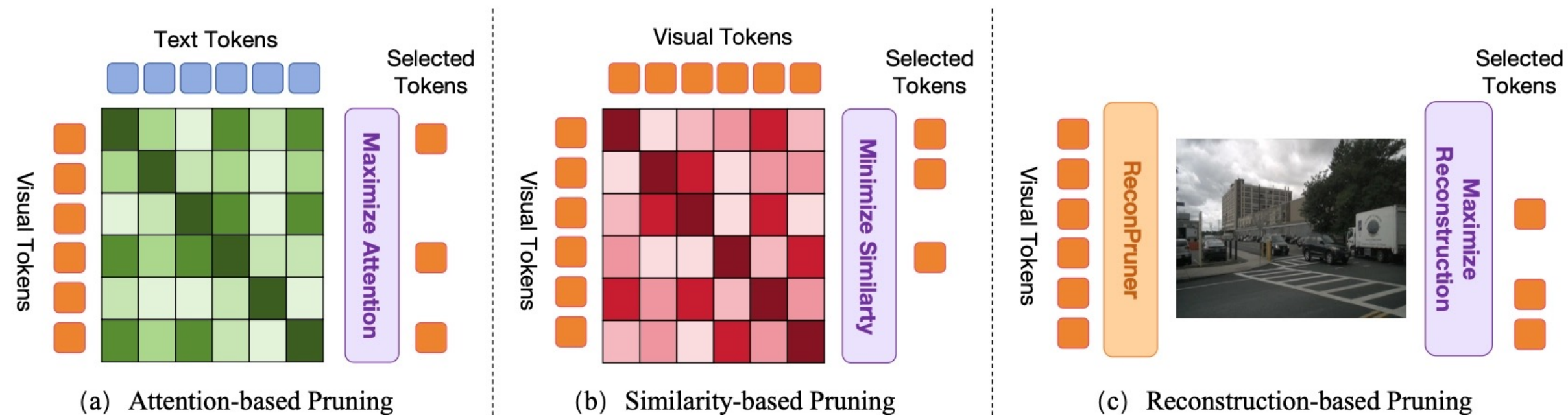


Figure 1: Comparison of different visual token pruning strategies.

사람은 운전할 때 배경보다 전경에 집중한다 — 모델도 그래야 하지 않을까?

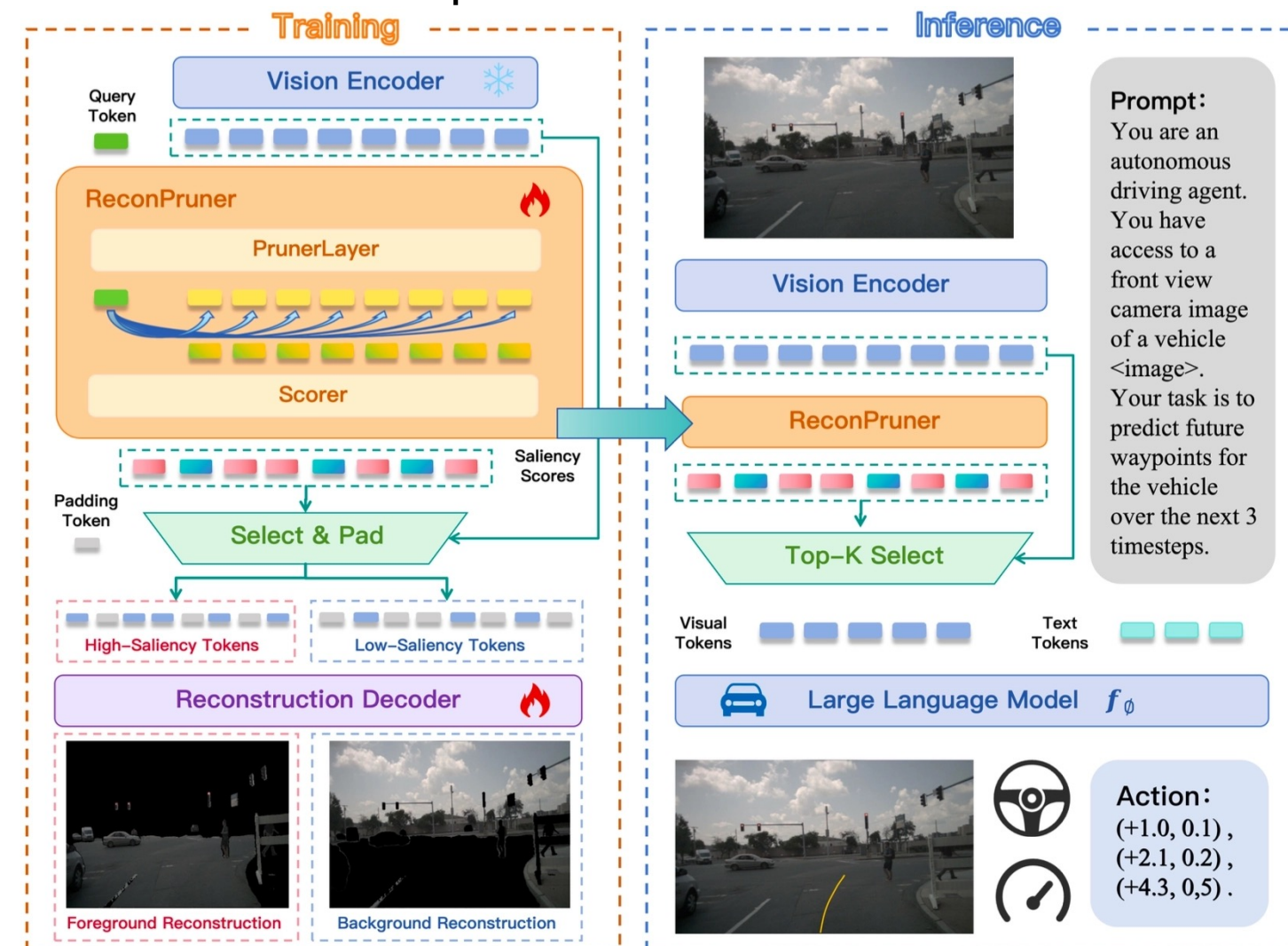
핵심 아이디어

MAE 방식 픽셀 재구성으로 전경 토큰을 우선 보존

- ① ReconPruner(MAE 스타일 픽셀 재구성으로 각 visual token을 채점하는 plug-and-play pruner)
 - 전경 영역 복원에 기여하는 토큰은 높은 점수, 배경 패치는 조기 탈락.
- ② 적대적 전경-배경 재구성 전략
 - 전경과 배경 사이에 판별적 attention을 강제해 ReconPruner가 local optima에 빠지는 것을 방지.
- ③ nuScenes-FG(학습용 데이터셋) 구축



Figure 2: **nuScenes-FG**. It contains 241k foreground segmentation annotations for scenes in the nuScenes dataset.



