KE:SAI

**Open Science** Autonomy Lab

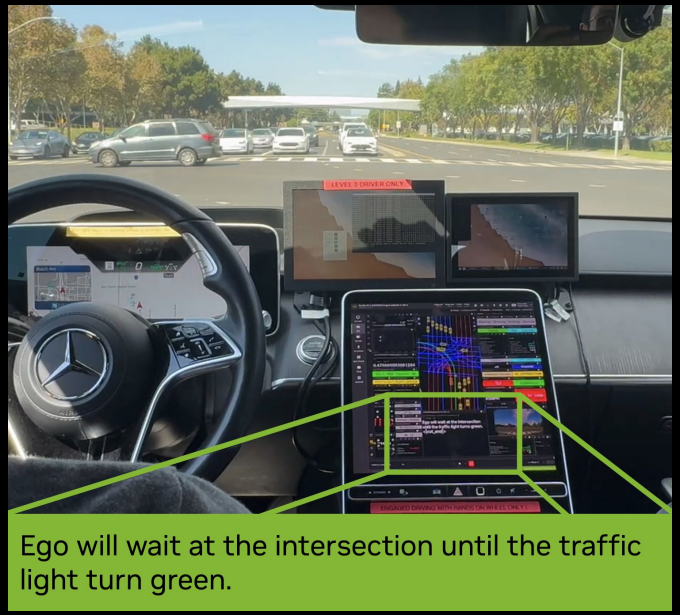
# Democratizing Autonomous Driving with **Latent World Models**

kashyap7x.github.io

Our mission is to advance the efficiency frontier
of robust and safe physical AI
through fully open and reproducible research.

# Current frontier: Vision Language Action models

Predicting an **expert action** conditioned on a **visual world state** and a **reasoning trace** in **natural language**




- **Reasoning, rare case handling, arbitration**
- Semi-automatic labeling
- **Misalignment** of reasoning to actions
- Hard to incorporate **full sensor suite**
- Many parameters dedicated to **knowledge memorization**
- …

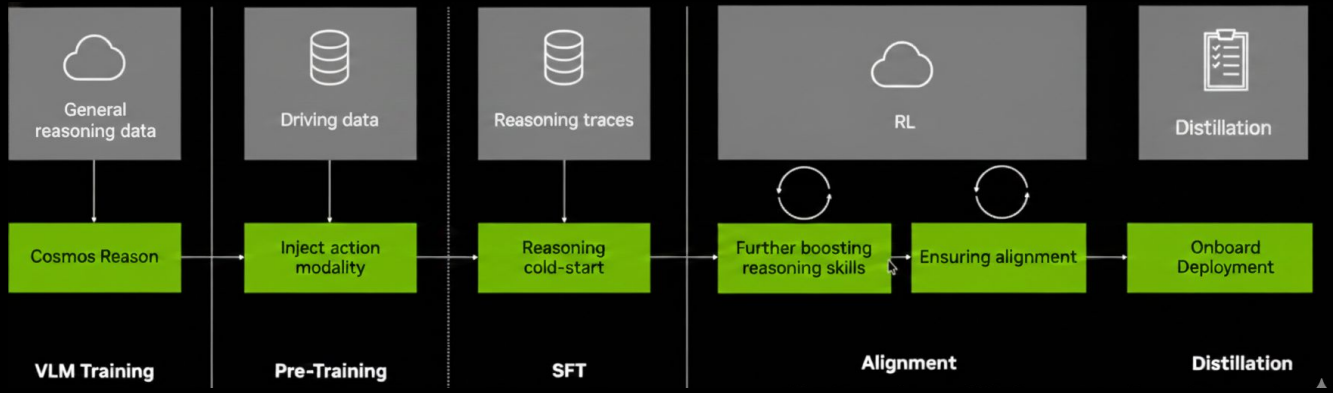


Wang et al., Alpamayo-R1: Bridging Reasoning and Action Prediction for Generalizable Autonomous Driving in the Long Tail. ArXiv 2025

# The current frontier is inefficient

End-to-end VLAs are the dominant paradigm since 2025, but...

