

当世界模型遇到VLA

范略

中国科学院自动化研究所

世界模型的前世：状态转移



中国科学院自动化研究所
模式识别实验室
New Laboratory of Pattern Recognition

□ 传统意义上的世界模型是对环境状态的建模和状态转移函数的刻画。

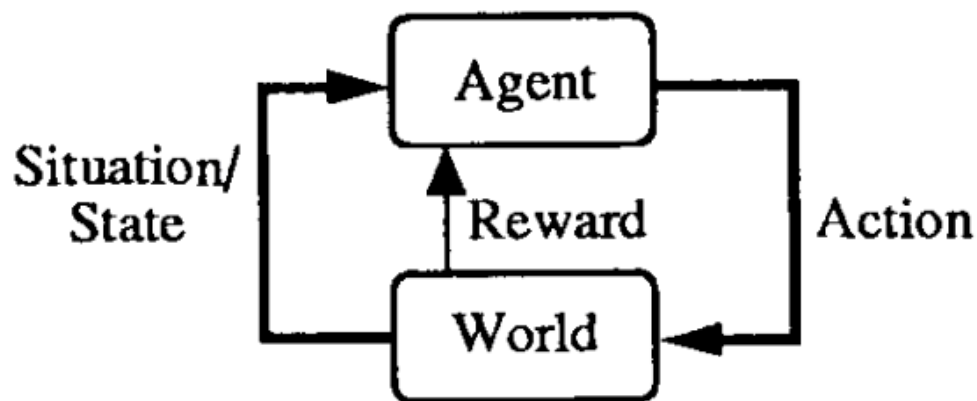


Figure 1: The Problem Formulation Used in Dyna. The agent's object is to maximize the total reward it receives over time.¹

Sutton, Richard (1990). Integrated Architectures for Learning, Planning and Reacting based on Dynamic Programming. Proceedings of the 7th International Workshop on Machine Learning.

世界模型的今生：更加丰富的外延与内涵



中国科学院自动化研究所
模式识别实验室
New Laboratory of Pattern Recognition

世界模型

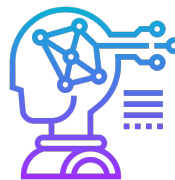


现实世界



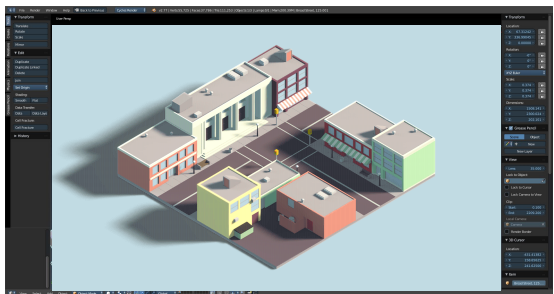
吸收环境反馈
知识交互发现

智能体



场景理解感知
具身规划决策

- 可视化
- 可推理
- 可交互
- 可预测



物理建模仿真



因果逻辑推演



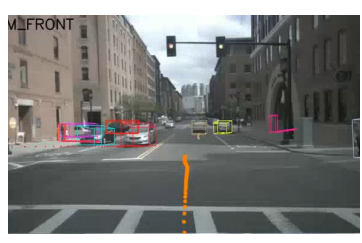
医疗机器人



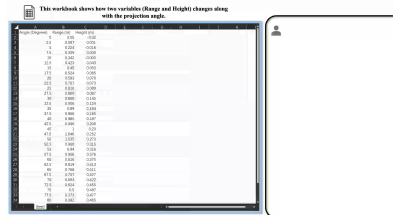
工业自主检修



服务机器人



自动驾驶



网络智能体



自主无人机

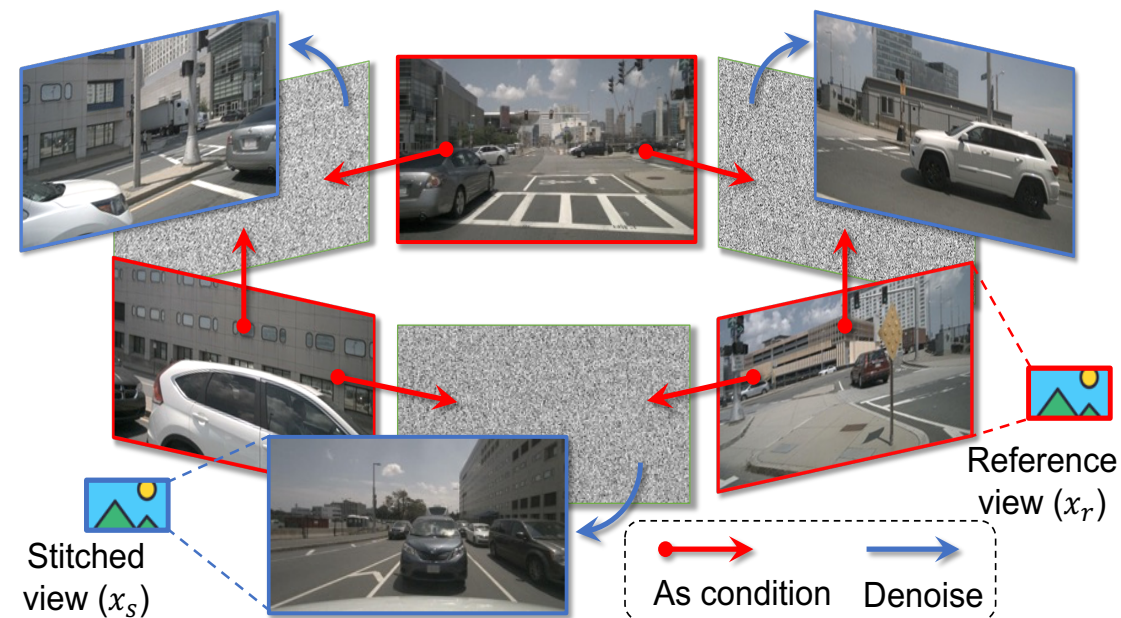
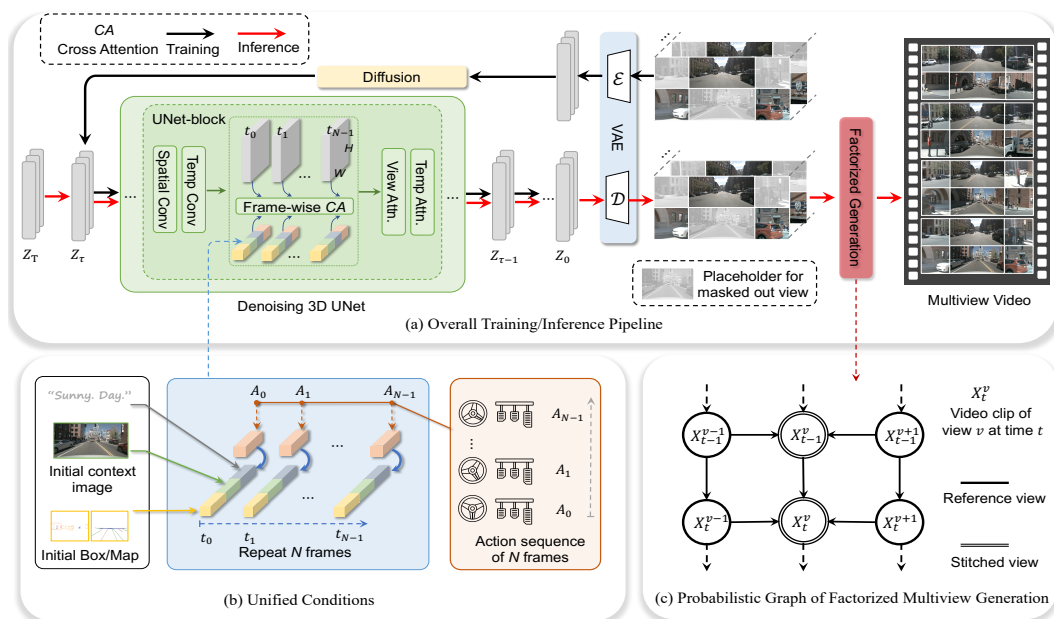
衔接与人类认知的对齐；衔接图文大数据与大模型；衔接普适具身智能应用

生成式世界模型Drive-WM



中国科学院自动化研究所
模式识别实验室
New Laboratory of Pattern Recognition

基于扩散模型的视频生成

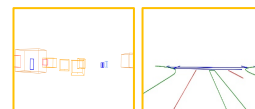


统一的多模态条件控制模块引导自动驾驶场景多视图视频生成模型

文本
控制

"Sunny. Day."

布局
控制



运动
控制



生成式世界模型Drive-WM



中国科学院自动化研究所
模式识别实验室
New Laboratory of Pattern Recognition

□ 基于扩散模型的视频生成：文本控制

晴天



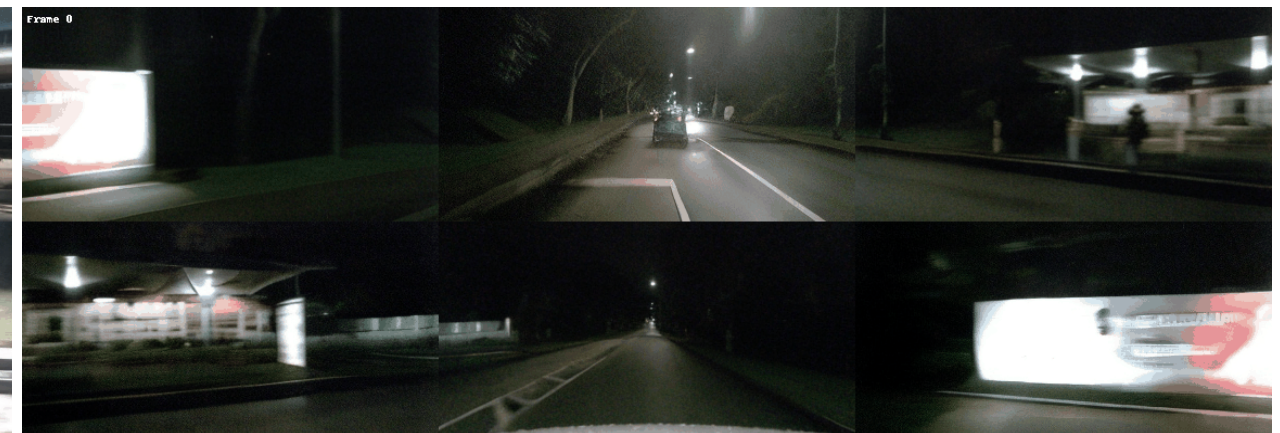
雨天



白天

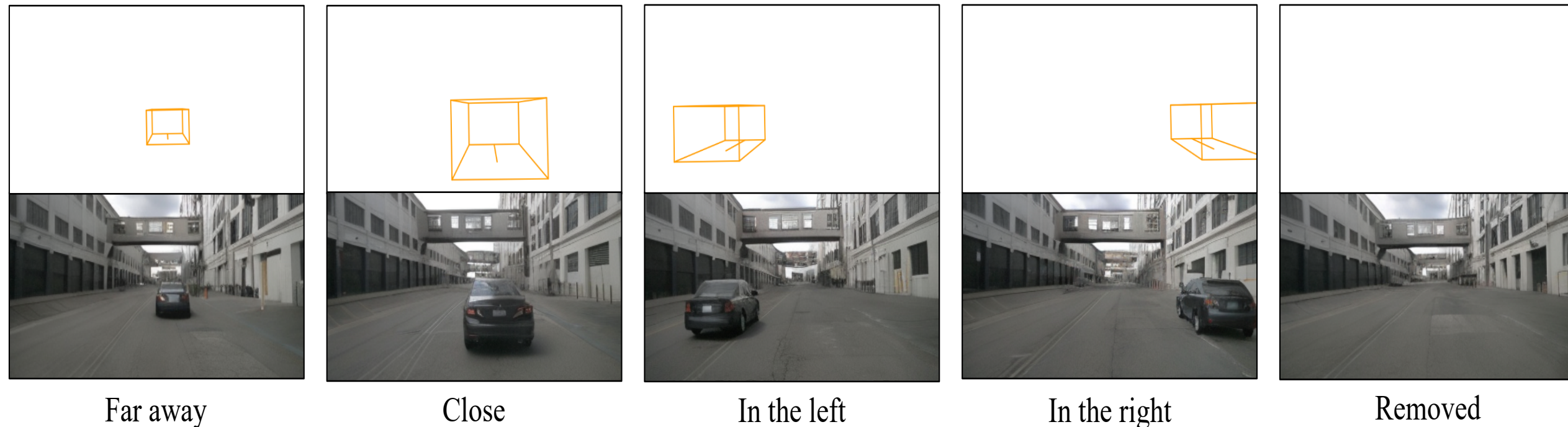


黑夜



□ 基于扩散模型的视频生成：布局控制

○ 3D物体框控制



生成式世界模型Drive-WM



中国科学院自动化研究所
模式识别实验室
New Laboratory of Pattern Recognition

□ 自车运动行为的模拟

白天场景



前进



不同的
驾驶行为



停车



左转



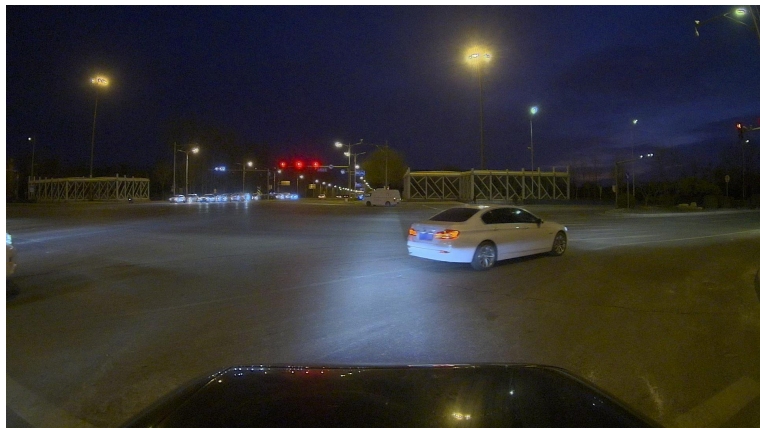
生成式世界模型Drive-WM



中国科学院自动化研究所
模式识别实验室
New Laboratory of Pattern Recognition

□ 自车运动行为的模拟

夜晚场景



前进



不同的
驾驶行为



右转

左转



生成式世界模型Drive-WM



中国科学院自动化研究所
模式识别实验室
New Laboratory of Pattern Recognition

□ 自车运动行为的模拟

驾驶行为的泛化



实现异常驾驶行为的模拟

正常行驶应是向前行驶，控制自车上马路牙子

驾驶场景的泛化



实现驾驶场景的泛化

在其他数据集中实现一些原本不存在的驾驶行为，如倒车

□ 其他道路使用者行为的模拟



初始状态

减速避让：右前车超车



正常行驶：右前车避让



核心挑战：动作控制精度



中国科学院自动化研究所
模式识别实验室
New Laboratory of Pattern Recognition



停止



前进



左转

2m/s or 1.9m/s ?