핵심 기여

- 기존 attention 기반·similarity 기반과 차별화되는 새로운 reconstruction-based token pruning 프레임워크 FastDriveVLA를 제안했다.
- MAE 스타일 픽셀 재구성으로 학습되는 플러그앤플레이 pruner인 ReconPruner를 설계하고, 가치 있는 토큰 식별 능력 향상을 위해 적대적 전경-배경 재구성 전략을 도입했다.
- 자율주행 시나리오를 위한 전경 세그멘테이션 주석 241k image-mask 쌍의 nuScenes-FG 데이터셋을 구축했다.
- nuScenes open-loop planning 벤치마크에서 다양한 pruning 비율 전반에 걸쳐 SOTA 달성.
→ visual token을 3,249 → 812개로 줄여 약 7.5배 계산량 감소하면서도 planning 정확도 유지.

Methods	L2 (cm) ↓					Collision (%) ↓					Intersection (%) ↓				
	1s	2s	3s	Avg.	Rel.	1s	2s	3s	Avg.	Rel.	1s	2s	3s	Avg.	Rel.
<i>Input size 1596 × 1596, 3249 tokens (100%)</i>															
Impromptu-VLA (NeurIPS25)	13.97	28.38	53.13	31.83	100%	0.00	0.13	0.60	0.24	100%	0.53	2.34	5.52	2.80	100%
<i>Retain 2436 Tokens (↓ 25%)</i>															
FastV (ECCV25)	14.23	28.85	53.80	32.29	98.6%	0.00	0.18	0.74	0.31	79.3%	0.52	2.44	5.65	2.87	97.4%
SparseVLM (ICML25)	14.09	28.72	53.74	32.18	98.9%	0.00	0.17	0.67	0.28	86.9%	0.51	2.41	5.52	2.81	99.4%
VisPruner (ICCV25)	14.02	28.50	53.44	31.99	99.5%	0.00	0.17	0.61	0.26	93.6%	0.51	2.40	5.51	2.81	99.6%
DivPrune (CVPR25)	14.17	28.83	53.72	32.24	98.7%	0.00	0.17	0.73	0.30	81.1%	0.50	2.47	5.61	2.86	97.8%
FastDriveVLA (Ours)	13.99	28.36	53.04	31.80	100.1%	0.00	0.15	0.63	0.26	93.6%	0.53	2.36	5.42	2.77	101.0%
<i>Retain 1624 Tokens (↓ 50%)</i>															
FastV (ECCV25)	14.29	29.14	54.33	32.59	97.7%	0.00	0.20	0.79	0.33	73.7%	0.52	2.67	5.77	2.99	93.6%
SparseVLM (ICML25)	14.24	28.97	54.17	32.46	98.0%	0.00	0.18	0.73	0.30	80.2%	0.53	2.62	5.73	2.96	94.5%
VisPruner (ICCV25)	14.16	28.77	53.82	32.25	98.7%	0.00	0.17	0.65	0.27	89.0%	0.52	2.54	5.78	2.95	94.9%
DivPrune (CVPR25)	14.20	28.98	54.12	32.43	98.1%	0.00	0.20	0.78	0.33	74.5%	0.50	2.63	5.72	2.95	94.8%
FastDriveVLA (Ours)	14.08	28.65	53.57	32.10	99.1%	0.00	0.15	0.60	0.25	97.3%	0.55	2.49	5.78	2.94	95.1%
<i>Retain 812 Tokens (↓ 75%)</i>															
FastV (ECCV25)	14.63	29.54	54.97	33.05	96.3%	0.00	0.21	0.79	0.33	73.0%	0.58	2.63	5.76	2.99	93.5%
SparseVLM (ICML25)	14.58	29.47	54.81	32.95	96.6%	0.00	0.21	0.75	0.32	76.0%	0.57	2.58	5.74	2.96	94.4%
VisPruner (ICCV25)	14.42	29.38	54.52	32.77	97.1%	0.00	0.19	0.73	0.31	79.3%	0.52	2.57	5.72	2.94	95.2%
DivPrune (CVPR25)	14.50	29.46	54.57	32.84	96.9%	0.00	0.20	0.76	0.32	76.0%	0.55	2.54	5.70	2.93	95.4%
FastDriveVLA (Ours)	14.28	29.18	54.46	32.64	97.5%	0.00	0.18	0.70	0.29	83.0%	0.55	2.50	5.68	2.91	96.1%

Table 1: Performance comparison of different pruning methods on Impromptu-VLA. Input images are of resolution 1596×1596, resulting in 3249 visual tokens. Here, Rel. represents the average percentage of performance maintained, and the underlined values indicate improvements over the original unpruned model.