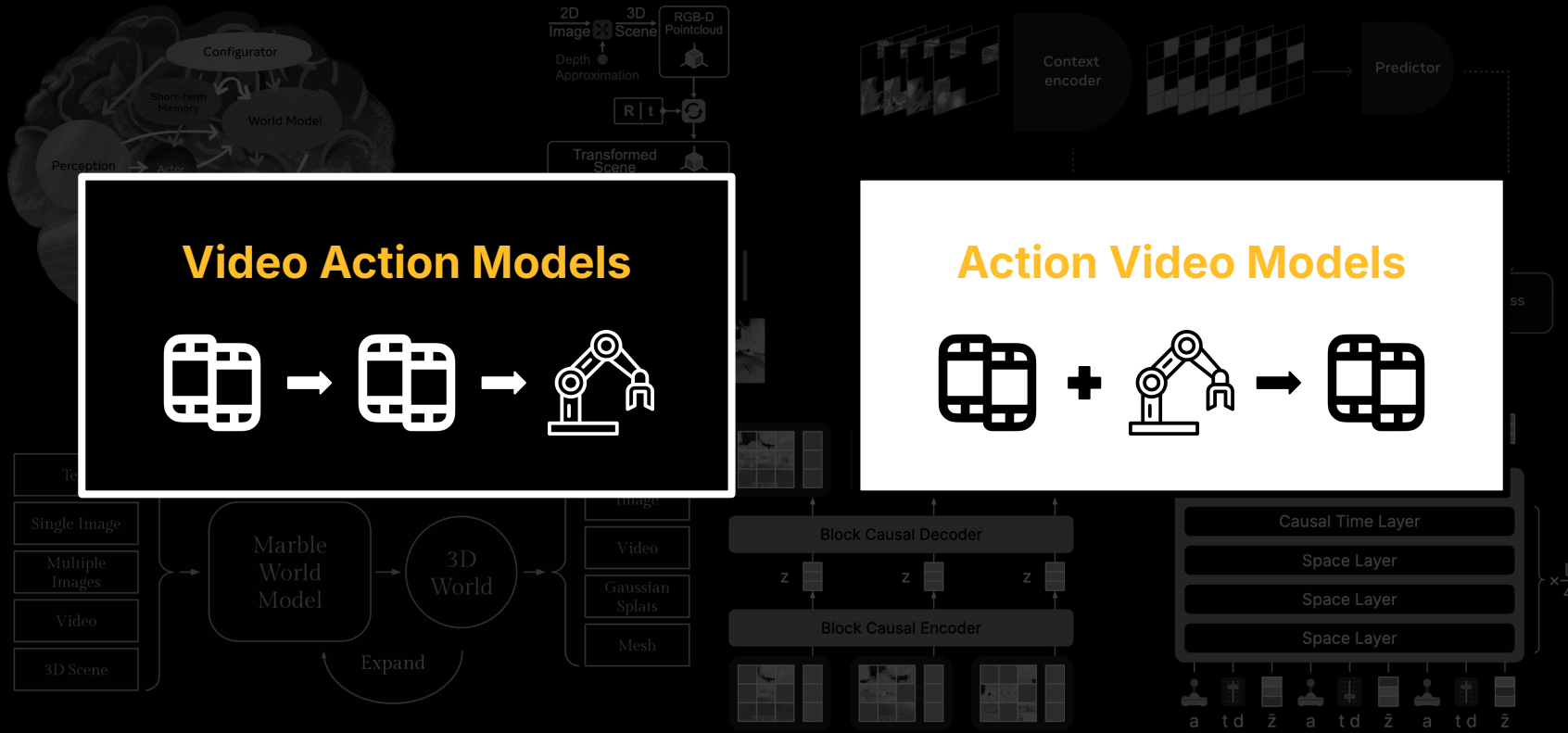
- Data inefficient (80k hours of curated expert data)
- Compute inefficient (e.g., 60k GPU hours for in-domain “Pre-Training”)
- Inference inefficient (100ms on server GPU after quantization and distillation)






Wang et al., Alpamayo-R1: Bridging Reasoning and Action Prediction for Generalizable Autonomous Driving in the Long Tail. ArXiv 2025

# World models?



# Video Action Models (VAMs)

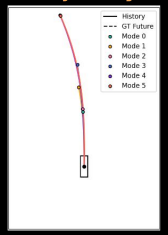
Predicting an **expert action** conditioned on an **initial world state** and a **reasoning trace** comprising the **next plausible world state(s)**

