# 基于视频世界模型的闭环端到端自动驾驶

陈韫韬

yuntao.chen@cair-cas.org.hk

2024-5-7

# 内容提要

- 自动驾驶中信息利用的新趋势
- 闭环端到端驾驶的概念
- 端到端驾驶与生成世界模型

# 自动驾驶中信息利用的新趋势

## 2D感知时代

从图像中提取目标区域信息

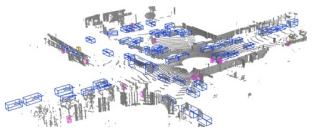
对驾驶至关重要的测距，测速，预测，规划依赖**目标区域信息**

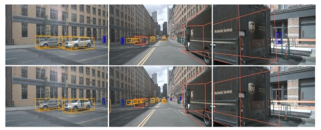
2D感知

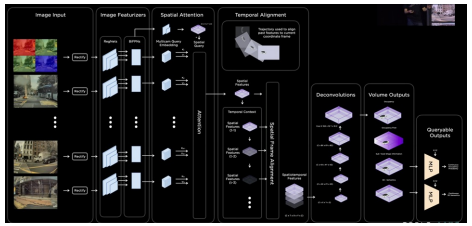

## BEV感知时代

从图像中提取封闭类别物体的位置，速度，未来轨迹等矢量化信息，规划依赖**矢量化信息**

3D感知








## 端到端时代

从图像中直接提取**决策规划信息**

大模型



端到端模型



2012 2021 2023

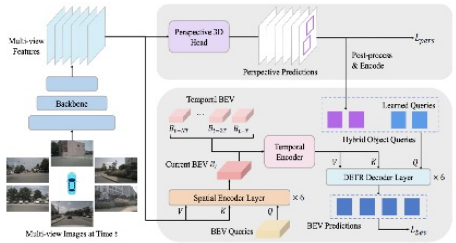
重要使能技术：**大数据+大算力+大模型**

主要趋势：从**感知**到**认知**，规划输入信息由**低维**到**高维**

# 回顾感知智能时代

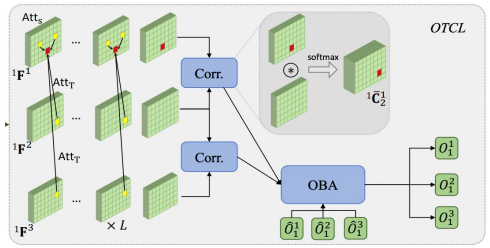
## 激光点云与视觉感知

### BEVFormer-V2 [CVPR 2023]



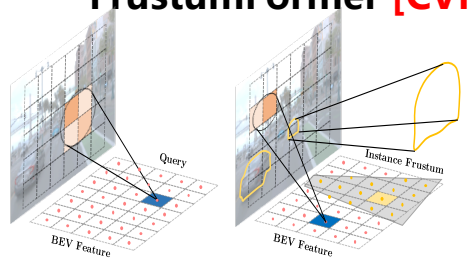

2D与3D通路整合的全视角BEV感知模型。

### BA-Det [CVPR 2023]



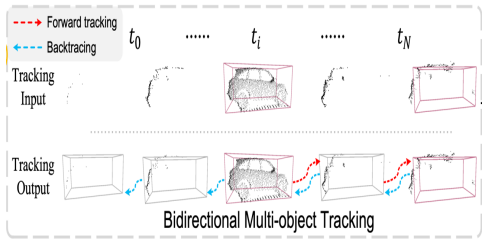

以物体重建为约束的多帧三维物体检测。

### FrustumFormer [CVPR 2023]



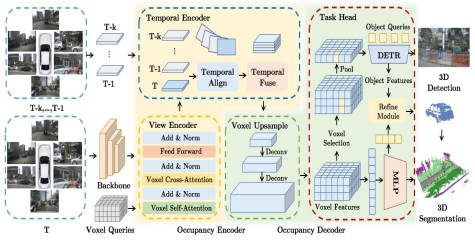

基于视锥约束关系的的2D-3D通路整合感知。

### CTRL [ICCV 2023]



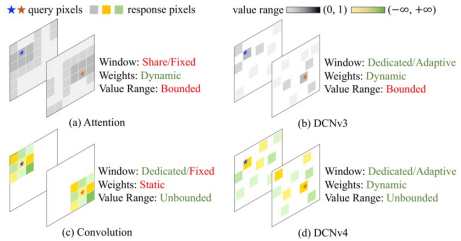

全局时空信息引导的长时目标检测跟踪。

### PanoOcc [CVPR 2024]




基于统一占用栅格表示的3D全景分割。

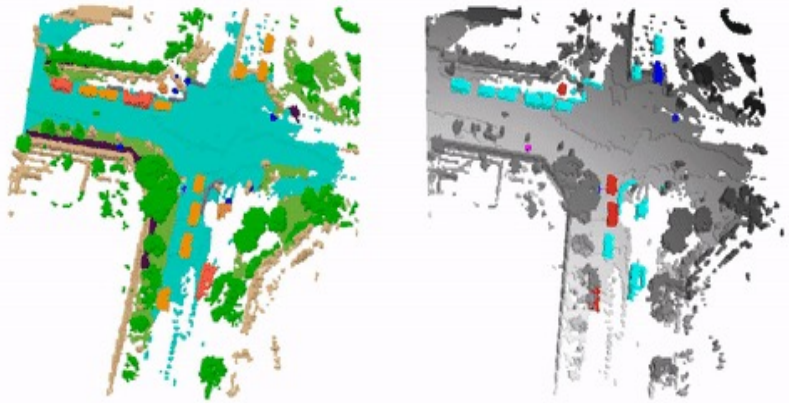
### DCNv4 [CVPR 2024]