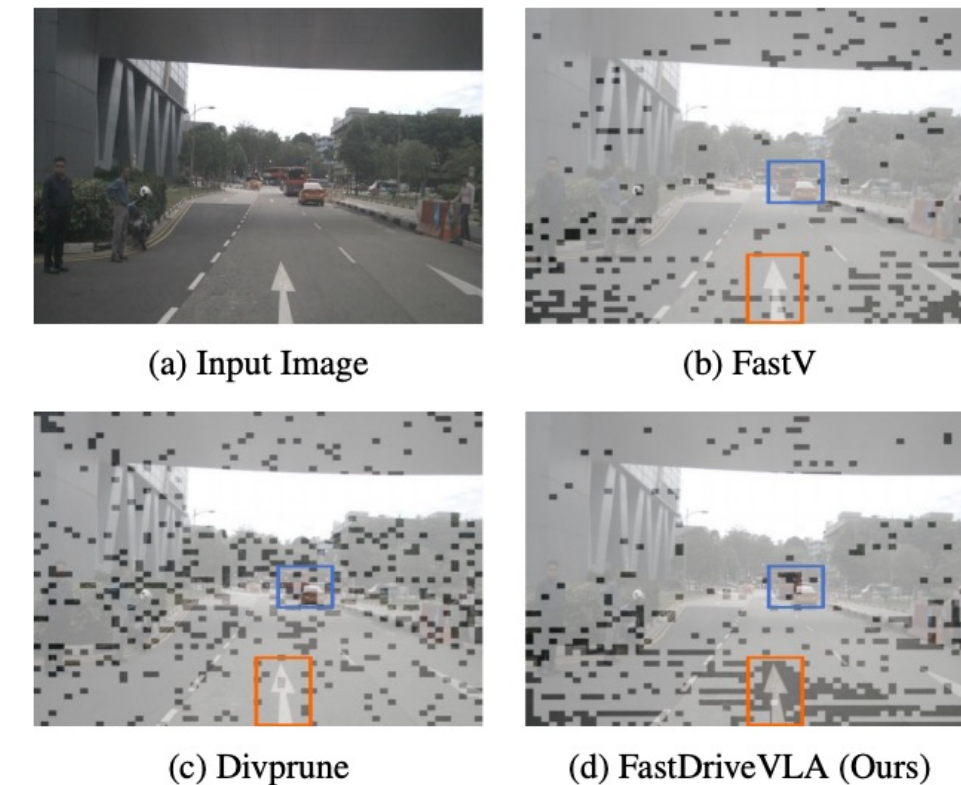


Figure 5: Visual comparison of visual tokens retained by different visual token pruning methods.

02

DriveSuprim: Towards Precise Trajectory Selection for End-to-End Planning

**Wenhao Yao^{1, 2}, Zhenxin Li^{1, 2}, Shiyi Lan³, Zi Wang³,
Xinglong Sun³, Jose M. Alvarez³, Zuxuan Wu^{1, 2*}**

¹Shanghai Key Lab of Intell. Info. Processing, School of CS, Fudan University

²Shanghai Collaborative Innovation Center of Intelligent Visual Computing

³NVIDIA

{whyao23, lizx23}@m.fudan.edu.cn, zxwu@fudan.edu.cn

AAAI 2026 (인용 73회)

<https://arxiv.org/abs/2506.06659>

<https://github.com/William-Yao-2000/DriveSuprim>

문제 상황

기존의 Selection-based 방법

Regression 기반은 단일 expert 궤적을 모방할 뿐 예측 궤적의 안전성을 명시적으로 평가하지 않는다. Selection 기반은 여러 후보를 생성·채점해 이를 보완하지만, 수천 개 후보 중 최선을 정확히 고르는 최적화가 어렵고, 특히 드물고 어려운 시나리오에서 미묘하지만 안전에 결정적인 차이를 구분하지 못한다.

① 😊 Easy Negative



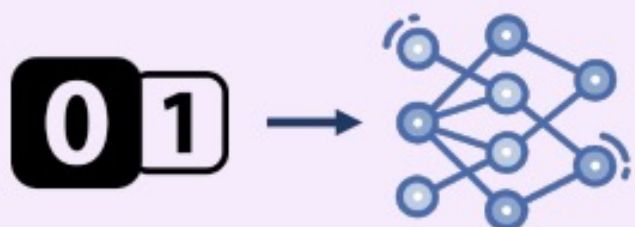
😞 Hard Negative



② 😞 Sharp Turns



③ 😞 Hard Label

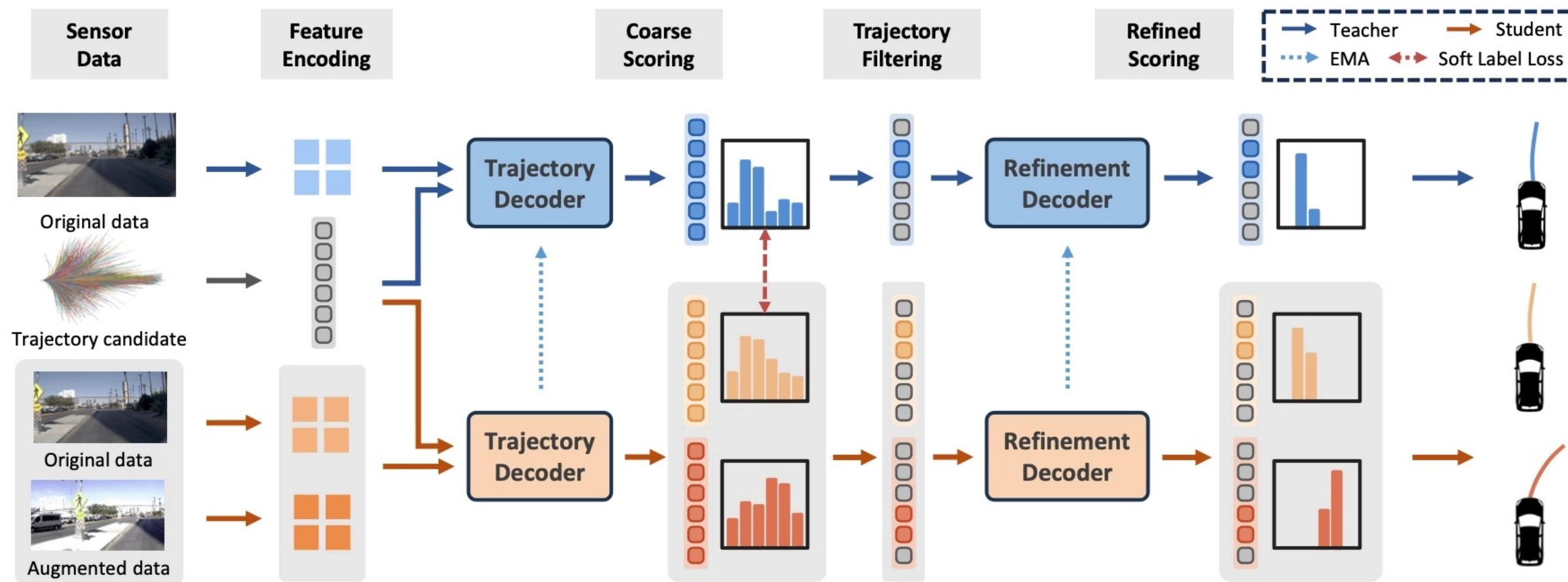


비슷비슷한 궤적 수천 개 중에서 "진짜 안전한 것"을 어떻게 골라낼까?