Input Observation

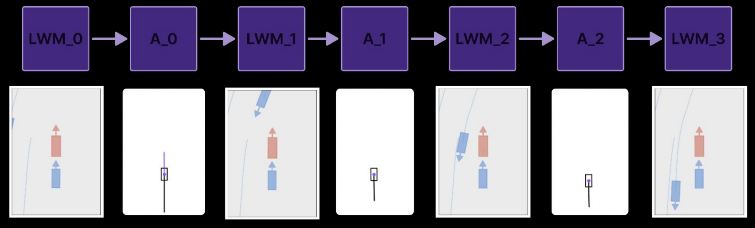
**Text CoT Reasoning (90 tokens)**

Ego is waiting at a red light, observing the traffic signal. **The light turns green, and the ego vehicle begins to accelerate forward**, indicating it is proceeding straight through the intersection. **There are no other vehicles directly in the ego's lane, so no yielding or nudging is necessary**.

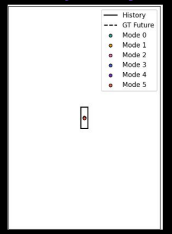
Decoded Trajectory



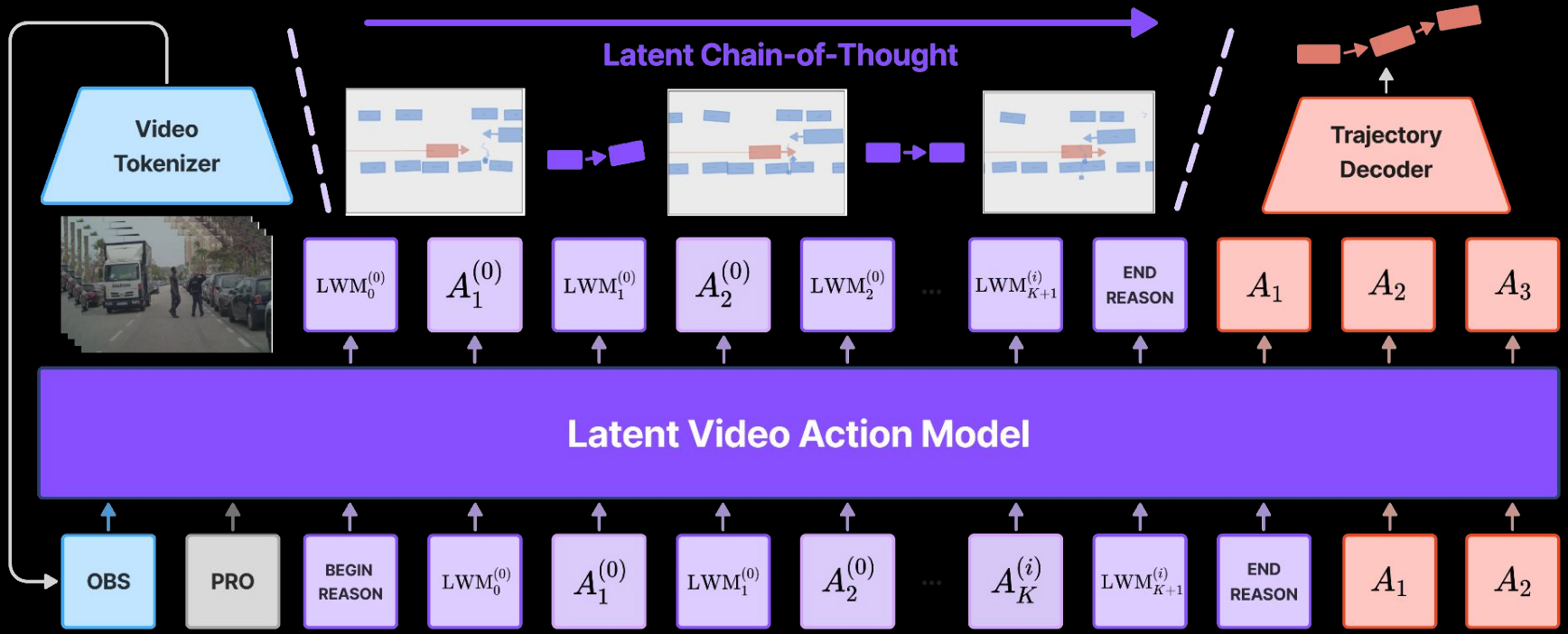

**Latent CoT Reasoning (38 tokens)**




Decoded Trajectory



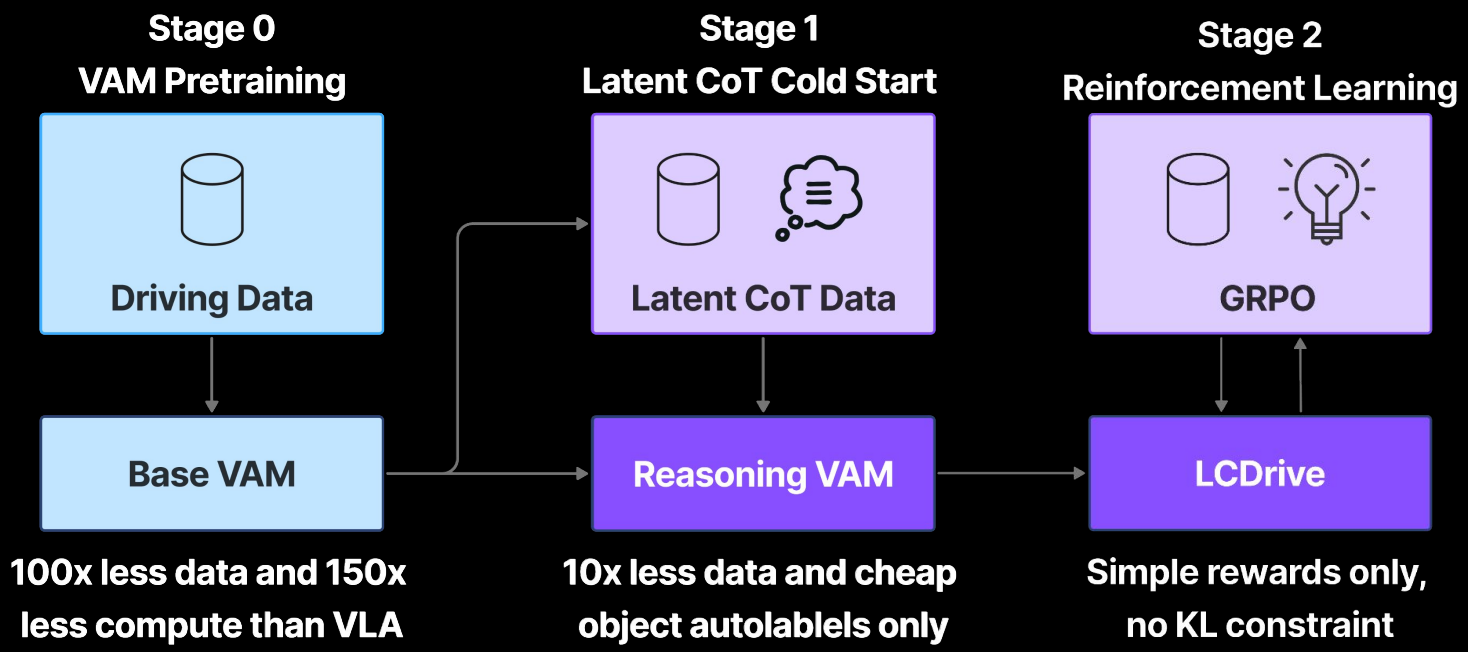
# From textual to spatio-temporal reasoning




Tan, Chitta et al., Latent Chain-of-Thought World Modeling for End-to-End Driving. CVPR 2026

# Surprising data and compute efficiency!

In Total,
**8 A100 GPUs**
**x 6 days**




Tan, Chitta et al., Latent Chain-of-Thought World Modeling for End-to-End Driving. CVPR 2026

# Was language relevant for VLA performance?

- Our results (50 hours of curated test data) indicate: **not significantly**

| Method | Collision @ 5s (%) | Off Road @ 2.5s (%) |
|---|---|---|
| Stage 0 | 2.207 | 1.753 |
| Stage 1 | 1.591 | 1.268 |
| Stage 2 | **0.836** | **1.219** |
| VLA | 0.905 | 1.391 |