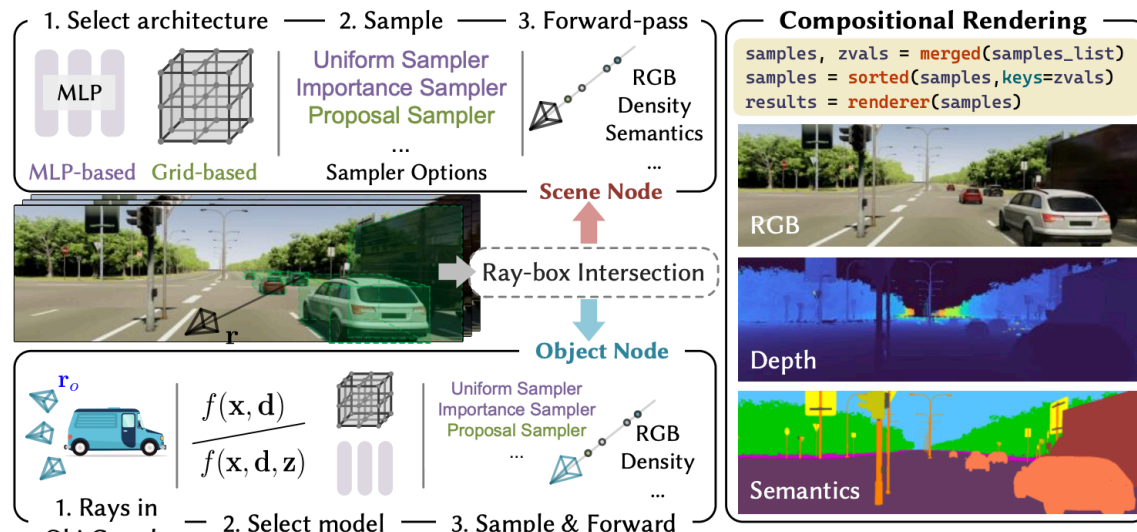
20 deg or 21 deg ?



重建方法的高精度动作控制



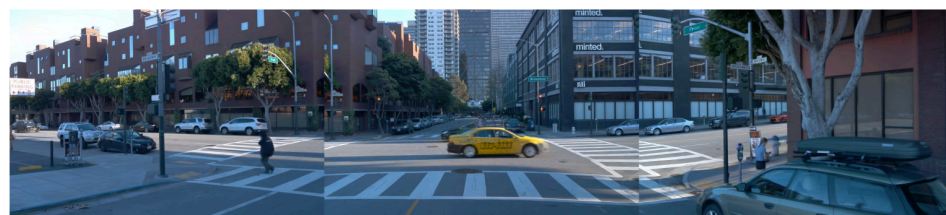
中国科学院自动化研究所
模式识别实验室
New Laboratory of Pattern Recognition



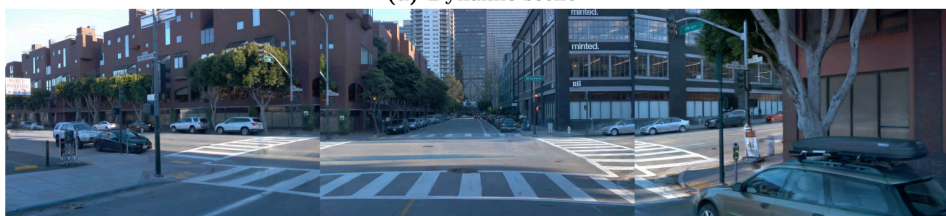
MARS



UniSim

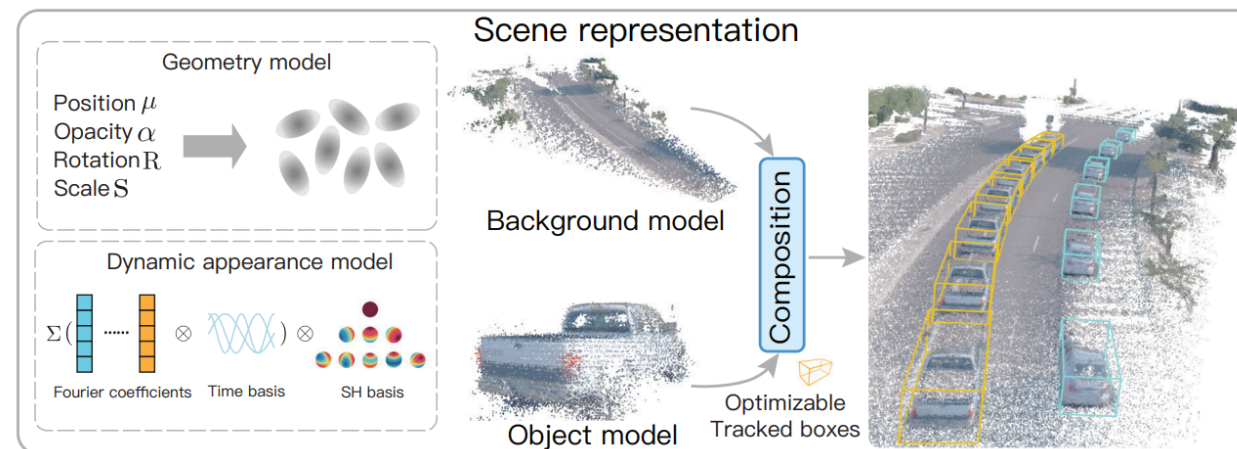


(a) Dynamic scene



(b) Remove the dynamic

PVG



StreetGaussian

空间泛化性缺陷



中国科学院自动化研究所
模式识别实验室
New Laboratory of Pattern Recognition



本质原因

无轨迹外数据

- 大范围视角变换质量仍旧受限
- 严重遮挡区域无法解决

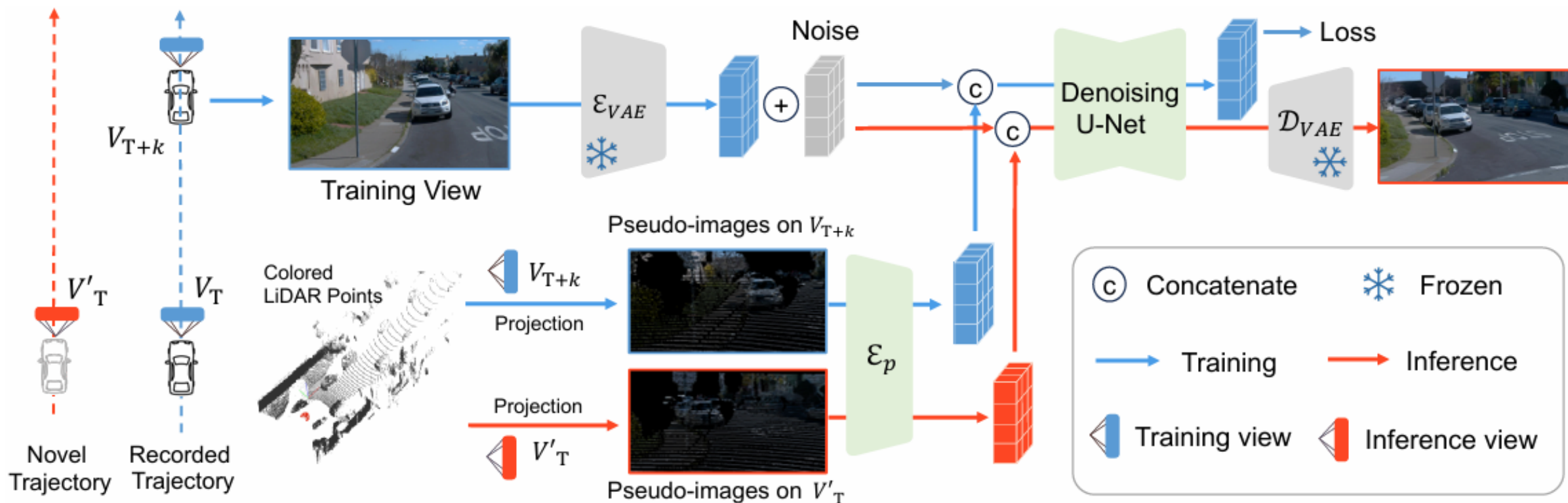
重建结合生成？

—— *几何结构 + 纹理表观*

生成重建一体化：FreeVS



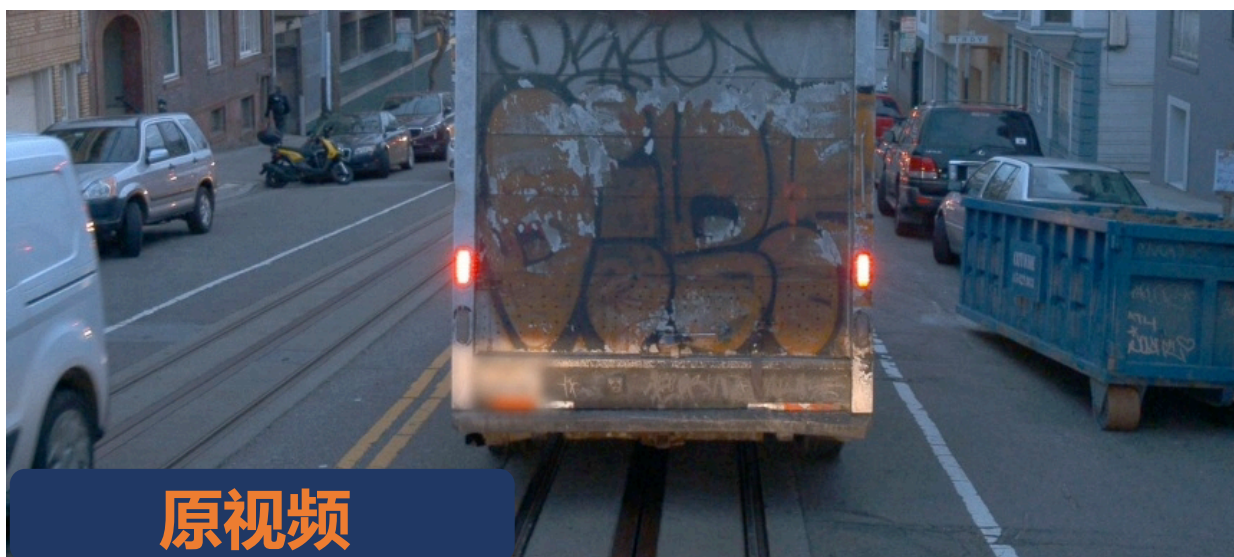
中国科学院自动化研究所
模式识别实验室
New Laboratory of Pattern Recognition



生成重建一体化：FreeVS



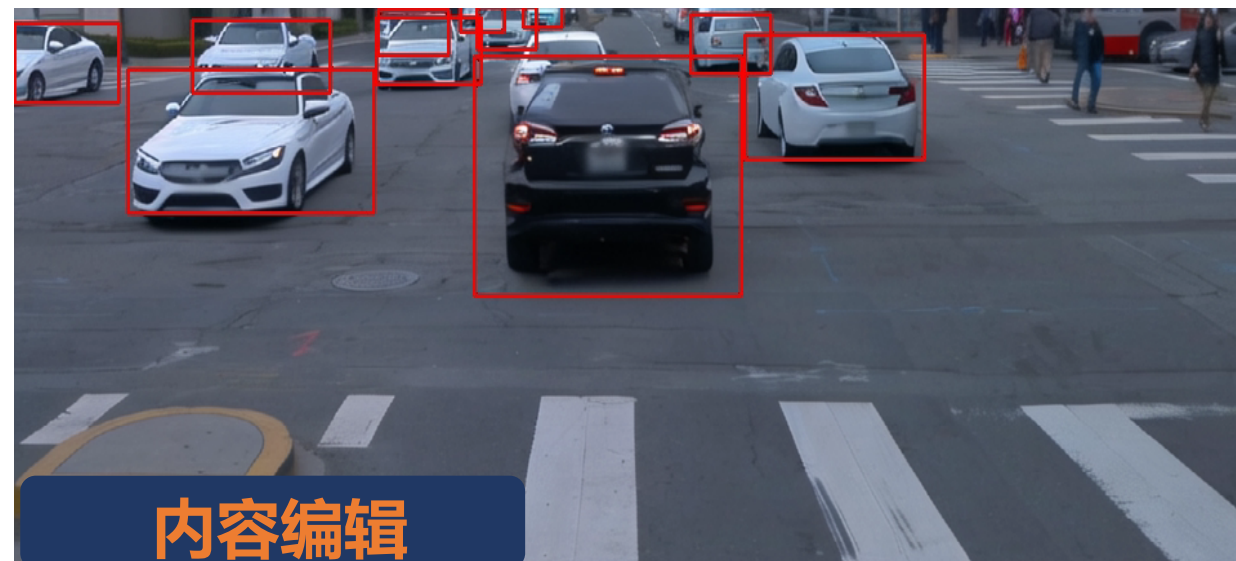
中国科学院自动化研究所
模式识别实验室
New Laboratory of Pattern Recognition