핵심 아이디어

Coarse-to-Fine으로 점진적 후보 필터링

- ① Trajectory Decoder가 점수 분포에 따라 유망한 후보만 1차 선별
- ② Refinement Decoder가 남은 어려운 후보들에 세밀한 점수 부여
- ③ Rotation 기반 데이터 증강으로 회전 시나리오 보강
- ④ Self-distillation으로 teacher의 soft pseudo-label 활용



한 번에 다 고르지 말고, 걸러낸 뒤 어려운 것만 자세히 보자!

핵심 기여

- 유사 궤적 구분 문제를 해결하는 **coarse-to-fine refinement 제안했다**.
→ 유망 후보를 필터링한 뒤 가장 어려운 선택지에 세밀한 점수를 매겨 미묘한 차이의 판별력을 크게 향상시켰다.
- Rotation 기반 데이터 증강 + self-distillation을 결합한 통합 학습 파이프라인**
→ 방향 편향(directional bias)과 hard decision boundary 문제를 동시에 해결.
→ 어려운 회전 시나리오를 합성하고 teacher의 soft pseudo-label을 활용해 궤적 분포를 균형 있게 만들었다.
- 공개 데이터셋만으로 NAVSIM 벤치마크 SOTA 달성
(v1 93.5% PDMS, v2 87.1% EPDMS, Bench2Drive 83.02 DS / 60.00 SR)

Method	Backbone	NC ↑	DAC ↑	EP ↑	TTC ↑	C ↑	PDMS ↑
Human Agent	—	100	100	87.5	100	99.9	94.8
Transfuser (2022)	ResNet34	97.7	92.8	79.2	92.8	100	84.0
UniAD (2023)	ResNet34	97.8	91.9	78.8	92.9	100	83.4
VADv2 (2024)	ResNet34	97.9	91.7	77.6	92.9	100	83.0
LAW (2024b)	ResNet34	96.4	95.4	81.7	88.7	99.9	84.6
DRAMA (2024)	ResNet34	98.0	93.1	80.1	94.8	100	85.5
Hydra-MDP (2024c)	ResNet34	98.3	96.0	78.7	94.6	100	86.5
DiffusionDrive (2025)	ResNet34	98.2	96.2	82.2	94.7	100	88.1
DriveSuprim	ResNet34	97.8	97.3	86.7	93.6	100	89.9 (+1.8)
Hydra-MDP (2024c)	V2-99	98.4	97.8	86.5	93.9	100	90.3
DriveSuprim	V2-99	98.0	98.2	90.0	94.2	100	92.1 (+1.9)
Hydra-MDP (2024c)	ViT-L	98.4	97.7	85.0	94.5	100	89.9
DriveSuprim	ViT-L	98.6	98.6	91.3	95.5	100	93.5 (+3.6)

Table 2: Evaluation on NAVSIM v1. Results are grouped by backbone types.



Figure 4: Visualization results across various challenging scenarios. In each example, the green trajectory represents the ground truth from the human expert, the red trajectory is generated by Hydra-MDP++, and the blue trajectory is produced by DriveSuprim.

***ResAD*: Normalized Residual Trajectory Modeling for End-to-End Autonomous Driving**

Zhiyu Zheng^{1,◇} Shaoyu Chen^{2,†} Haoran Yin² Xinbang Zhang² Jialv Zou^{3,◇}
Xinggang Wang³ Qian Zhang² Lefei Zhang^{1, ✉}

¹ School of Computer Science, Wuhan University ² Horizon Robotics

³ School of EIC, Huazhong University of Science & Technology

<https://duckyee728.github.io/ResAD>

CVPR 2026 (인용 13회)

<https://arxiv.org/abs/2510.08562>

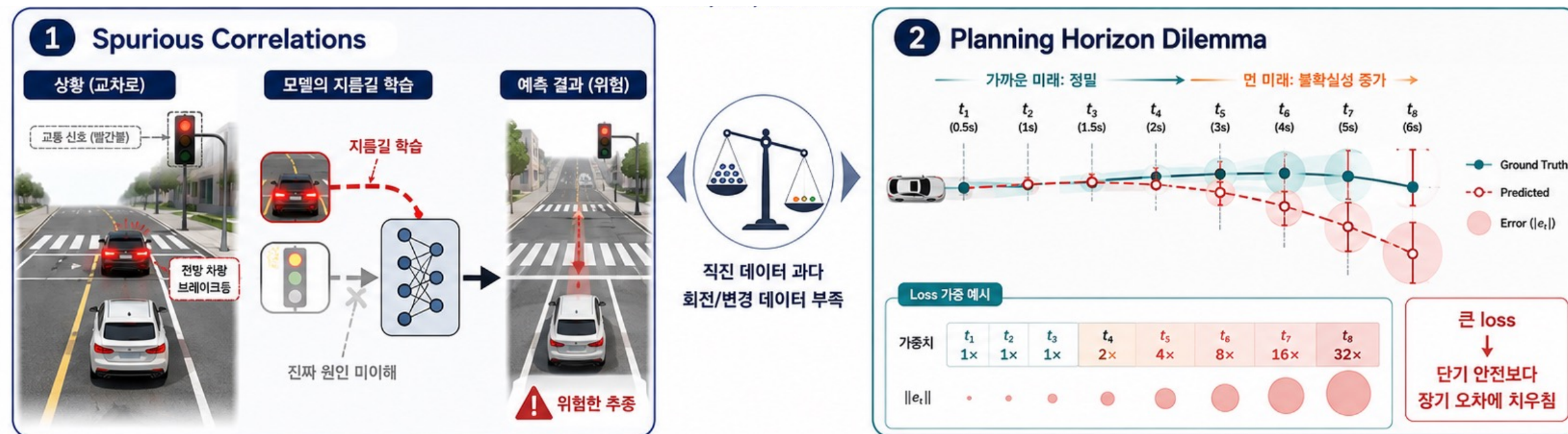
<https://duckyee728.github.io/ResAD>

문제 상황

궤적 데이터의 시공간적 불균형

E2E 자율주행은 센서 데이터에서 미래 궤적을 직접 예측하는데,
궤적 데이터 자체에 내재된 시공간적 불균형 때문에 근본적인 어려움을 겪는다.