硬件高效的可变形卷积基础网络结构实现

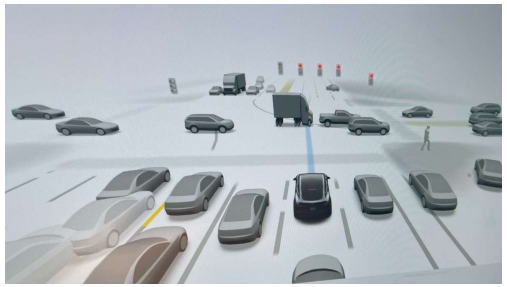
# 回顾感知智能时代

## 感知智能 = 真实世界的矢量化孪生建模



- 感知智能能力取决于标注数据规模化程度

Super Intelligent

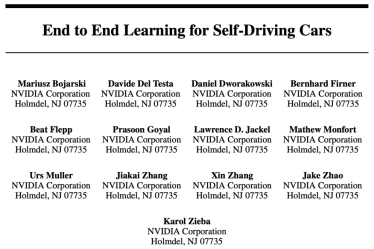


Not so Intelligent

- 矢量化孪生建模丢失原始图像信息


# 展望行为智能时代

- **端到端自动驾驶**重掀浪潮，在学界和业界都受到广泛关注


End to End Learning for Self-Driving Cars

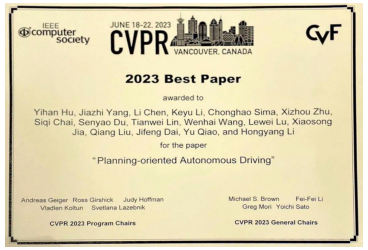
Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D. Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, Xin Zhang, Jake Zhao, Karol Zieba
NVIDIA Corporation, Holmdel, NJ 07735

DAVE-2 (2016)


End-to-end Driving via Conditional Imitation Learning

Felipe Codevilla, Matthias Müller, Antonio López, Vladlen Koltun, Alexey Dosovitskiy

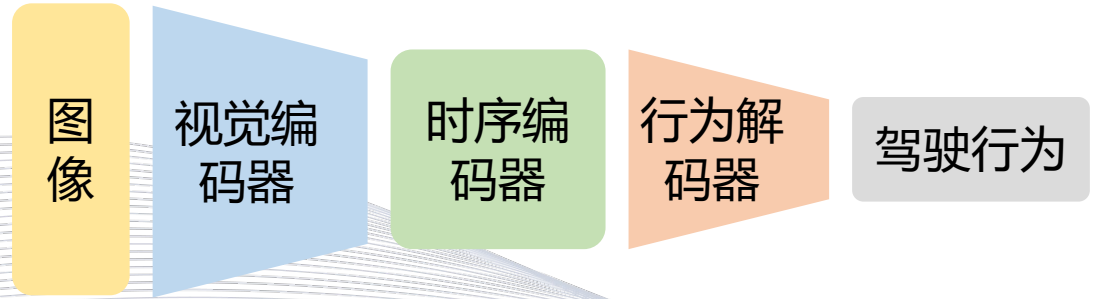
Conditional E2E(2018)