# Was language relevant for VLA performance?

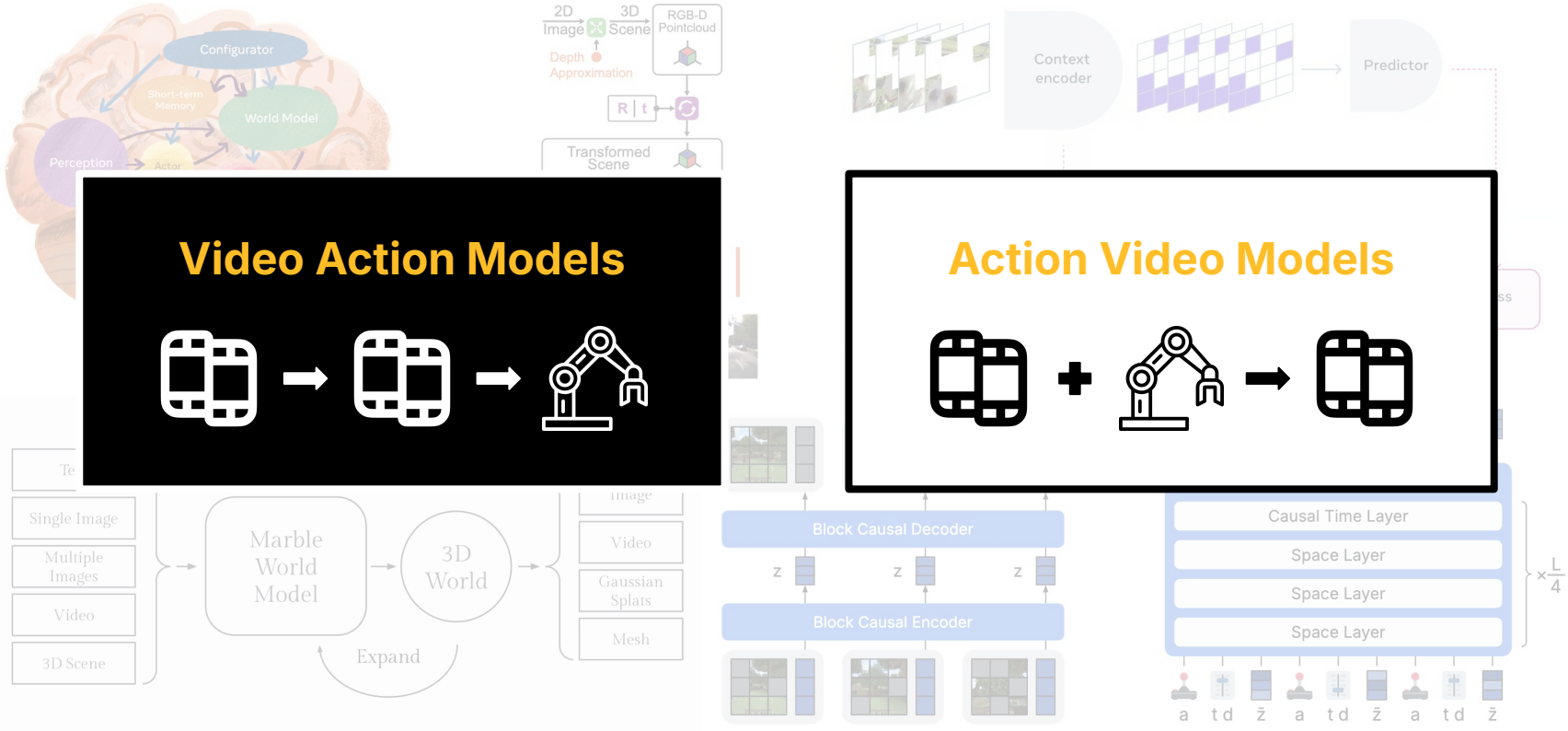
- Independent results (500 closed-loop scenarios): similar conclusion (?)

| Method | Collision (%) | Off Road (%) |
|---|---|---|
| **Alpamayo 1.5** (80k hours, 10B params) | 7.5 | **1.6** |
| **OmniDreams** (20k hours, 2B params) | **2.6** | 2.1 |



Basant et al., NVIDIA OmniDreams: Real-Time Generative World Model for Closed-Loop Autonomous Vehicle Simulation. ArXiv 2026

# World models?



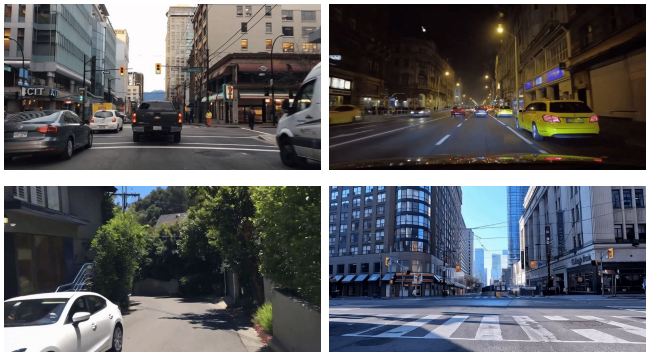
# Action Video Models (AVMs)

Predicting the **next plausible world state** (or a longer duration of states) conditioned on an **action** (or a sequence of actions)

What happens if I **go straight**?