- 이 불균형은 ① 모델이 견고한 주행 논리 대신 허위 상관 관계를 학습하게 만들고, — (Spurious Correlation)
② 불확실한 먼 미래 예측을 우선시하게 만들어 당장의 안전을 해친다. — (Planning Horizon Dilemma)

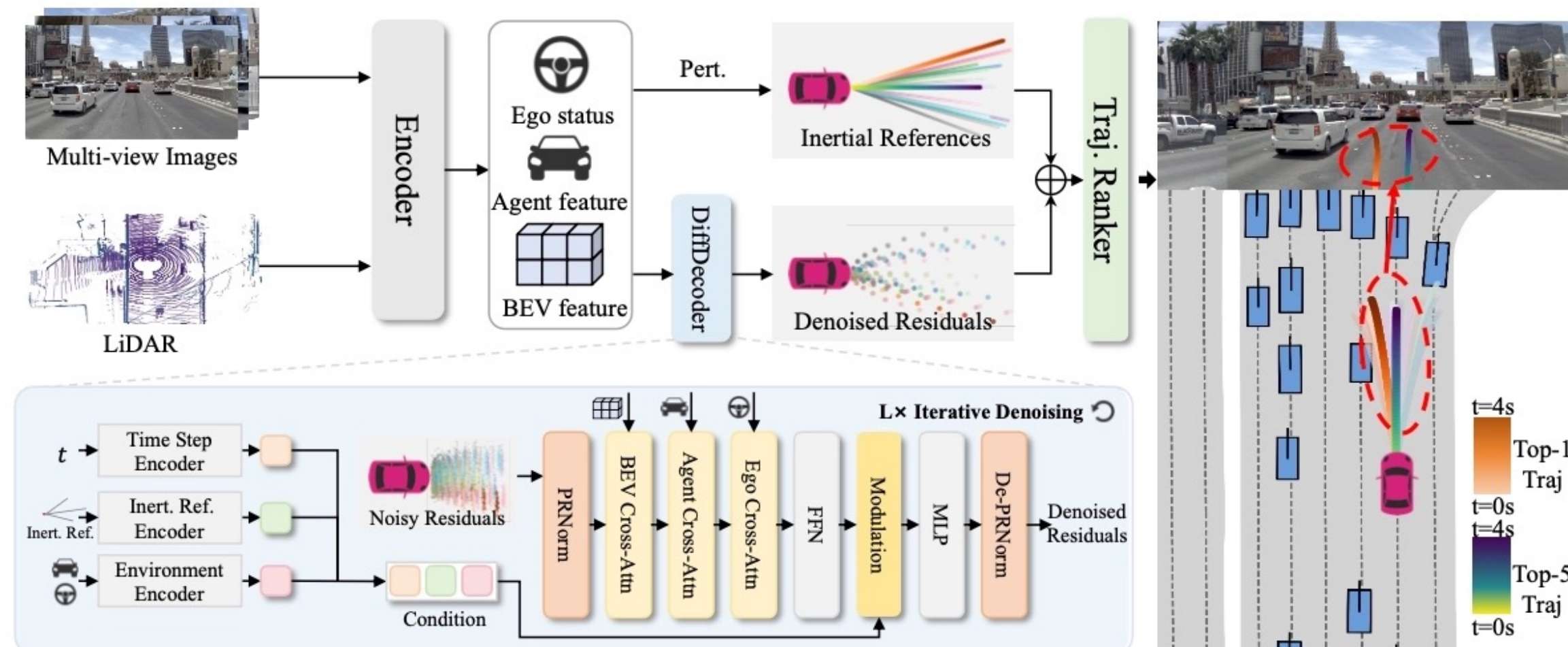


먼 미래의 큰 오차가 학습을 지배하면, 당장의 안전을 놓친다!

핵심 아이디어

미래 궤적이 아니라 관성 기준선으로부터의 잔차(residual)를 예측

- ① 결정론적 관성 기준(inertial reference) 을 먼저 계산 — "아무것도 안 하면 갈 경로"
- ② 모델은 궤적 전체가 아니라 그 기준으로부터의 편차(residual)만 예측
- ③ 관성 기준을 통해, 단순 패턴 매칭 대신 "왜 이 경로에서 벗어나야 하는가"(교통 규칙, 장애물 등 인과 요인)에 집중하게
- ④ Point-wise Normalization으로 먼 waypoint의 큰 오차가 학습 신호를 지배하지 않도록 재가중



핵심 기여

- 미래 궤적을 직접 예측하는 대신 관성 **기준으로부터의 정규화된 잔차를 예측하는 Normalized Residual Trajectory Modeling 프레임워크를 제안**해 학습 과제를 크게 단순화했다.
- 관성 기준이 강한 물리적 prior이자 counterfactual 로 작동해, 모델이 spurious correlation 대신 맥락에 기반한 인과 요인(교통 규칙, 장애물)을 학습하도록 강제한다.
- **Point-wise Normalization** 으로 long-horizon의 최적화 불균형을 완화했다.
→ 먼 waypoint의 큰 오차가 학습 신호를 지배하는 것을 방지한다.
- 단 2번의 denoising step만으로 NAVSIM v1 88.8 PDMS, v2 85.5 EPDMS 달성했다.

Table 3. Ablation study on the influence of each component.

Model	Description	NC $\uparrow$	DAC $\uparrow$	EP $\uparrow$	TTC $\uparrow$	C $\uparrow$	PDMS $\uparrow$
$\mathcal{M}_0$	Base Model	97.8	94.2	78.1	93.4	100	84.9
$\mathcal{M}_1$	$\mathcal{M}_0$ + Ranker	98.3	94.3	77.8	94.6	100	85.1
$\mathcal{M}_2$	$\mathcal{M}_1$ + TRM	97.4	96.6	80.3	93.2	100	86.3
$\mathcal{M}_3$	$\mathcal{M}_2$ + PRNorm	97.6	96.7	81.4	93.3	100	87.2
$\mathcal{M}_4$	$\mathcal{M}_3$ + IRP	98.0	97.5	83.4	94.1	100	88.8