UniAD（2023）


TESLA FSD 12（2024）

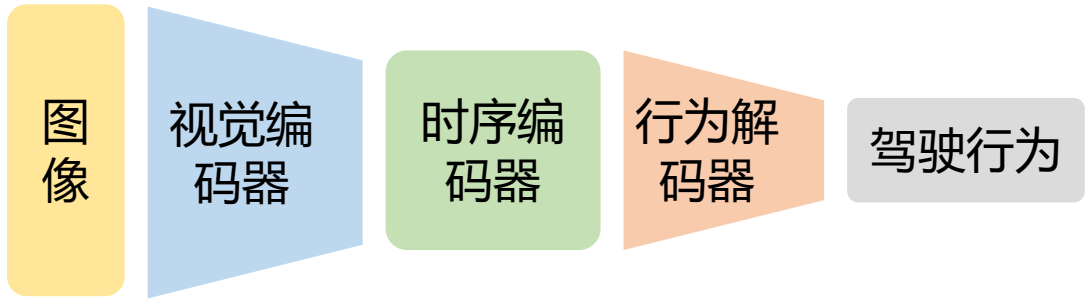
- 什么是端到端自动驾驶 – 一个极简端到端模型




输入：历史 $T_1$ 时刻的前视图像
输出：未来 $T_2$ 时刻的横纵向三维路点
标签：对齐后的（图像，行为）序列
损失：多模轨迹损失

# 展望行为智能时代

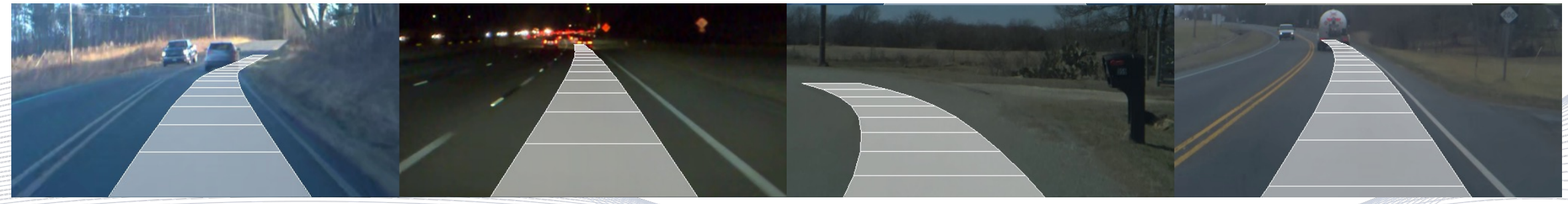
- 什么是端到端自动驾驶 – 一个极简端到端模型




$T_1 \times H \times W \times C$

$T_2 \times 6$

输入：历史 $T_1$ 时刻的前视图像
输出：未来 $T_2$ 时刻的横纵向三维路点
标签：对齐后的（图像，行为）序列
损失：多模轨迹损失



# 展望行为智能时代

- 端到端自动驾驶的**优势**

  - 通过（图像，动作）统一建模行为智能


极罕见语义

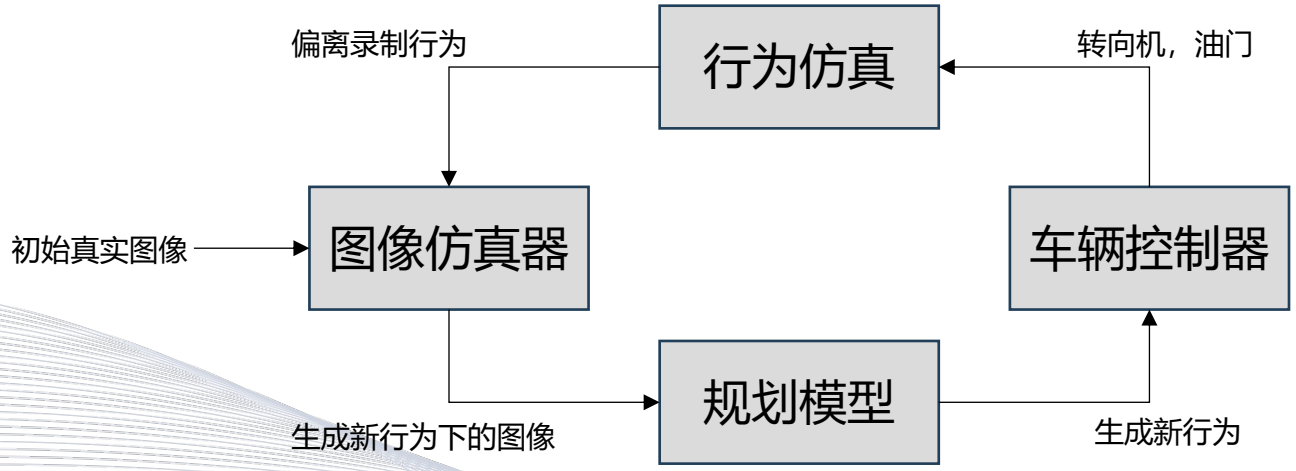
交警意图理解

细碎抛洒物

非结构道路规划

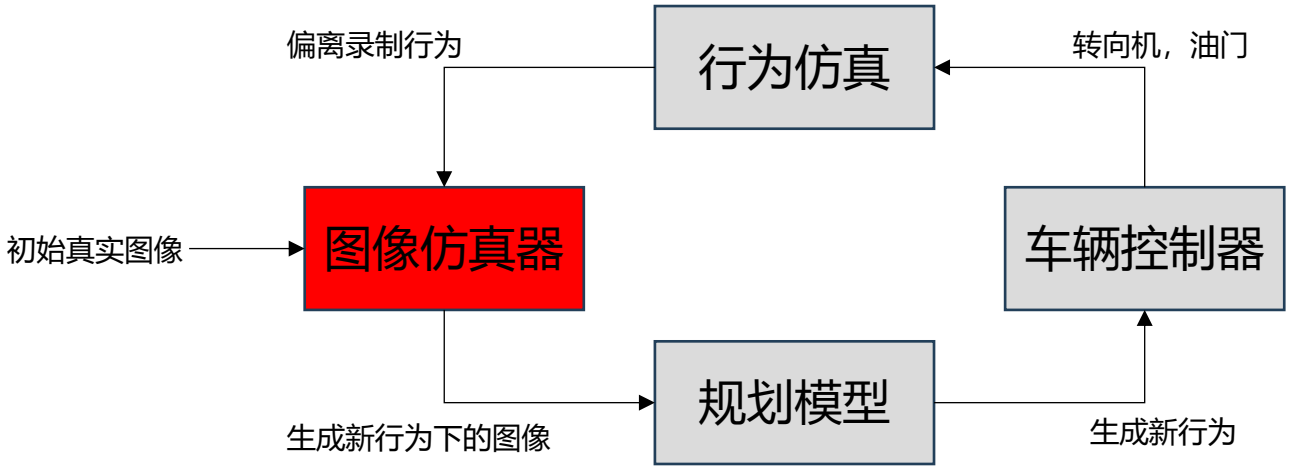
# 展望行为智能时代

- 端到端自动驾驶的**挑战**
  - 可解释性下降，需要更为系统的测试手段
  - 模型行为影响所获取的图像，无法像感知智能一般进行开环评测



# 展望行为智能时代

- 端到端自动驾驶的**挑战**
  - 端到端自动驾驶需要有力的图像仿真器




CARLA 仿真

三维重建仿真

世界模型仿真

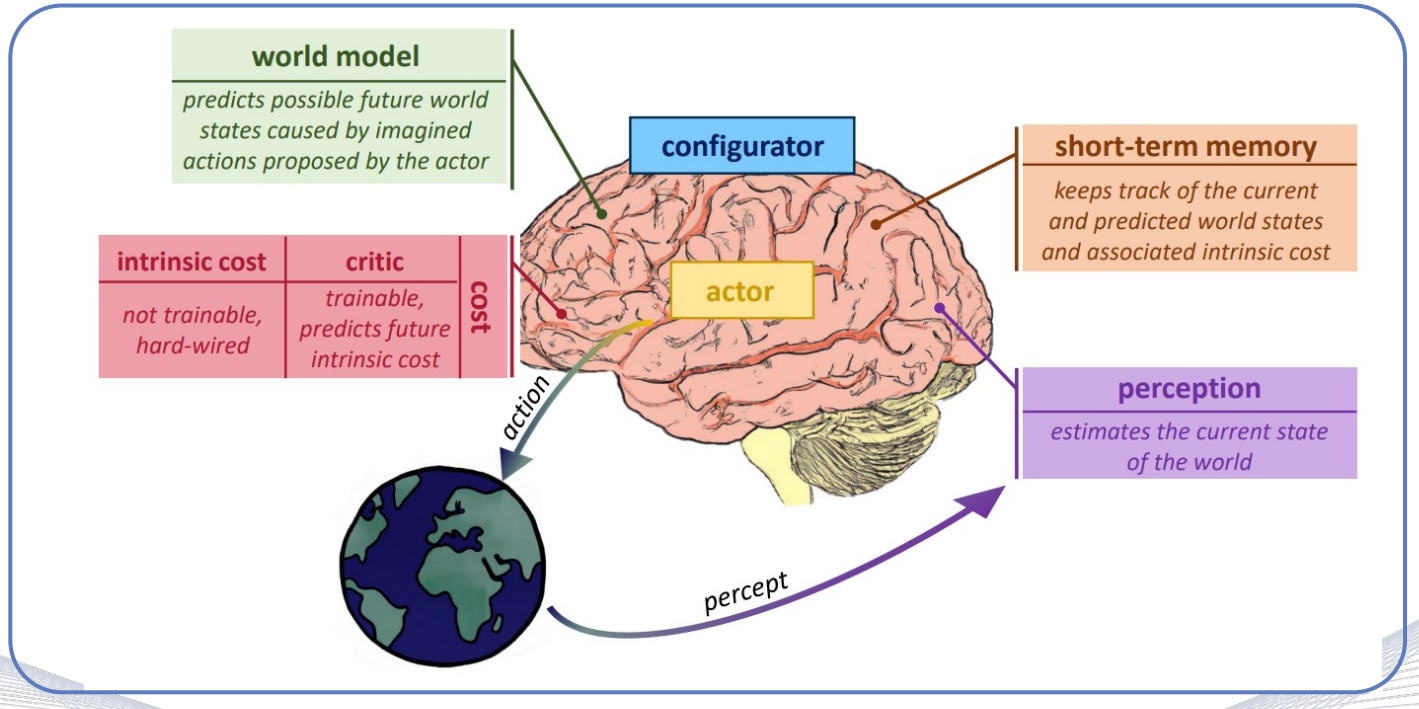
# 生成世界模型应运而生

## ❑ 构建世界模型：从数据中学习到世界的运转规律，掌握知识

*“要实现 AGI，最关键的一点是让机器了解世界是如何运转的，掌握广泛的现实知识”*

——图灵奖得主、Meta AI负责人**Yann LeCun**

**自主智能**构架包括**配置器模块**、**感知模块**、**世界模型模块**、**短时记忆模块**和**执行模块**




**LeCun**, et.al, *A Path Towards Autonomous Machine Intelligence, 2022*

# 生成世界模型的概念

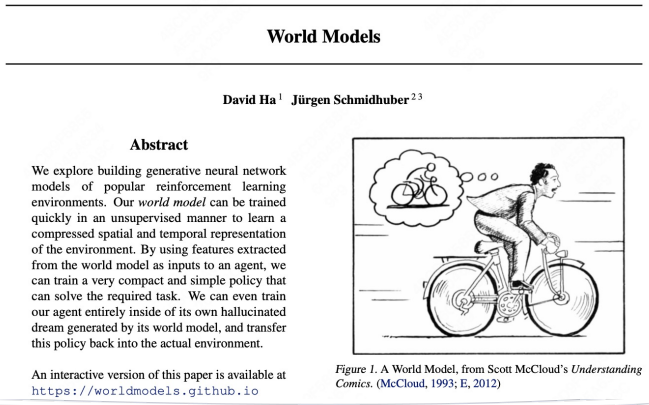