生成重建一体化：FreeVS



中国科学院自动化研究所
模式识别实验室
New Laboratory of Pattern Recognition



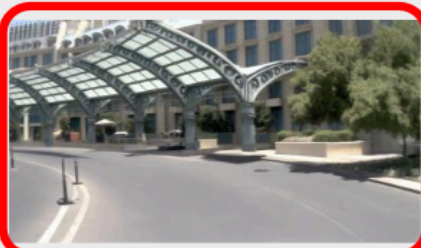
生成重建一体化：FreeVS



中国科学院自动化研究所
模式识别实验室
New Laboratory of Pattern Recognition

□ 利用世界模型合成的各类Corner Case

Diverse
Backgrounds



Synthetic Normal Scenes



Synthetic Corner Scenes



Wrong-way Vehicle



Obstructing Vehicle



Aggressive Lane-cutting Vehicles



Irregular Vehicle

生成重建一体化：FreeVS



中国科学院自动化研究所
模式识别实验室
New Laboratory of Pattern Recognition

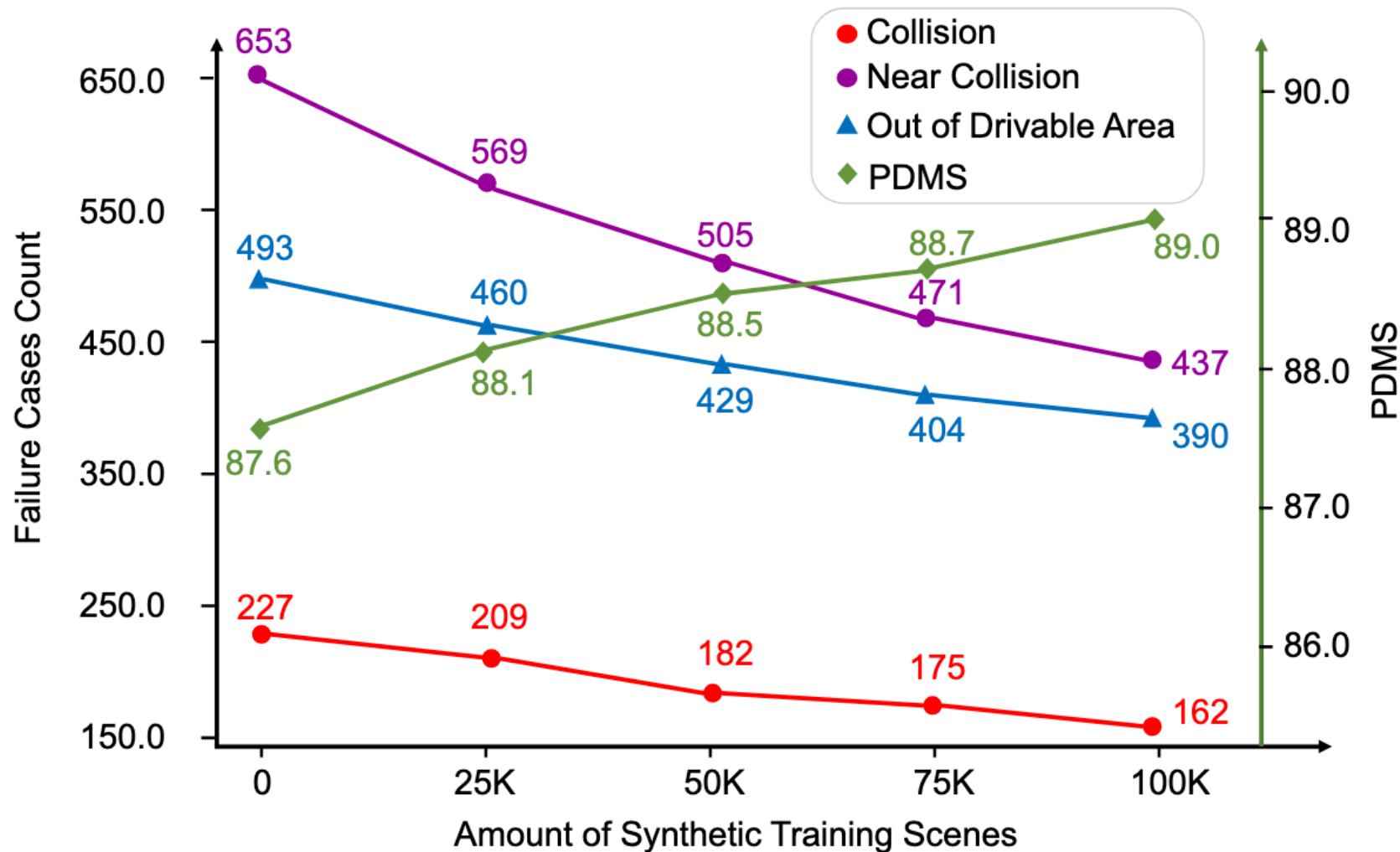
□ 利用世界模型合成的各类Corner Case



生成重建一体化：FreeVS



中国科学院自动化研究所
模式识别实验室
New Laboratory of Pattern Recognition

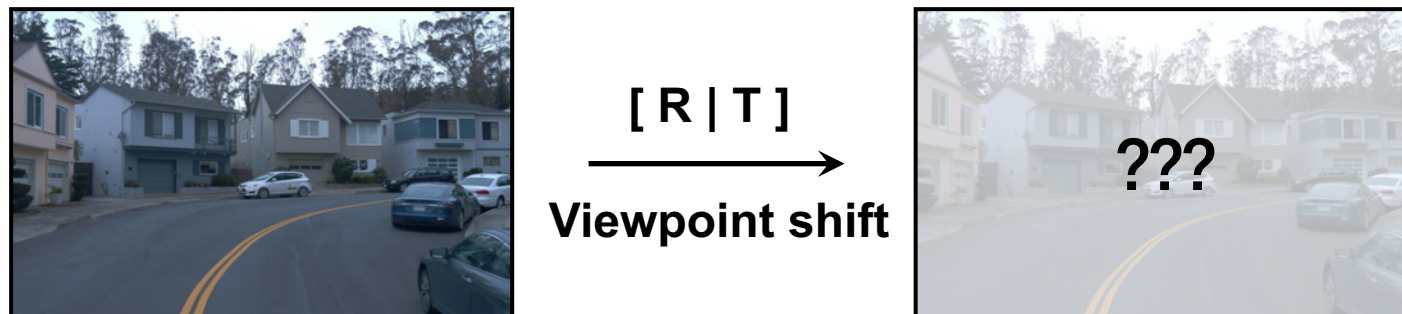


随着合成数据增加，各类failure case迅速减少

生成重建一体化：FreeSim



中国科学院自动化研究所
模式识别实验室
New Laboratory of Pattern Recognition



❑ 由于数据缺失，直接以姿态变换为条件的生成方式不可行



轨迹外渲染



图像增强



生成重建一体化：FreeSim



中国科学院自动化研究所
模式识别实验室
New Laboratory of Pattern Recognition

样本构造



degrade
→



生成重建一体化：FreeSim



中国科学院自动化研究所
模式识别实验室
New Laboratory of Pattern Recognition

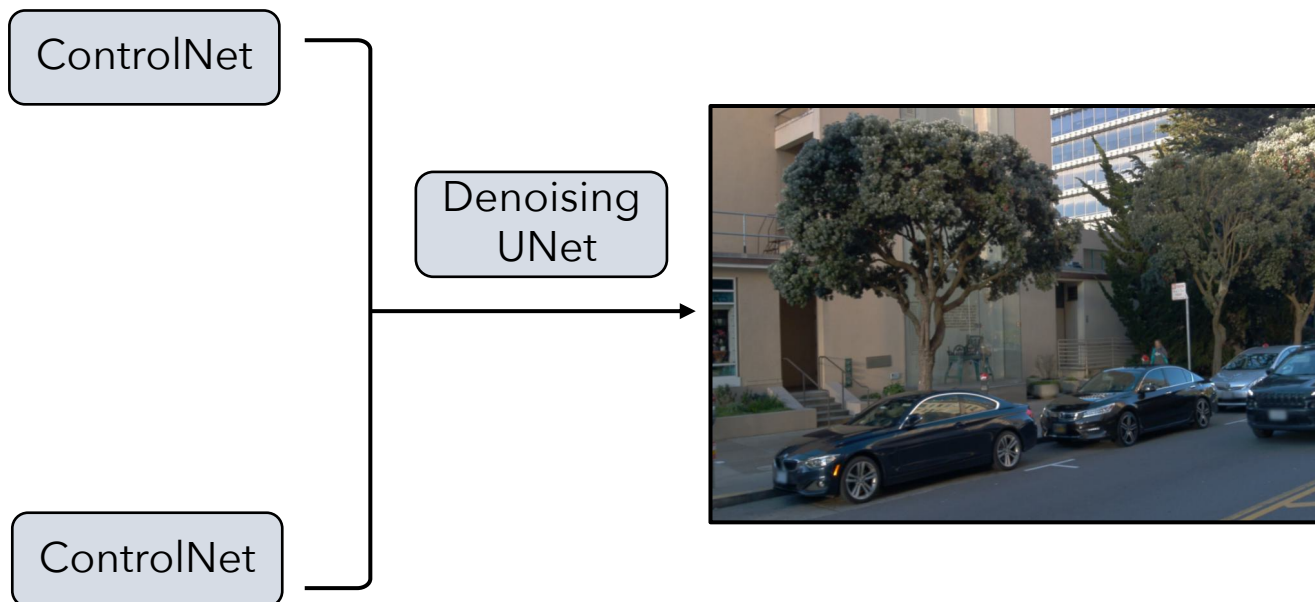
生成框架



Degraded rendering



Sparse LiDAR projection

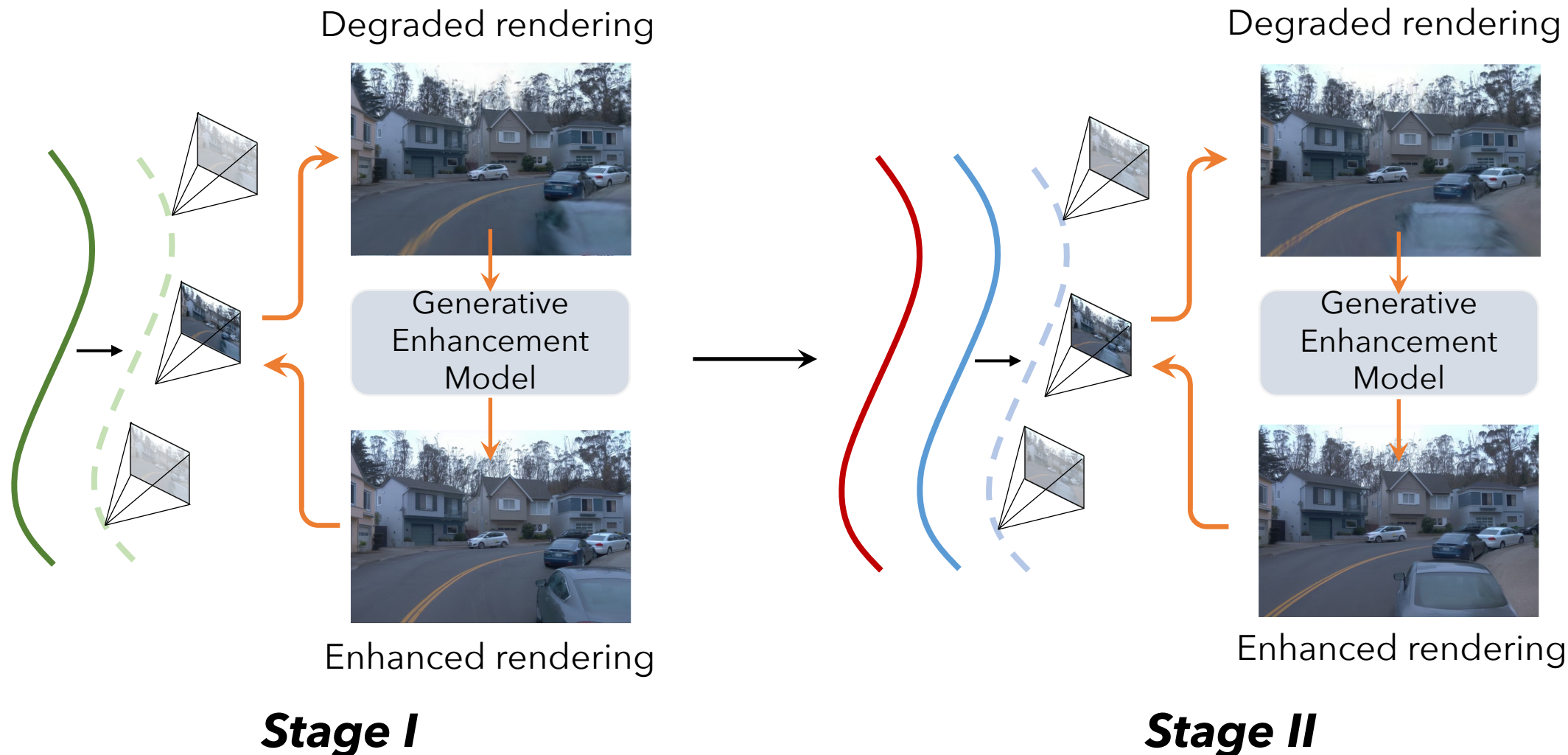


生成重建一体化：FreeSim



中国科学院自动化研究所
模式识别实验室
New Laboratory of Pattern Recognition

渐进式重建



生成重建一体化：FreeSim



中国科学院自动化研究所
模式识别实验室
New Laboratory of Pattern Recognition

Original viewpoint



1-meter left



2-meter left



3-meter left



Step size 1.5 m



Step size 3 m
(non-progressive)