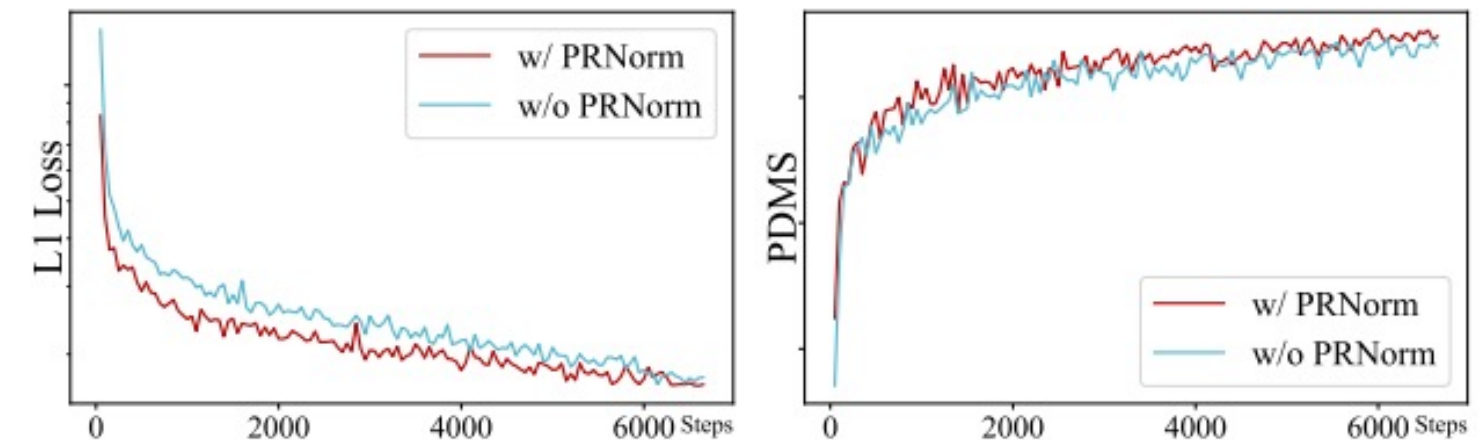


Figure 4. **PRNorm stabilizes optimization and accelerates convergence.** This figure compares the training dynamics of *ResAD* with and without PRNorm.

04

SafeDrive: Fine-Grained Safety Reasoning for End-to-End Driving in a Sparse World

Jungho Kim Jiyong Oh Seunghoon Yu Hongjae Shin Donghyuk Kwak Jun Won Choi[†]
Seoul National University, Republic of Korea

{jhhkim, jyoh, shyu, hjshin, dhkwak}@adr.snu.ac.kr junwchoi@snu.ac.kr

<https://spa-junghokim.github.io/SafeDrive-Page/>

CVPR 2026 Highlight (인용 6회)

<https://arxiv.org/abs/2602.18887>

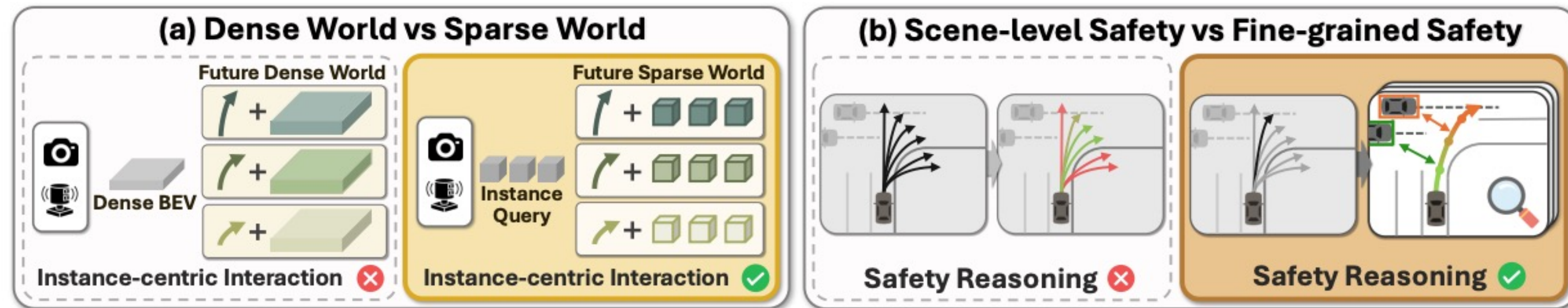
<https://github.com/SPA-junghokim/SafeDrive>

문제 상황

E2E 프레임워크에서의 안전성 확보

E2E 패러다임은 통합적 모델링과 확장성으로 주목받고 있지만, 안전성 보장이 가장 critical한 과제로 남아 있다.

기존 방식의 한계는 두 가지 ① Dense world model은 instance 중심 상호작용을 제대로 모델링하지 못한다.
 ② Scene 수준의 안전 평가는 너무 거칠어서 "어떤 agent가, 언제 위험한지"를 짚어내지 못한다.



"위험하다"가 아니라 "누가, 언제 위험한지"까지 말할 수 있어야 한다!

핵심 아이디어

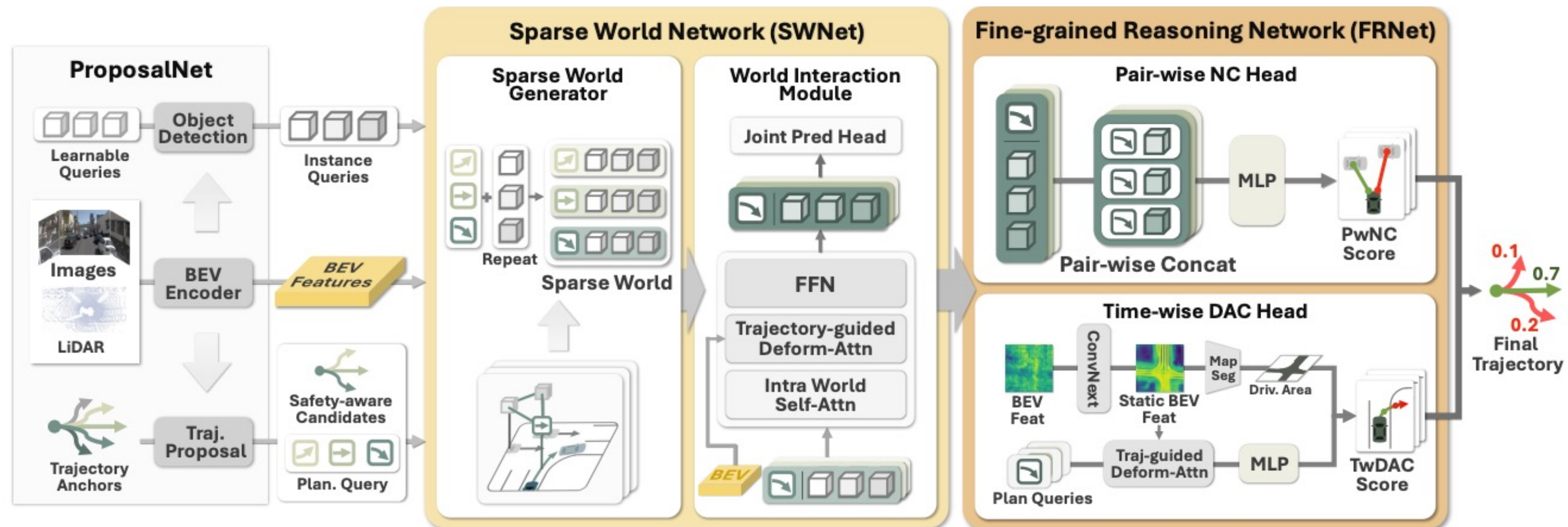
Trajectory-conditioned Sparse World Model로 명시적 안전 추론