**Encoder + Predictor**

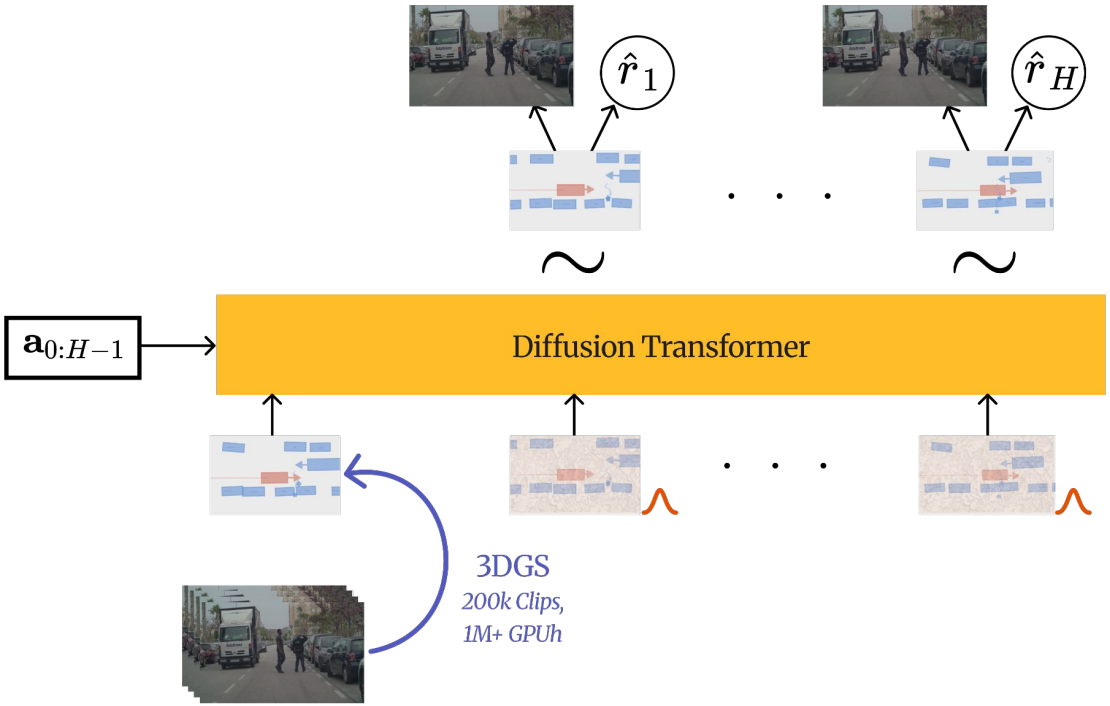
Imagined future + 🏆 Reward





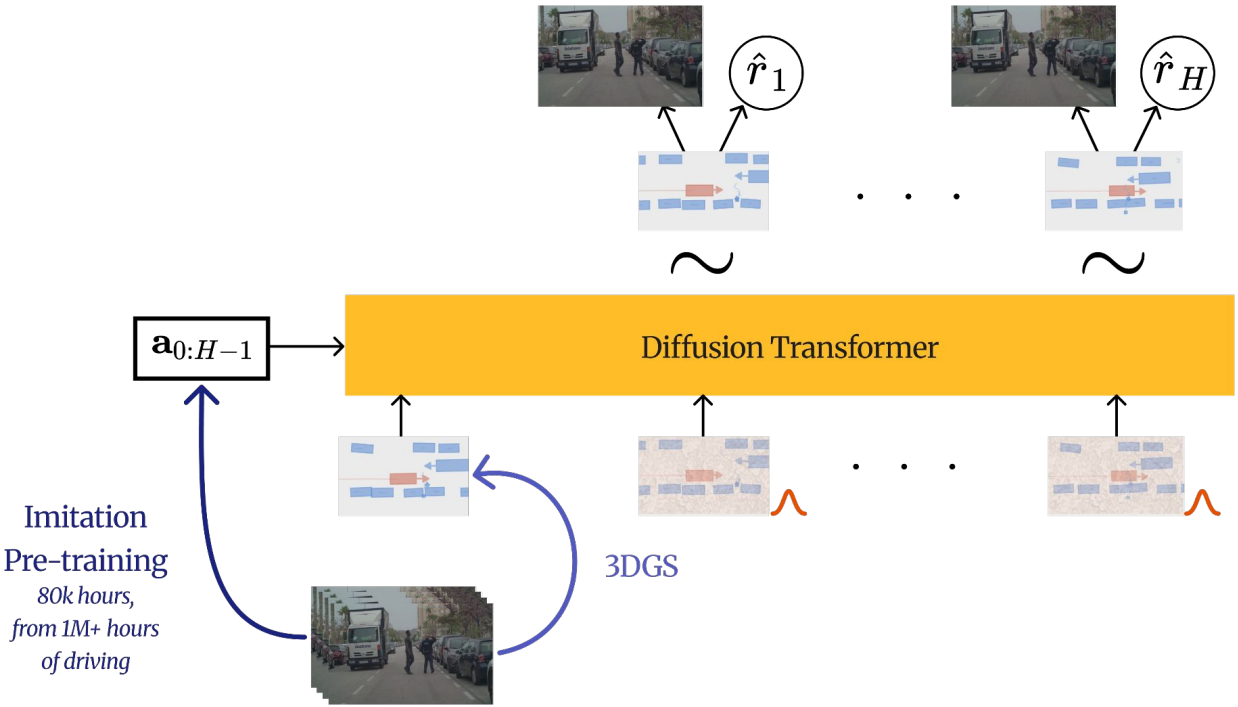
Gao et al., Vista: A Generalizable Driving World Model with High Fidelity and Versatile Controllability. NeurIPS 2024

# Combining 3DGS with traffic diffusion




Li et al., World Engine: Towards the Era of Post-Training for Physical AI. ArXiv 2026