- 视频世界模型是一个**预测模型**：旨在合理地预测事物在特定**动作**下的演变

The **world model module** predicts possible **future world states** as a function of imagined **actions sequences** proposed by the actor.

$$I_{t+1} = f(S_t, I_t)$$

——Yann Lecun

- 以人为启发的**智能学习**方式：人类使用有限的感官感知世界，并基于这些感知建立起一个内部的、简化的世界模型，我们所做的决策和行动都是基于这个内部模型



World Models

David Ha [1] Jürgen Schmidhuber [2 3]

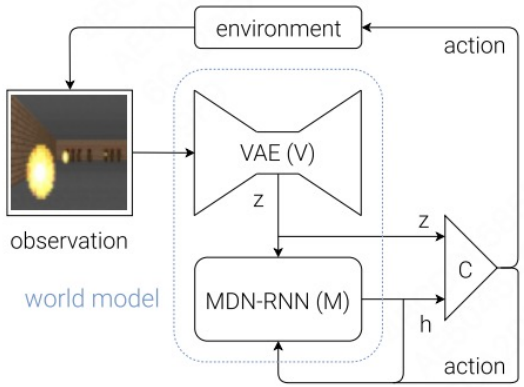
Abstract

We explore building generative neural network models of popular reinforcement learning environments. Our *world model* can be trained quickly in an unsupervised manner to learn a compressed spatial and temporal representation of the environment. By using features extracted from the world model as inputs to an agent, we can train a very compact and simple policy that can solve the required task. We can even train our agent entirely inside of its own hallucinated dream generated by its world model, and transfer this policy back into the actual environment.