① SWNet (Sparse World Network)

- 후보 궤적을 조건으로 삼아 sparse world를 구성
- >> 주요 동적 agent와 도로 요소의 미래 행동을 시뮬레이션해 상호작용 중심 표현을 만든다.

② FRNet (Fine-grained Reasoning Network)

- agent별 충돌 위험(PwNC)과 주행가능영역 준수 여부(TwDAC)를 미래 timestep마다 평가해 안전-critical 사건을 특정한다.



핵심 기여

- Trajectory-conditioned Sparse World Model을 통해 명시적이고 해석 가능한 안전 추론을 수행하는 E2E planning 프레임워크 SafeDrive를 제안했다.
- SWNet은 후보 궤적을 조건으로 주요 동적 agent와 도로 요소의 미래를 시뮬레이션해 상호작용 중심 표현을 제공한다.
- FRNet은 agent별 충돌 위험과 시간별 주행가능영역 준수를 평가해, 미래 timestep 전반에서 안전-critical 사건을 정밀하게 식별한다.
- Open-loop와 closed-loop 양쪽에서 SOTA를 달성했다.
(NAVSIM PDMS 91.6 / EPDMS 87.5, 12,146개 중 충돌 61건(0.5%), Bench2Drive 66.8% Driving Score)

Table 1. **Evaluation of PDMS performance on the NAVSIM dataset.** "C" and "L" denote Camera and LiDAR inputs, respectively. The best and second-best results are highlighted in bold and underline, respectively. Unless noted, models use ResNet-34 as the image backbone, and "†" marks the one using V2-99. The "-" denotes that the associated results are not available.

Method	Venue	Input	NC	DAC	TTC	EP	Comfort	PDMS
Human	-	-	100.0	100.0	100.0	87.5	99.9	94.8
UniAD [9]	CVPR 2023	C	97.8	91.9	92.9	78.8	100	83.4
Transfuser [29]	CVPR 2021	C&L	97.7	92.8	92.8	79.2	100	84.0
PARA-Drive [38]	CVPR 2024	C	97.9	92.4	93.0	79.3	99.8	84.0
LAW [21]	ICLR 2025	C	96.4	95.4	88.7	81.7	<u>99.9</u>	84.6
DistillDrive [45]	ICCV 2025	C&L	98.1	94.6	93.6	81.0	100	86.2
Hydra-MDP [23]	arXiv 2024	C&L	98.3	96.0	94.6	78.7	100	86.5
DiffusionDrive [26]	CVPR 2025	C&L	98.2	96.2	94.7	82.2	100	88.1
WoTE [22]	ICCV 2025	C&L	<u>98.5</u>	96.8	<u>94.9</u>	81.9	<u>99.9</u>	88.3
GaussianFusion [27]	NeurIPS 2025	C&L	98.3	97.2	94.6	83.0	-	88.8
SeerDrive [47]	NeurIPS 2025	C&L	98.4	97.0	<u>94.9</u>	83.2	<u>99.9</u>	88.9
GoalFlow† [41]	CVPR 2025	C&L	98.4	<u>98.3</u>	94.6	85.0	100	<u>90.3</u>
SafeDrive (Ours)	-	C&L	99.5	99.0	97.2	<u>84.3</u>	100	91.6

Table 4. Ablation study for evaluating the main components. "Diff" denotes the DiffusionDrive [26] head, "Prop" indicates the ProposalNet, and "SL Eval" refers to the scene-level safety evaluation.

Diff	Prop	SL Eval	SWNet	FRNet	NC	DAC	TTC	EP	PDMS
✓					98.8	96.3	95.8	82.3	88.6
	✓				98.9	97.1	95.1	83.2	89.1
	✓	✓			99.0	97.7	96.3	83.5	90.1
	✓	✓	✓		99.2	98.2	96.7	84.4	90.9
	✓	✓		✓	99.2	98.5	96.0	84.6	90.9
	✓	✓	✓	✓	99.5	99.0	97.2	84.3	91.6

05

SIMSCALE: Learning to Drive via Real-World Simulation at Scale

Haochen Tian^{1,2,3◇} Tianyu Li^{2†} Haochen Liu³ Jiazhi Yang² Yihang Qiu^{2,3◇}
Guang Li³ Junli Wang^{1,3} Yinfeng Gao^{1,3} Zhang Zhang^{1§} Liang Wang^{1§}
Hangjun Ye³ Tieniu Tan¹ Long Chen^{3§} Hongyang Li²

¹ MAIS, Institute of Automation, Chinese Academy of Sciences

² OpenDriveLab at The University of Hong Kong ³ Xiaomi EV

<https://opendrive-lab.com/SimScale>

CVPR 2026 Oral (인용 26회)

<https://arxiv.org/abs/2511.23369>

<https://github.com/OpenDriveLab/SimScale>

문제 상황

실세계 데이터의 근본적 한계

안전-critical하고 OOD(분포 밖)인 상황은 인간 전문가가 수집한 실세계 데이터에 과소 표현되어 있다.
그런데 이런 희귀한 상호작용이 학습된 정책의 실질적 안전 경계를 결정한다.
문제는 이를 실세계에서 대규모로 수집하기가 극도로 어렵다는 것이다.