生成重建一体化：FreeSim



中国科学院自动化研究所
模式识别实验室
New Laboratory of Pattern Recognition



PVG



StreetGaussian



S3Gaussian



FreeSim (ours)

生成重建一体化：FreeSim



中国科学院自动化研究所
模式识别实验室
New Laboratory of Pattern Recognition



PVG



StreetGaussian



S3Gaussian



FreeSim (ours)

生成重建一体化：FreeSim



中国科学院自动化研究所
模式识别实验室
New Laboratory of Pattern Recognition



PVG



StreetGaussian



S3Gaussian



FreeSim (ours)

生成重建一体化：FreeSim

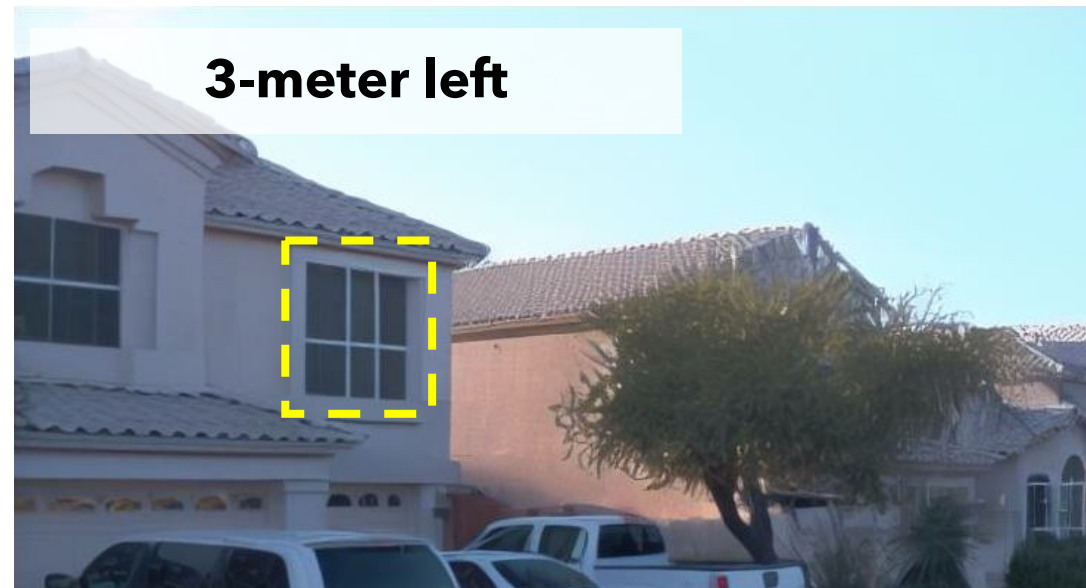


中国科学院自动化研究所
模式识别实验室
New Laboratory of Pattern Recognition

GT (front view)



3-meter left



GT (front right view)



3-meter right



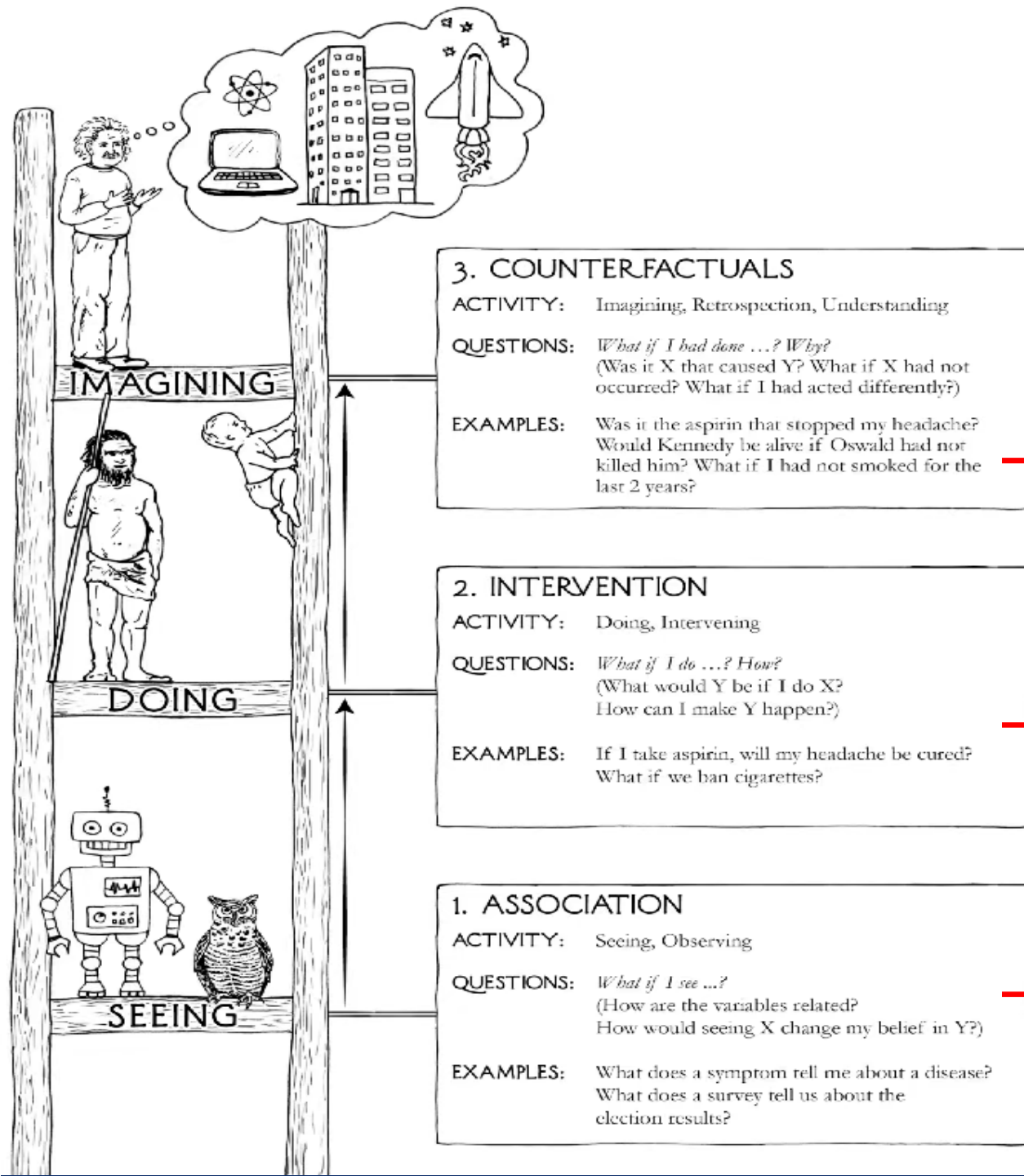
当世界模型遇到VLA



中国科学院自动化研究所
模式识别实验室
New Laboratory of Pattern Recognition

Everything about the World Model seems perfect. However, the autonomous driving community has begun debating: ***World Model or VLA?***





In this debate, someone believes that world model goes beyond VLAs.

Is that true?

World model
(imaging, modeling, reasoning)

VLA
(doing, mimicking)

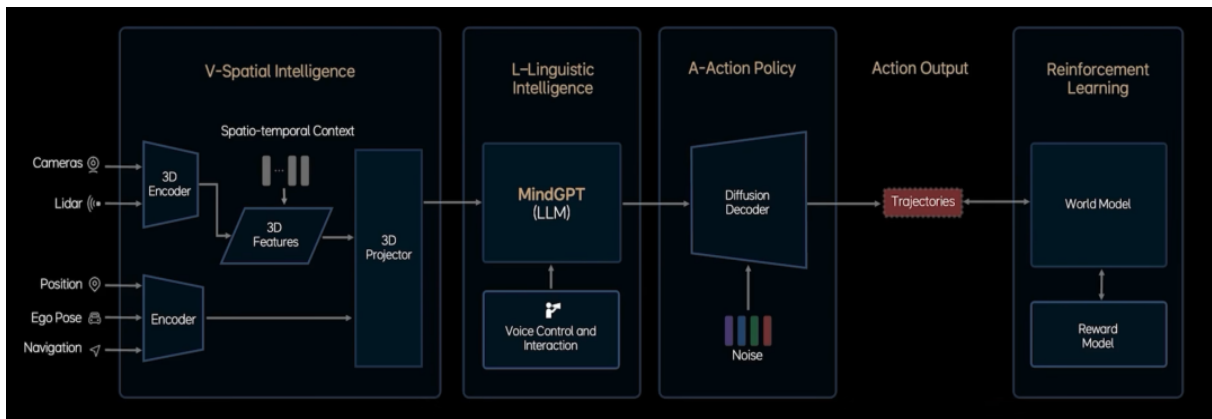
Perception models
(seeing, perception)

当世界模型遇到VLA



中国科学院自动化研究所
模式识别实验室
New Laboratory of Pattern Recognition

巨头的押注



Li Auto Mind-VLA

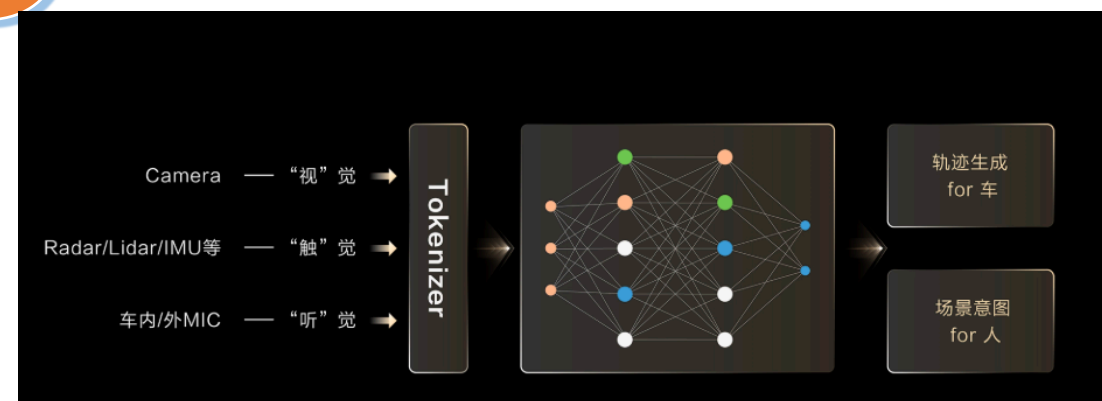
- Input: Vision / Language; Output: Action
- Usually Autoregression Backbone with additional MoE Action Decoder

VLA

V/S

World Model

- Action-conditioned State transition
- Simulation
- Action-conditioned Video Generation
- ...



Huawei World Model

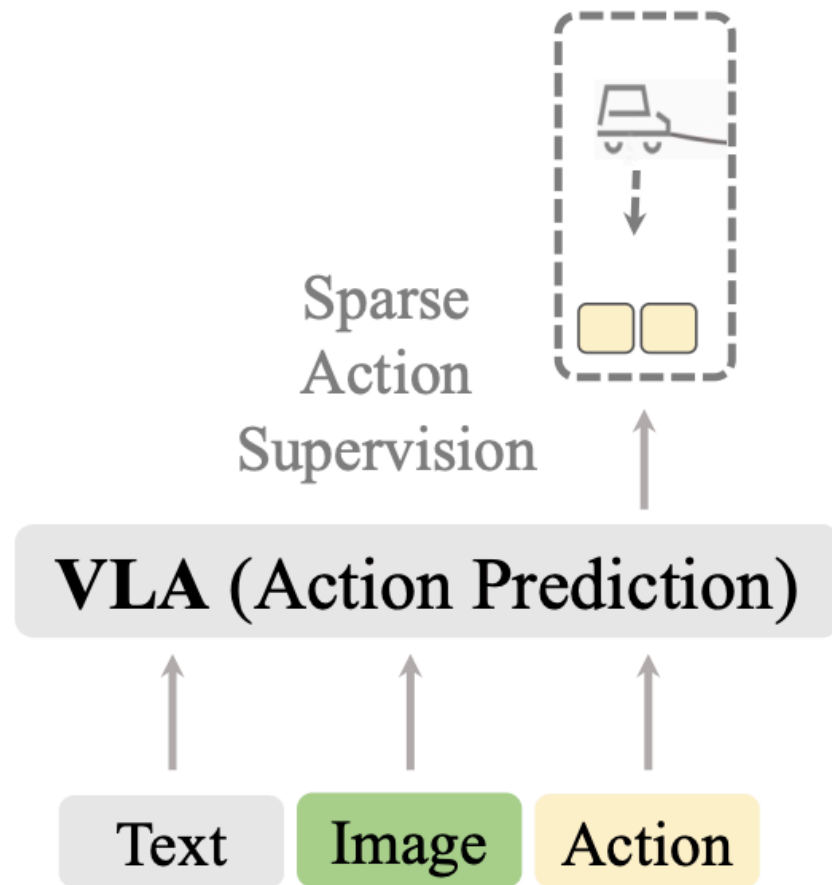
❑ 世界模型和VLA并非对齐

- World Model构建仿真环境训练VLA
- VLA显式预测未来，辅助决策
- World Model for VLA posting training (RL)
- ...