An interactive version of this paper is available at https://worldmodels.github.io



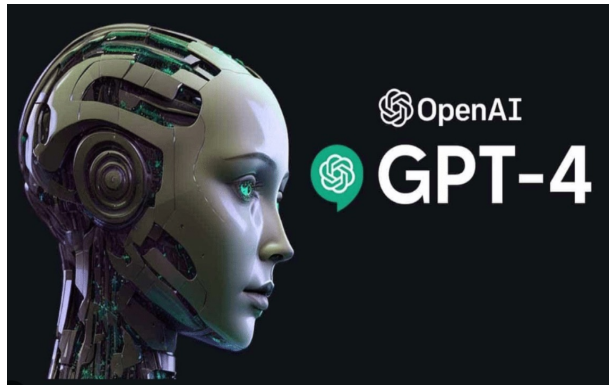
*Figure 1.* A World Model, from Scott McCloud's *Understanding Comics.* (McCloud, 1993; E, 2012)



# 生成世界模型的爆火

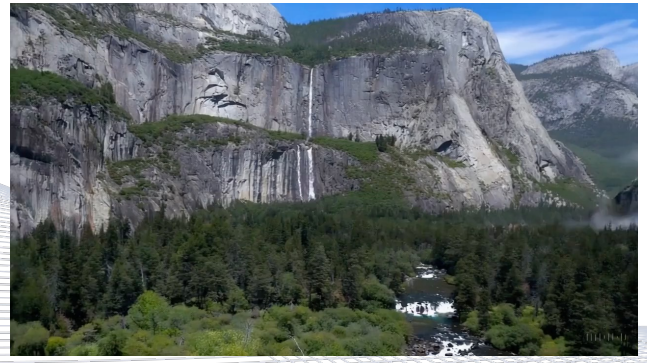
- 视频生成之于物理世界，就如同语言建模之于数字世界

**Predict next token**




**Predict next frame**

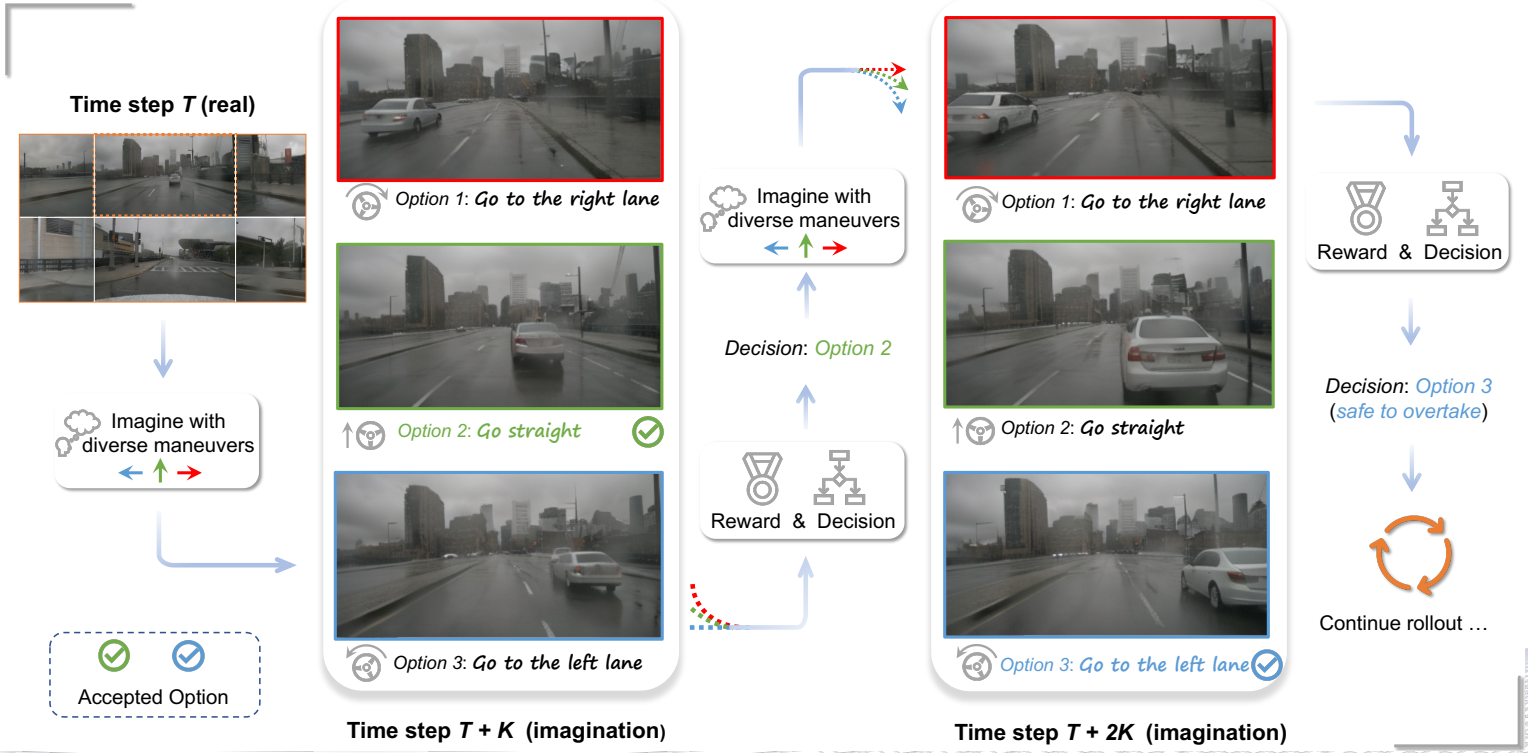

**Sora： Video generation models as world simulators**



# 世界模型与端到端自动驾驶

***Driving into the Future: Multiview Visual Forecasting and Planning with World Model for Autonomous Driving (CVPR 2024)***