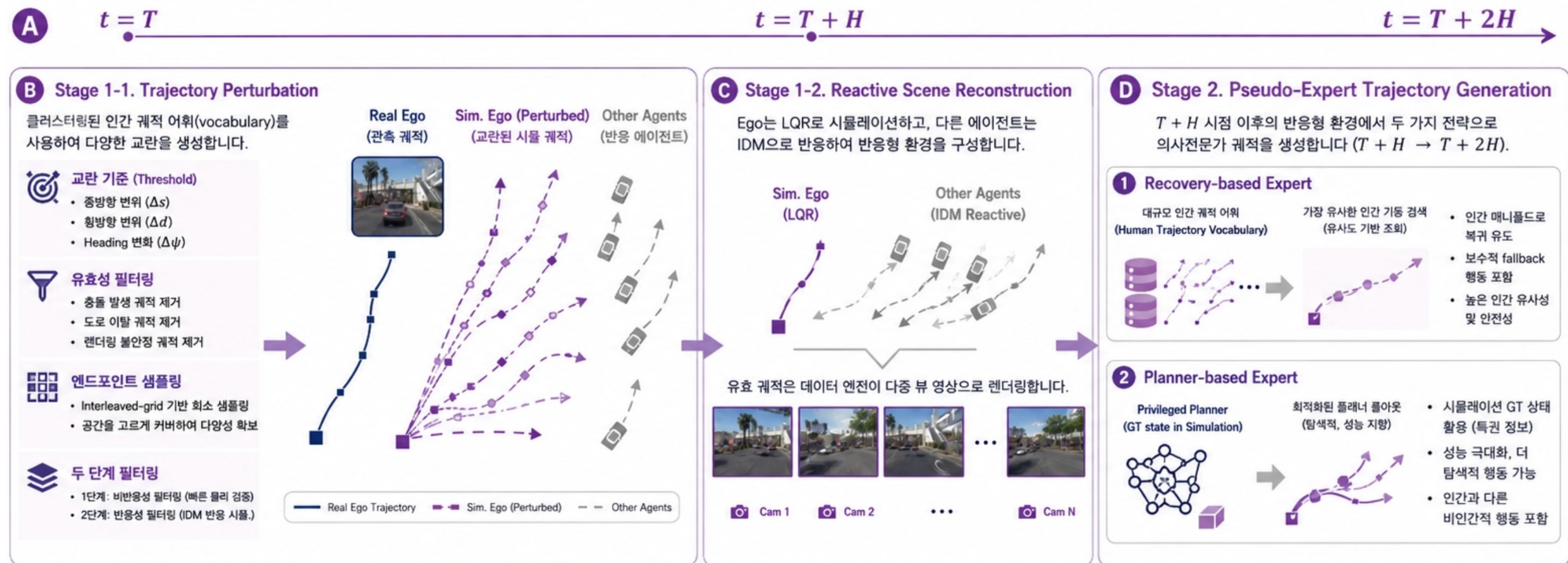


그런 데이터가 부족하면, 의도적인 시뮬레이션으로 만들어내면 되지 않을까?

핵심 아이디어

Real world 로그를 시뮬레이션으로 확장해 co-training

- ① 실세계 주행 로그를 상호작용 가능한 photorealistic 환경으로 재구성
- ② 그 환경에서 안전-critical 변형 시나리오를 대규모 합성
- ③ 실세계 데이터 + 시뮬레이션 데이터를 co-training



핵심 기여

- **실세계 로그를 상호작용 가능한 환경으로 재구성**하고 안전-critical 변형을 대규모 합성하는 real-world simulation 프레임워크를 제안했다.
- 단순 co-training만으로 navhard에서 +6.8 EPDMS, navtest에서 +2.9 향상했다.
- 추가 실세계 데이터 없이 시뮬레이션 데이터만 늘려도 성능이 매끄럽게 스케일링됨을 보였다. (데이터 수집의 병목을 우회할 수 있는 경로를 제시)
- **세 가지 planner 계열(regression, diffusion, vocabulary scoring) 모두에서 일관된 개선을 보였다.**
→ 특정 아키텍처에 종속되지 않는 범용 방법론임을 입증

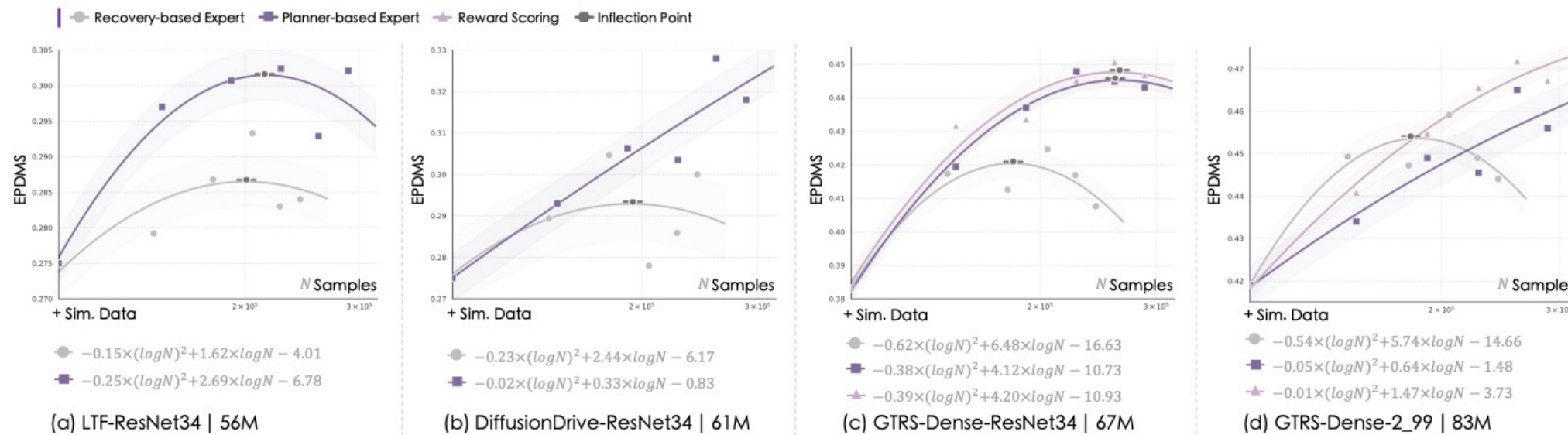


Figure 4. **Scaling dynamics across different planners and pseudo-expert trajectories.** We visualize how simulation data scale and supervision signals influence the driving performance of various planners, where the infection point indicates learning plateau.

Thank you

Question & Discussion



RILS LAB