□ VLA最需要什么

- VLA 模型容量极大，需要丰富的密集监督信号
- 驾驶动作数据天然稀疏

世界模型提供密集监督？



表征学习中的世界模型



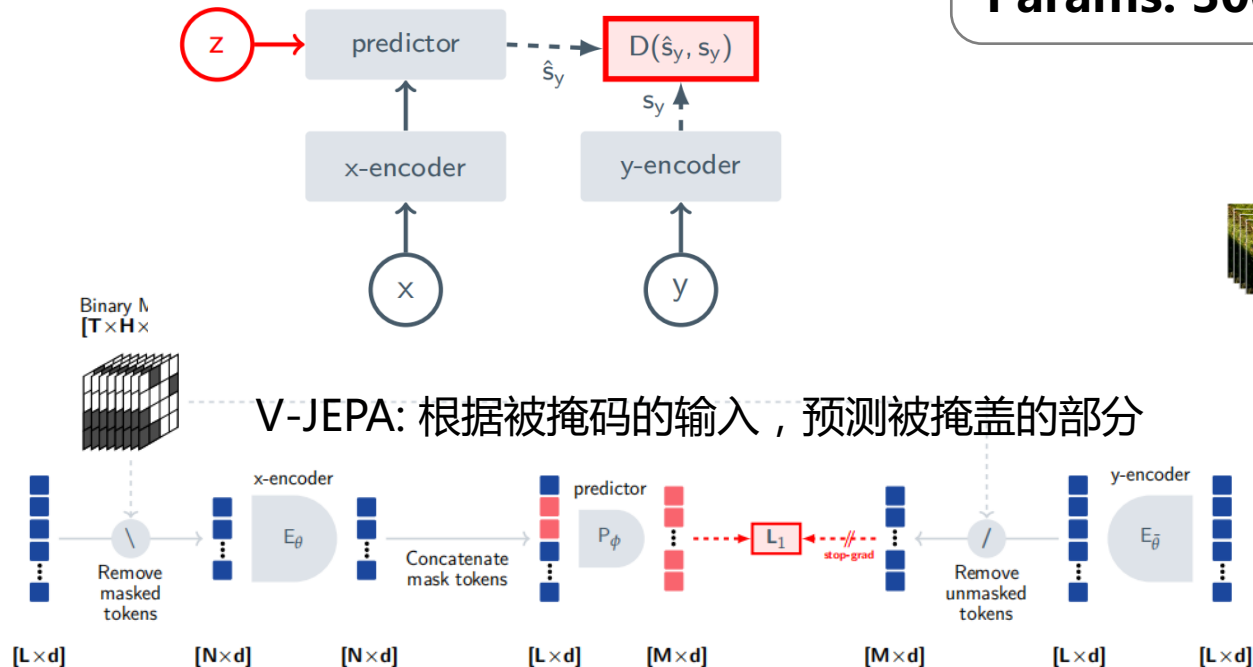
中国科学院自动化研究所
模式识别实验室
New Laboratory of Pattern Recognition

利用表征学习规避合成数据的bias

V-JEPA

基于海量视频数据，
利用**隐空间**进行**掩码**学习

JEPA: 在**隐空间**求损失



LeCun对自监督的坚持

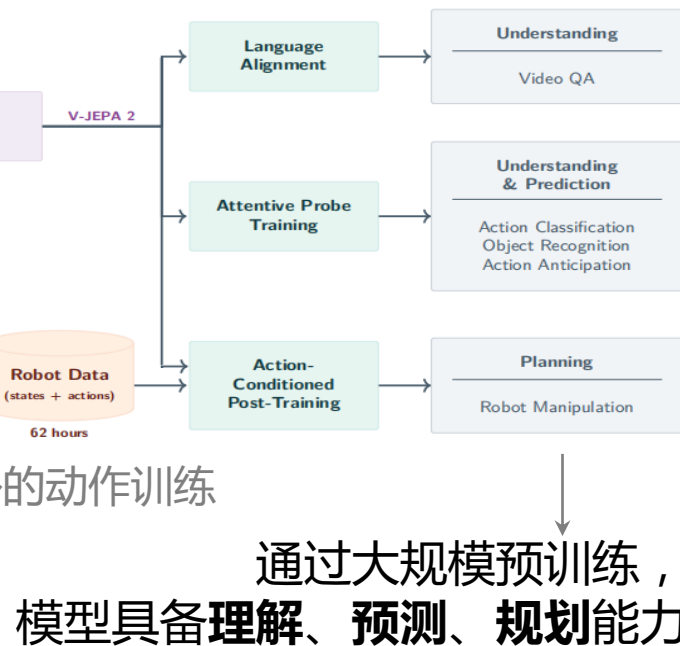
Videos: 2M to 22M
Iters: 90K to 252K
Params: 300M to 1B+