- 世界模型根据端到端规划模型预测的路线来想象未来情景
- 端到端规划模型根据世界模型想象场景的视觉反馈来选择最优决策



# 世界模型与端到端自动驾驶

***Driving into the Future: Multiview Visual Forecasting and Planning with World Model for Autonomous Driving (CVPR 2024)***

## ❑ 核心挑战

### ➢ 如何实现一个多视图的生成世界模型?

### ➢ 如何用生成世界模型来训练端到端规划器?

长尾案例

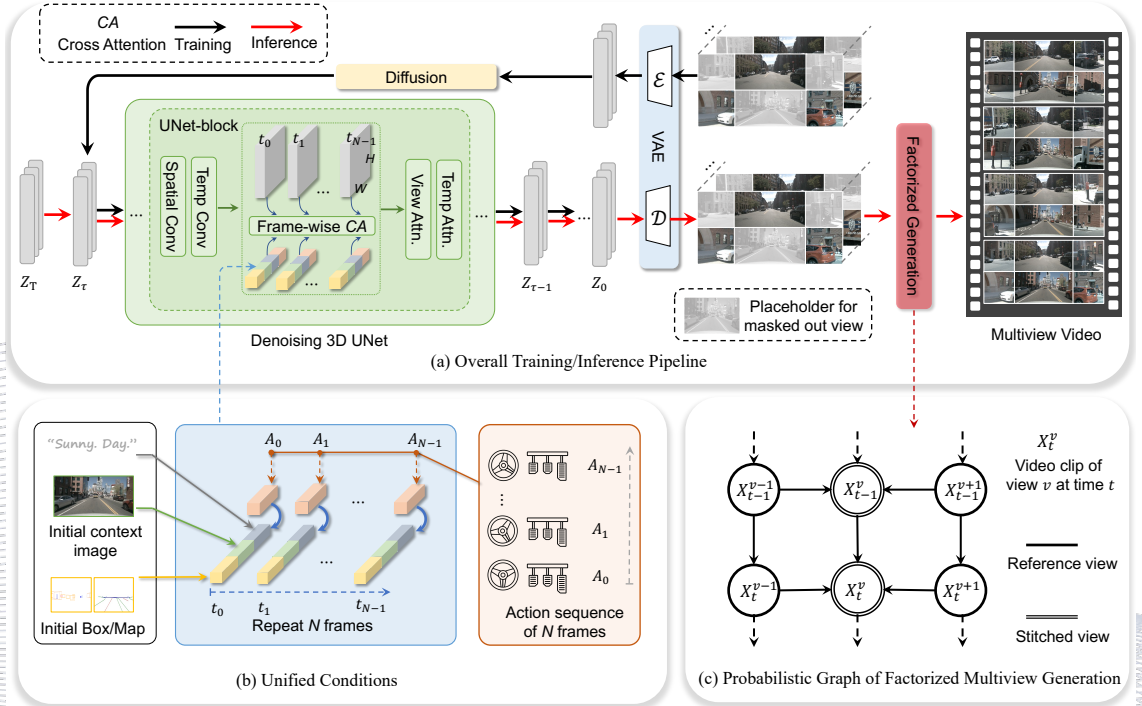
模型退化

# 世界模型与端到端自动驾驶

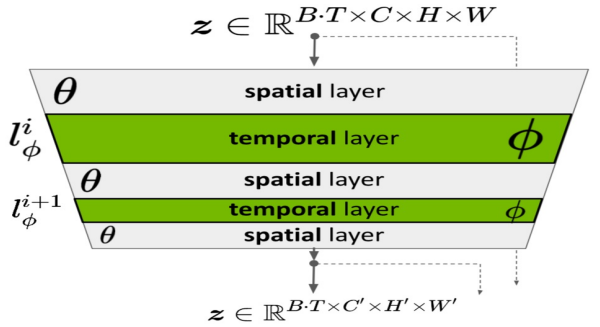
***Driving into the Future: Multiview Visual Forecasting and Planning with World Model for Autonomous Driving (CVPR 2024)***

## ❑ 挑战1: 构建多视图的视频生成模型

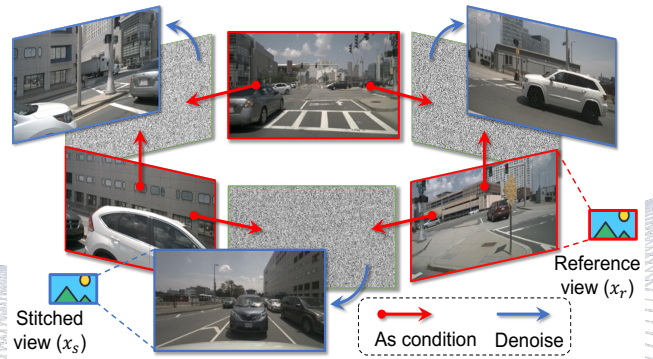
### 整体架构




### 时序层建模




### 多视图建模



# 世界模型与端到端自动驾驶

***Driving into the Future: Multiview Visual Forecasting and Planning with World Model for Autonomous Driving (CVPR 2024)***

## ❑ 多视图视频生成模型效果

| Method | Multi-view | Video | FID↓ | FVD↓ |
|---|---|---|---|---|
| BEVGen [53] | ✓ | | 25.54 | - |
| BEVControl [69] | ✓ | | 24.85 | - |
| MagicDrive [17] | ✓ | | 16.20 | - |
| Ours | ✓ | | **12.99** | - |
| DriveGAN [31] | | ✓ | 73.4 | 502.3 |
| DriveDreamer [63] | | ✓ | 52.6 | 452.0 |
| Ours | ✓ | ✓ | **15.8** | **122.7** |

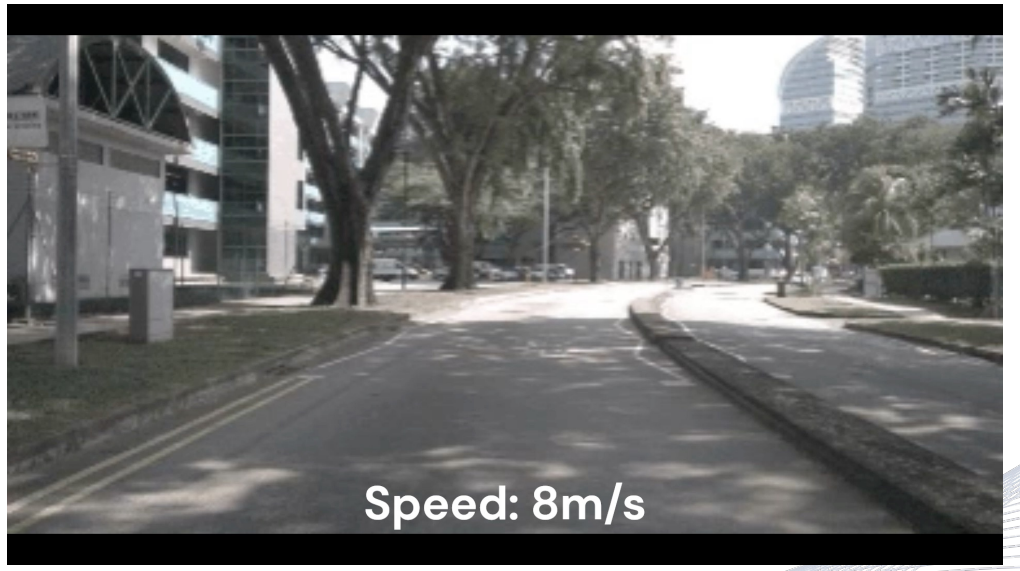
(a) **Generation quality.**

# 世界模型与端到端自动驾驶

***Driving into the Future: Multiview Visual Forecasting and Planning with World Model for Autonomous Driving (CVPR 2024)***

## ❑ 多视图视频生成模型可控性

- ➢ **自车行为的控制**

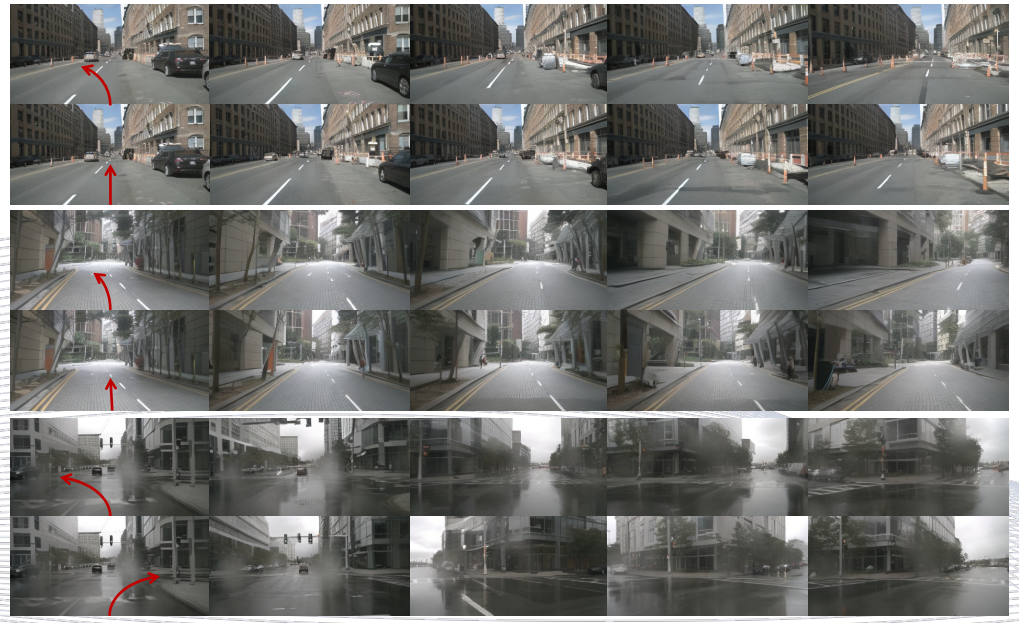
# 世界模型与端到端自动驾驶

***Driving into the Future: Multiview Visual Forecasting and Planning with World Model for Autonomous Driving (CVPR 2024)***