引入额外的动作训练

V-JEPA 2

更多数据，更大模型，
更长训练时间
最强Video QA 8B模型

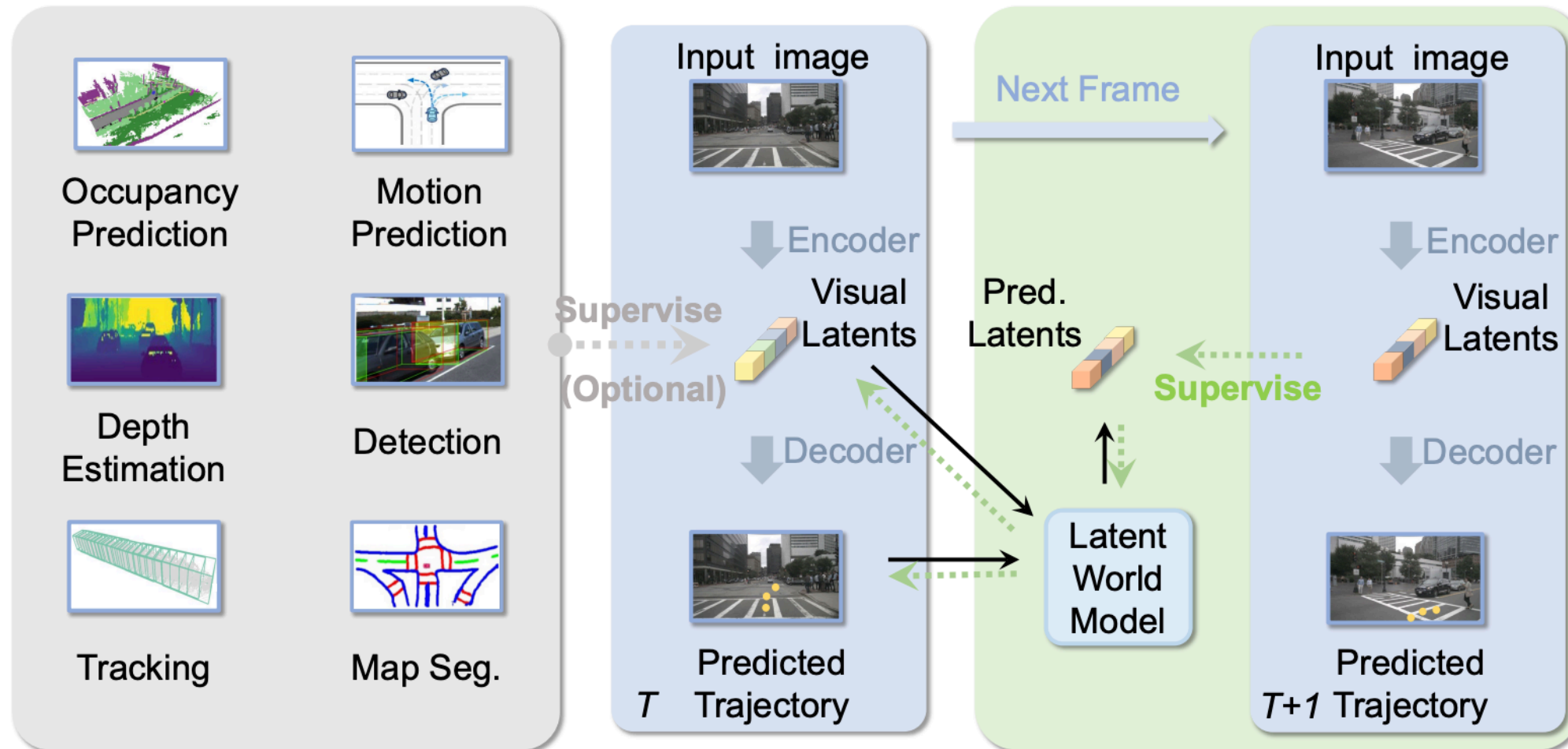


表征学习中的世界模型



中国科学院自动化研究所
模式识别实验室
New Laboratory of Pattern Recognition

□ 预测未来表征摆脱感知监督依赖



(a) Annotation-based Perception Tasks

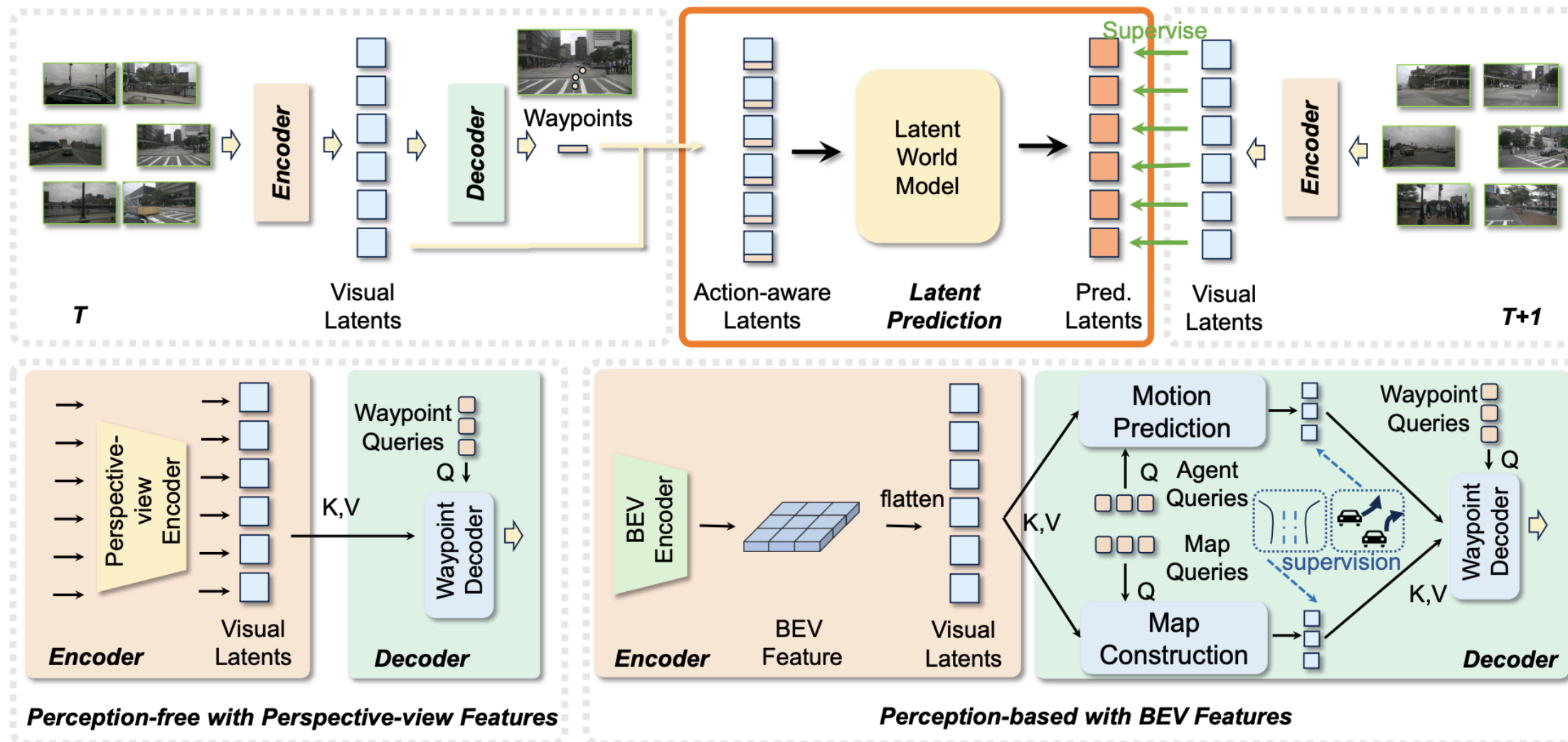
(b) Our Self-supervised Task

表征学习中的世界模型



中国科学院自动化研究所
模式识别实验室
New Laboratory of Pattern Recognition

预测未来表征摆脱感知监督依赖



表征学习中的世界模型



中国科学院自动化研究所
模式识别实验室
New Laboratory of Pattern Recognition



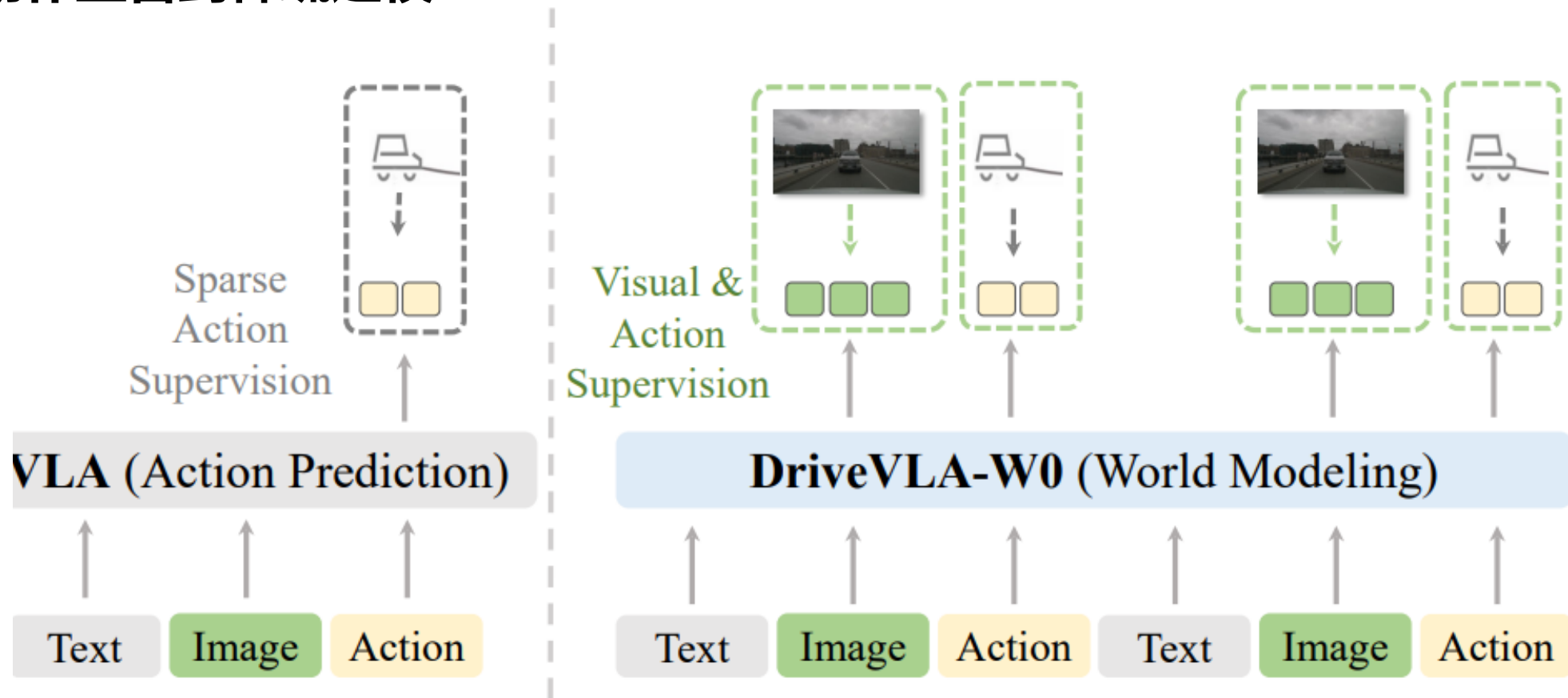
当世界模型遇到VLA



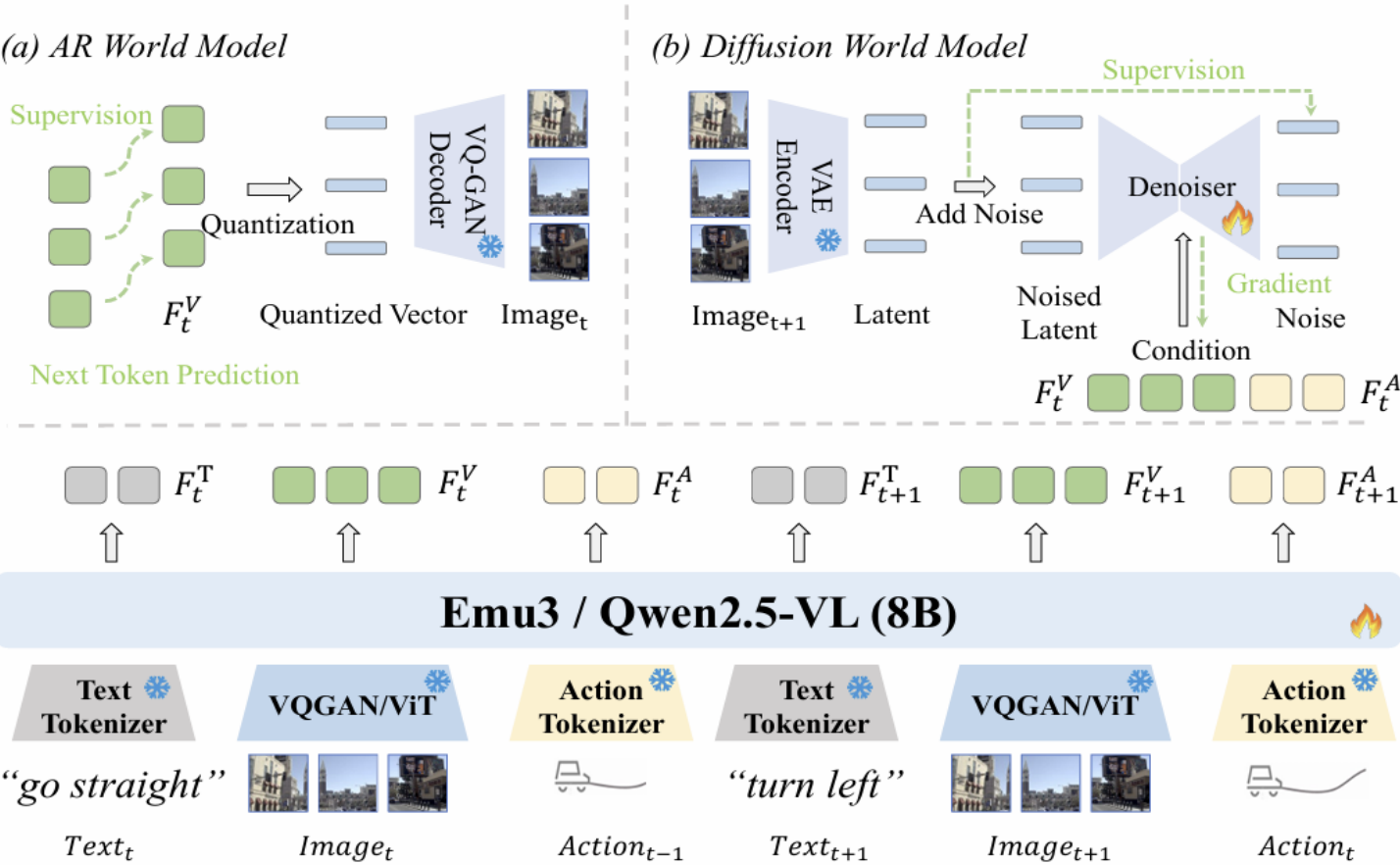
中国科学院自动化研究所
模式识别实验室
New Laboratory of Pattern Recognition

DriveVLA-W0: World Models Amplify Data Scaling Law in Autonomous Driving

□ 从动作监督到稀疏建模



整体框架



数据

7000万视觉帧 / 100万+ clips

架构

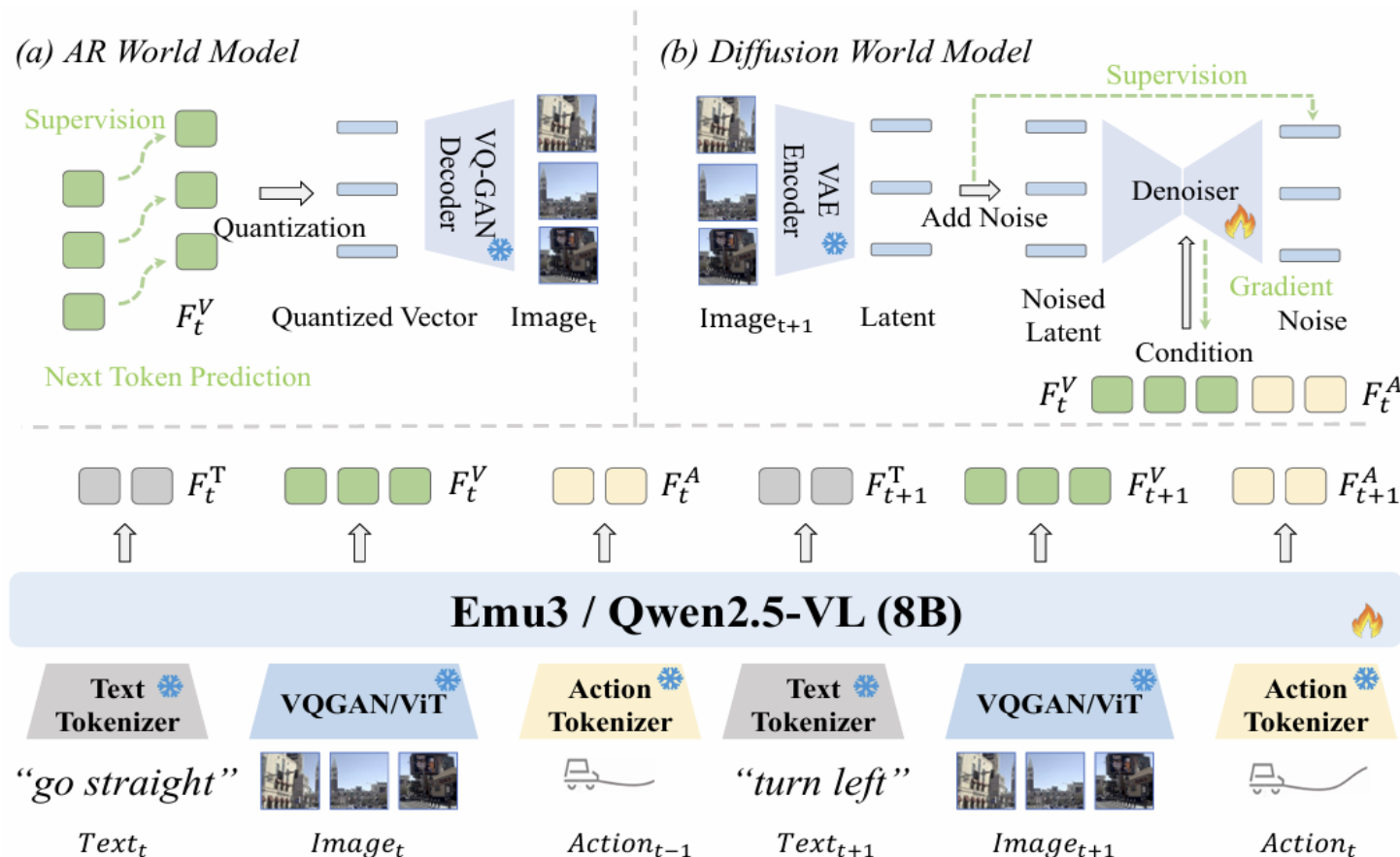
VLA + 高效混合专家架构

训练

World modeling 预训练

新范式，面向
下一代自动驾驶技术

□ 整体框架

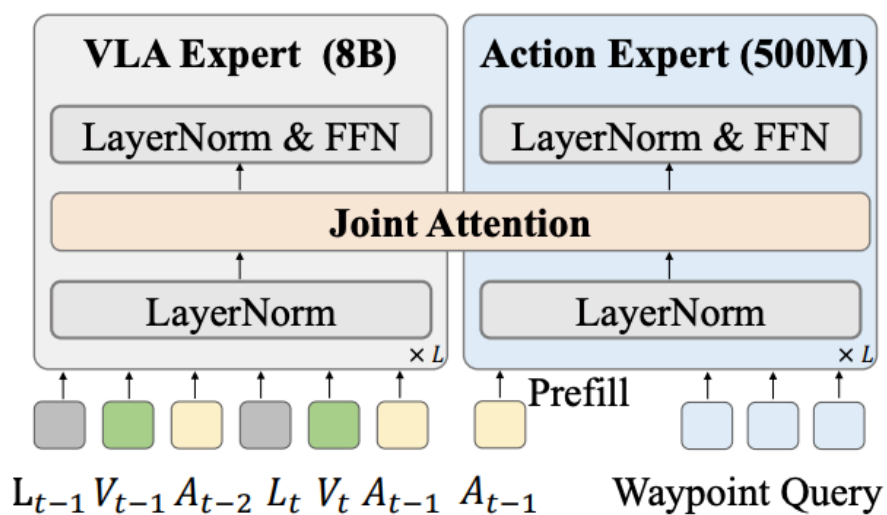


多种变体验证普适性

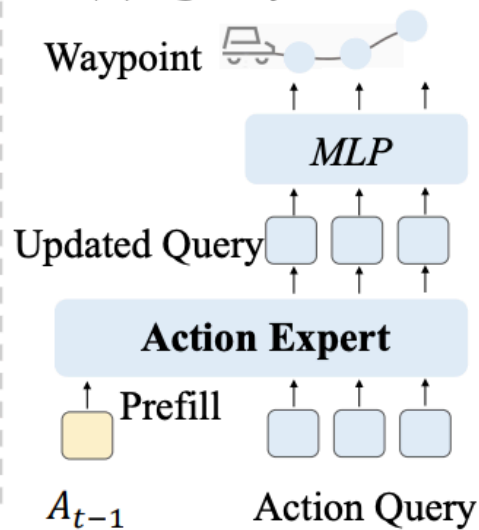
- 基础模型
Emu3 and Qwen2.5
- 世界模型实现
AR and Diffusion

□ 多种动作解码器设计

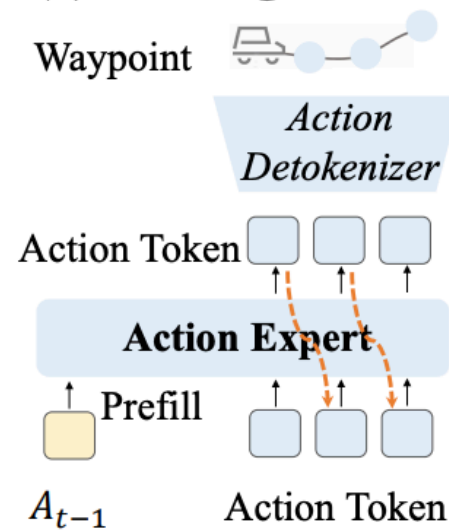
(a) MoE Architecture



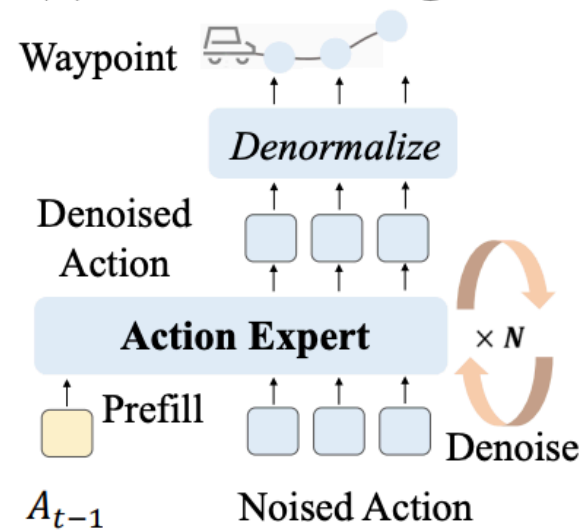
(b) Query-based



(c) Autoregressive



(d) Flow Matching



□ NavSim v1/v2 上实现最佳性能

Method	Ref	Sensors	NC ↑	DAC ↑	TTC ↑	C. ↑	EP ↑	PDMS ↑
Human	-	-	100	100	100	99.9	87.5	94.8
<i>BEV-based Methods</i>								
UniAD (Hu et al., 2022b)	CVPR'23	6x Cam	97.8	91.9	92.9	100.0	78.8	83.4
TransFuser (Prakash et al., 2021)	TPAMI'23	3x Cam + L	97.7	92.8	92.8	100.0	79.2	84.0
PARA-Drive (Weng et al., 2024)	CVPR'24	6x Cam	97.9	92.4	93.0	99.8	79.3	84.0
LAW (Li et al., 2024a)	ICLR'25	1x Cam	96.4	95.4	88.7	99.9	81.7	84.6
Hydra-MDP (Li et al., 2024b)	arXiv'24	3x Cam + L	98.3	96.0	94.6	100.0	78.7	86.5
DiffusionDrive (Liao et al., 2025)	CVPR'25	3x Cam + L	98.2	96.2	94.7	100.0	82.2	88.1
WoTE (Li et al., 2025a)	ICCV'25	3x Cam + L	98.5	96.8	94.4	99.9	81.9	88.3
<i>VLA-based Methods</i>								
AutoVLA (Zhou et al., 2025c)	NeurIPS'25	3x Cam	98.4	95.6	98.0	99.9	81.9	89.1
ReCogDrive (Li et al., 2025b)	arXiv'25	3x Cam	98.2	97.8	95.2	99.8	83.5	89.6
DriveVLA-W0*	-	1x Cam	98.7	99.1	95.3	99.3	83.3	90.2
AutoVLA† (Zhou et al., 2025c)	NeurIPS'25	3x Cam	99.1	97.1	97.1	100.0	87.6	92.1
DriveVLA-W0†	-	1x Cam	99.3	97.4	97.0	99.9	88.3	93.0

NavSim v1

NavSim v2

Method	NC ↑	DAC ↑	DDC ↑	TLC ↑	EP ↑	TTC ↑	LK ↑	HC ↑	EC ↑	EPDMS ↑
Ego Status	93.1	77.9	92.7	99.6	86.0	91.5	89.4	98.3	85.4	64.0
TransFuser (Prakash et al., 2021)	96.9	89.9	97.8	99.7	87.1	95.4	92.7	98.3	87.2	76.7
HydraMDP++ (Li et al., 2024b)	97.2	97.5	99.4	99.6	83.1	96.5	94.4	98.2	70.9	81.4
DriveSuprem (Yao et al., 2025)	97.5	96.5	99.4	99.6	88.4	96.6	95.5	98.3	77.0	83.1
ARTEMIS (Feng et al., 2025)	98.3	95.1	98.6	99.8	81.5	97.4	96.5	98.3	-	83.1
DiffusionDrive (Liao et al., 2025)	98.2	95.9	99.4	99.8	87.5	97.3	96.8	98.3	87.7	84.5
DriveVLA-W0	98.5	99.1	98.0	99.7	86.4	98.1	93.2	97.9	58.9	86.1

❑ 在超7000万视觉帧和100万场景的内部数据验证有效性

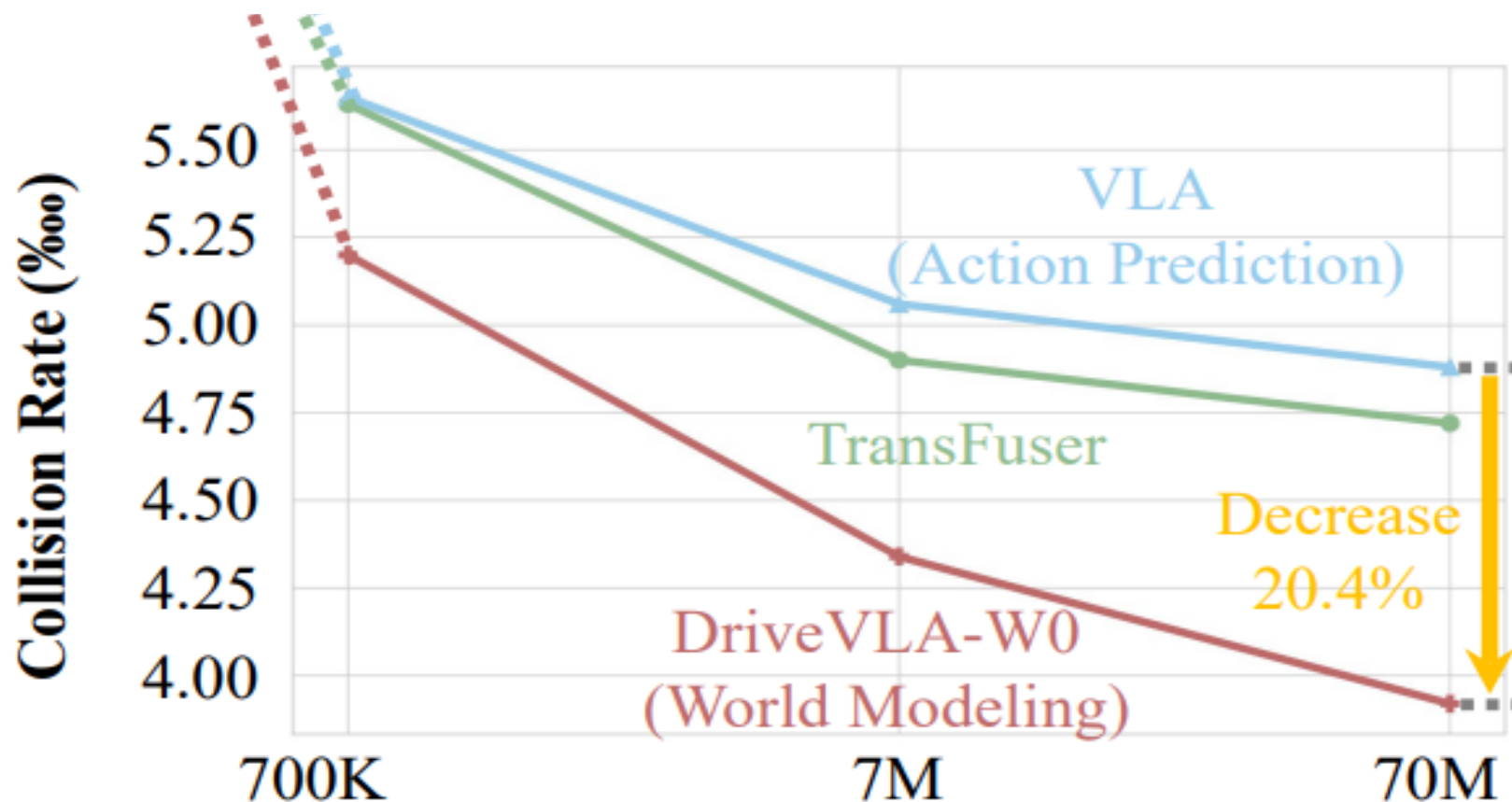
Model	In-house Dataset Scale (Number of Frames)					
	70k		700k		70M	
	ADE (m) ↓	Collision (%) ↓	ADE (m) ↓	Collision (%) ↓	ADE (m) ↓	Collision (%) ↓
TransFuser-50M	2.5893	0.0894	1.7464	0.0563	1.2627	0.0472
TransFuser-7B	2.5757	0.0839	2.1391	0.0710	1.2244	0.0539
VLA (VQ) (Baseline)	2.8520	0.0982	1.5424	0.0565	1.4829	0.0488
+ World Model (Ours)	2.7482 ↑3.6%	0.0956 ↑2.7%	1.5985 ↓3.6%	0.0520 ↑8.0%	1.0563 ↑28.8%	0.0392 ↑19.7%
VLA (ViT) (Baseline)	3.1524	0.0950	1.4202	0.0462	1.1051	0.0359
+ World Model (Ours)	2.5268 ↑19.9%	0.0834 ↑12.2%	1.3436 ↑5.4%	0.0513 ↓11.0%	1.0640 ↑3.7%	0.0302 ↑15.9%