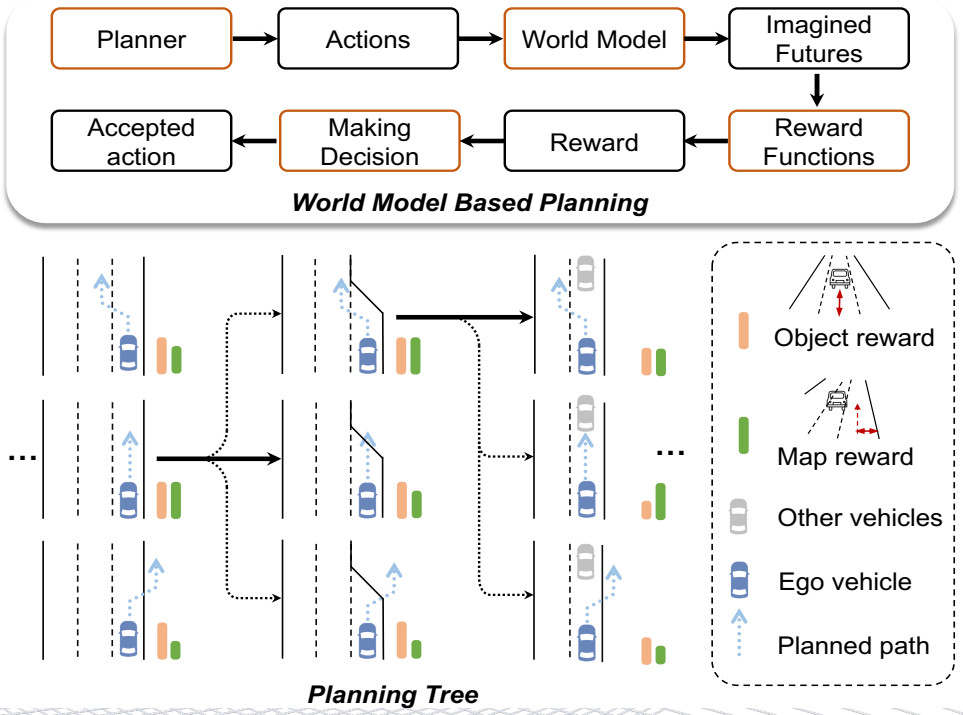
## ❑ 挑战2: 用世界模型来训练端到端规划器

### 前提

世界模型可以根据**不同的轨迹来生成未来**的情景



### 训练方案

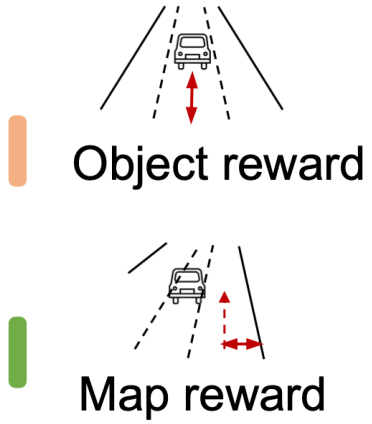


# 世界模型与端到端自动驾驶

***Driving into the Future: Multiview Visual Forecasting and Planning with World Model for Autonomous Driving (CVPR 2024)***

## ❑ 挑战2: 用世界模型来训练端到端规划器

### 视觉反馈设计




| Map Reward | Object Reward | L2 (m) ↓ 1s | 2s | 3s | Avg. | Collision (%) ↓ 1s | 2s | 3s | Avg. |
|---|---|---|---|---|---|---|---|---|---|
| | | 0.51 | 0.97 | 1.57 | 1.02 | 0.34 | 0.74 | 1.72 | 0.93 |
| ✓ | | 0.45 | 0.82 | 1.29 | 0.85 | 0.12 | 0.33 | 0.72 | 0.39 |
| | ✓ | 0.43 | 0.77 | 1.20 | 0.80 | 0.12 | 0.21 | 0.48 | 0.27 |
| ✓ | ✓ | **0.43** | **0.77** | **1.20** | **0.80** | **0.10** | **0.21** | **0.48** | **0.26** |

### 结合GPT4V（将来）

Describe the road surface condition first, please provide potential driving behaviors and the corresponding rewards.

The image depicts an overcast day with wet road conditions. There are **areas of standing water on the road**.
- **Avoiding Puddles**: Steering around large puddles. **Reward: Positive.**
- **Driving Through Puddles**: If unavoidable, **slowing down** before entering the puddle to reduce the risk. **Reward: Positive**; However, if done at **high speed**, it could lead to loss of control. **Reward: Negative**.

# 世界模型与端到端自动驾驶

***Driving into the Future: Multiview Visual Forecasting and Planning with World Model for Autonomous Driving (CVPR 2024)***

## ❑ 世界模型对于端到端驾驶的助力

➢ **1. 显著提升驾驶的安全性**： 引入未来场景评估对于端到端规划的提升作用

| Method | L2 (m) ↓ | | | | Collision (%) ↓ | | | |
|---|---|---|---|---|---|---|---|---|
| | 1s | 2s | 3s | Avg. | 1s | 2s | 3s | Avg. |
| VAD (GT cmd) | 0.41 | 0.70 | 1.05 | 0.72 | 0.07 | 0.17 | 0.41 | 0.22 |
| VAD (random cmd) | 0.51 | 0.97 | 1.57 | 1.02 | 0.34 | 0.74 | 1.72 | 0.93 |
| VAD (sampled cmd) | 0.46 | 0.83 | 1.31 | 0.87 | 0.23 | 0.42 | 0.76 | 0.47 |
| Ours | **0.43** | **0.77** | **1.20** | **0.80** | **0.10** | **0.21** | **0.48** | **0.26** |

L2: 评估预测轨迹与真实轨迹的差异 Collision: 评估轨迹潜在的碰撞风险

# 世界模型与端到端自动驾驶

***Driving into the Future: Multiview Visual Forecasting and Planning with World Model for Autonomous Driving (CVPR 2024)***

## ❑ 世界模型对于端到端驾驶的助力

➢ **2. 提升端到端规划器的鲁棒性**： 使用生成的OOD场景来训练端到端规划器，以解决现有端到端规划器面对OOD场景时的性能退化问题


现有端到端规划器面对OOD场景时的性能退化

| OOD | World Model f.t. | L2 (m) ↓ | | | | Collision (%) ↓ | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 1s | 2s | 3s | Avg. | 1s | 2s | 3s | Avg. |
| | | 0.41 | 0.70 | 1.05 | 0.72 | 0.07 | 0.17 | 0.41 | 0.22 |
| ✓ | | 0.73 | 0.99 | 1.33 | 1.02 | 1.25 | 1.62 | 1.91 | 1.59 |
| ✓ | ✓ | 0.50 | 0.79 | 1.17 | 0.82 | 0.72 | 0.84 | 1.16 | 0.91 |

利用世界模型的生成能力来提升鲁棒性

# 展望未来

- 更大规模、更丰富的数据来提升模型的生成能力

- 模拟更真实的物理交互

# 敬请批评指正！