当世界模型遇到VLA：结论（1）



中国科学院自动化研究所
模式识别实验室
New Laboratory of Pattern Recognition

□ 两者结合展现出相比 vanilla VLAs BEV端到端模型更好的data scalability

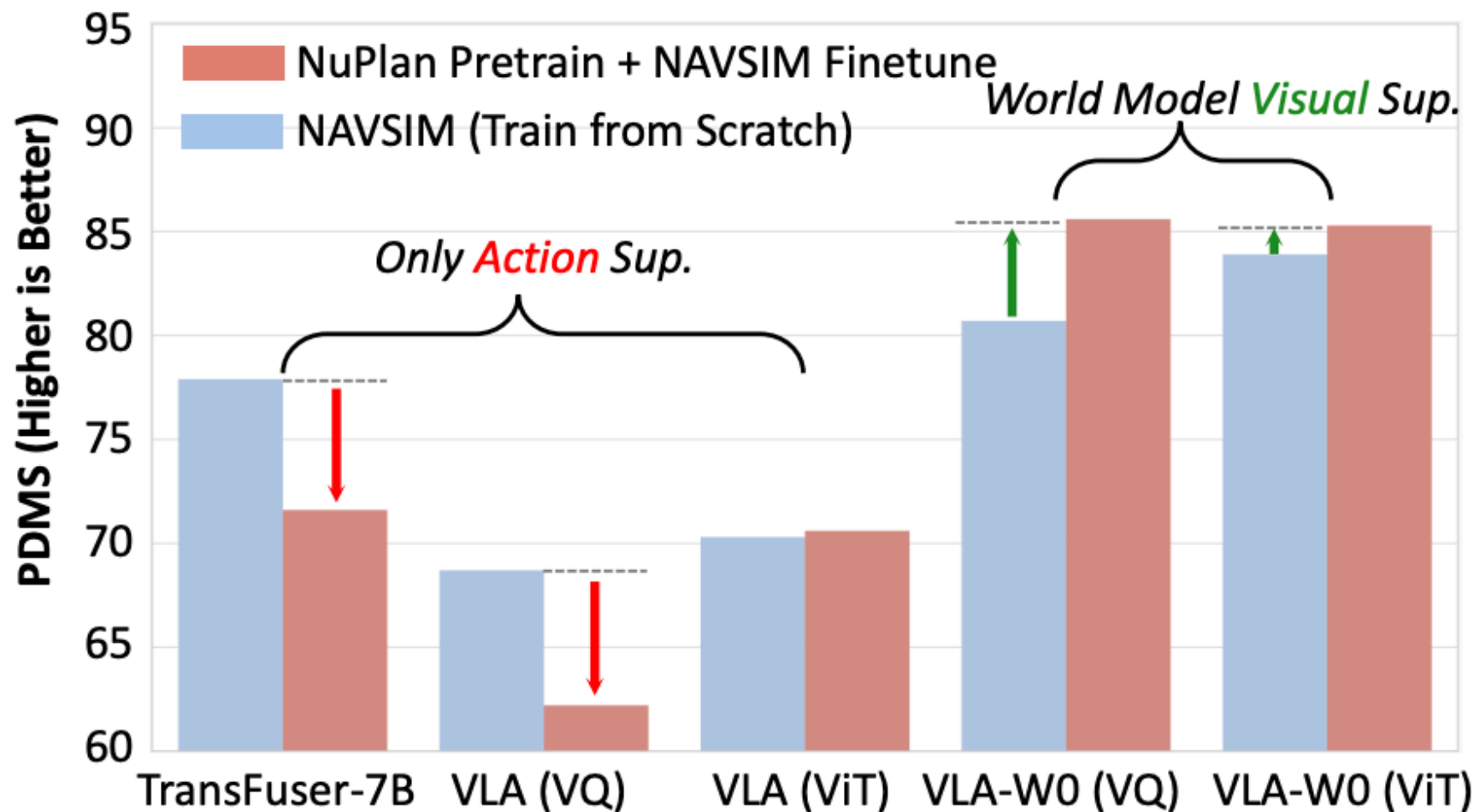


当世界模型遇到VLA：结论（2）



中国科学院自动化研究所
模式识别实验室
New Laboratory of Pattern Recognition

□ 两者结合可显著提升模型的迁移和泛化性能



当世界模型遇到VLA：结论（3）



中国科学院自动化研究所
模式识别实验室
New Laboratory of Pattern Recognition

□ 自回归（似乎）击败Diffusion

Action Expert	NAVSIM (103k Frames)						In-house Dataset (70M Frames)	
	NC ↑	DAC ↑	TTC ↑	C. ↑	EP ↑	PDMS ↑	ADE (m) ↓	Collision (%) ↓
Query-based	98.7	96.2	95.5	100.0	82.2	88.4	1.1248	0.0453
Flow Matching	98.4	95.3	95.2	100.0	80.9	87.2 ↓1.4%	1.0362 ↑7.9%	0.0398 ↑12.1%
Autoregressive	98.4	93.6	94.5	100.0	79.3	85.3 ↓3.6%	1.0069 ↑10.5%	0.0295 ↑34.9%

性能逆转

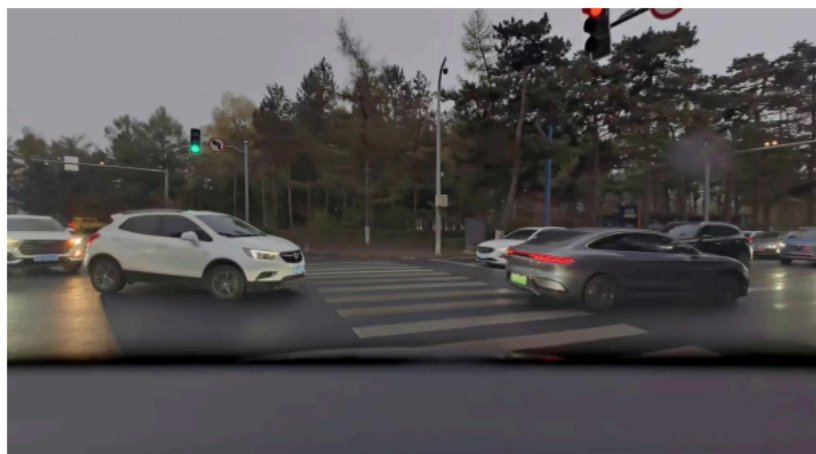
当世界模型遇到VLA：示例（1）



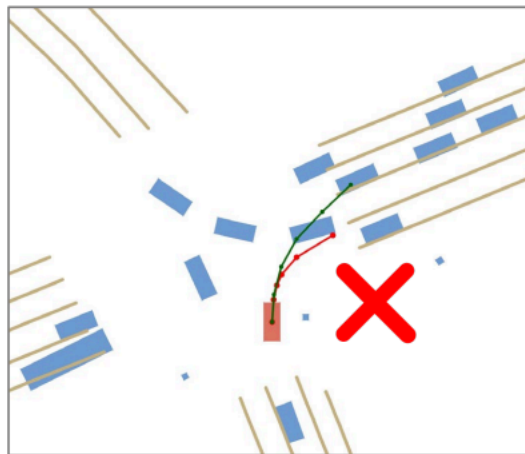
中国科学院自动化研究所
模式识别实验室
New Laboratory of Pattern Recognition

□ VLA + 世界模型在复杂场景表现更好

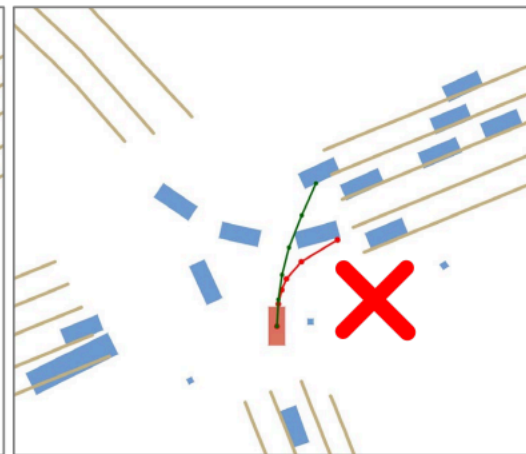
Front-view Image



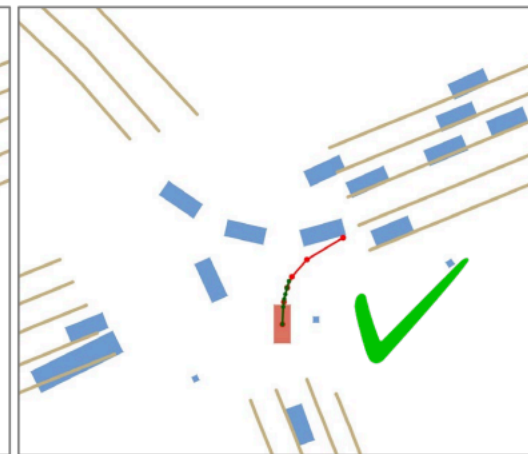
TransFuser



VLA Baseline



DriveVLA-W0



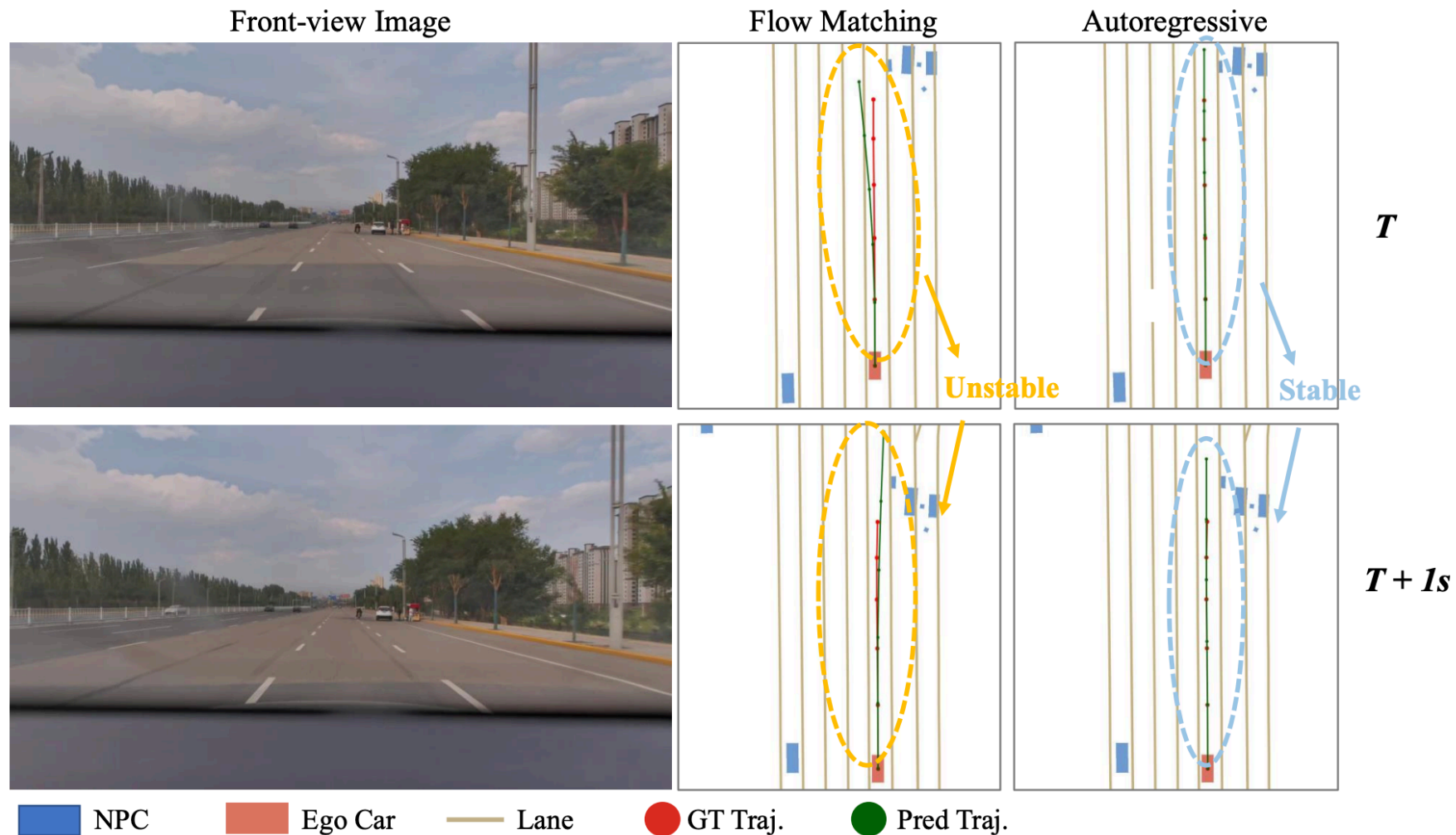
■ NPC ■ Ego Car — Lane ● GT Traj. ● Pred Traj.

当世界模型遇到VLA：示例（2）



中国科学院自动化研究所
模式识别实验室
New Laboratory of Pattern Recognition

□ 自回归相比Flow Matching更加鲁棒



当世界模型遇到VLA：示例（3）



中国科学院自动化研究所
模式识别实验室
New Laboratory of Pattern Recognition

□ 未来视觉生成

Input:

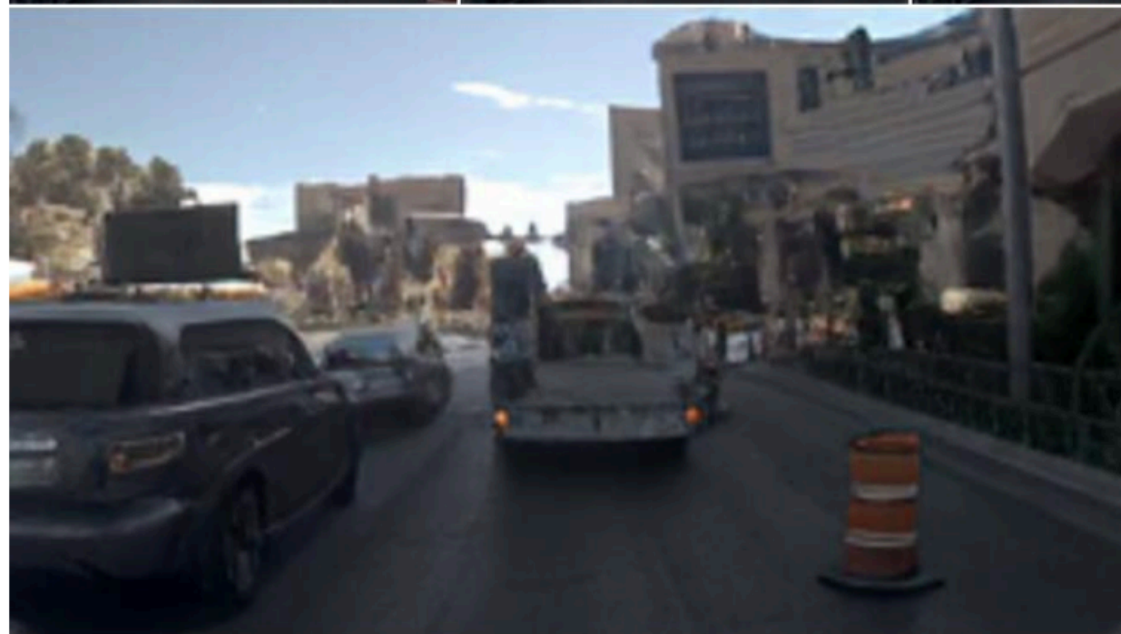
V_{t-4}

V_{t-3}

V_{t-2}

V_{t-1}

V_t



GT



Pred.

- Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving. CVPR 2024.
- FreeVS: Generative view synthesis on free driving trajectory. ICLR 2025.
- FreeSim: Toward Free-viewpoint Camera Simulation in Driving Scenes. CVPR 2025.
- Enhancing end-to-end autonomous driving with latent world model. ICLR 2025.
- DriveVLA-W0: World Models Amplify Data Scaling Law in Autonomous Driving



**谢谢，
敬请批评指正！**

T H A N K S