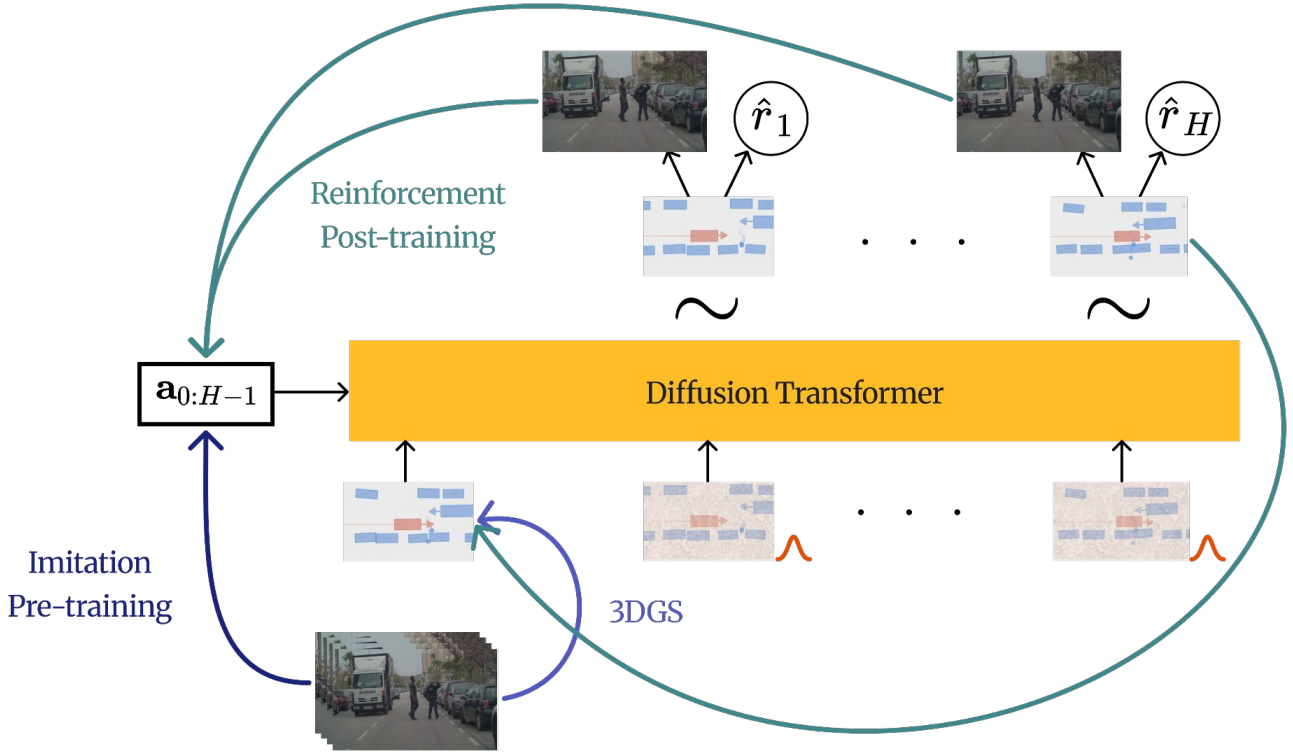
# Combining 3DGS with traffic diffusion






Li et al., World Engine: Towards the Era of Post-Training for Physical AI. ArXiv 2026

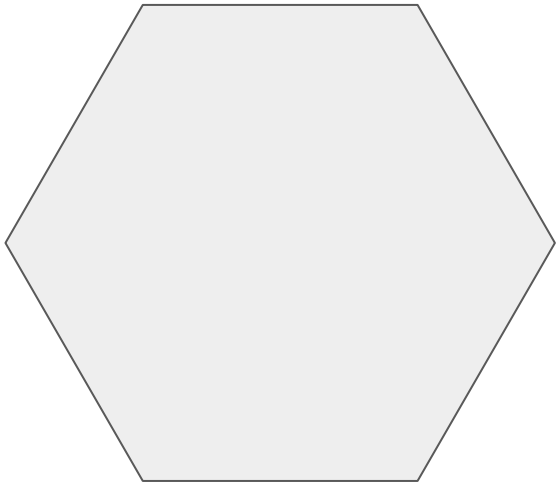
# Combining 3DGS with traffic diffusion

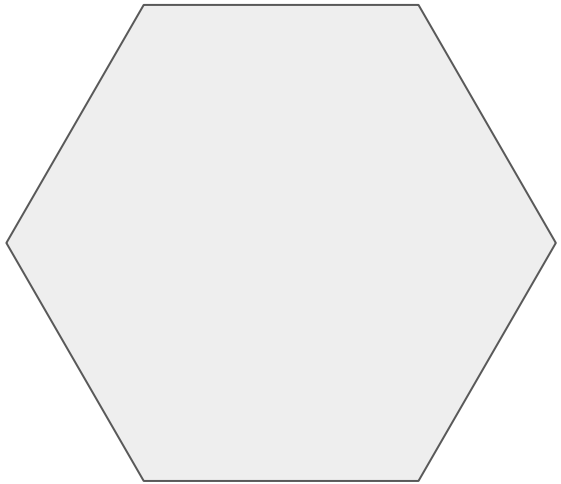




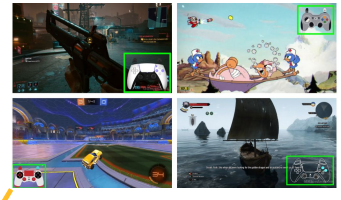

Li et al., World Engine: Towards the Era of Post-Training for Physical AI. ArXiv 2026

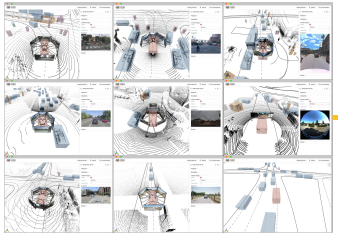
# Fully Open?

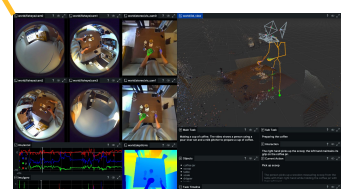





t = 0.0s








Physical AI AV Dataset


123D
Getting Started
Github
Documentation
123D Viser Server
Playback
Modalities
Camera Image
Visible
Camera
PCAM_F0
Image Scale
Render


World State
OmniDreams
Planned trajectory
NVIDIA NuRec

# Integrating our full toolkit latent world models

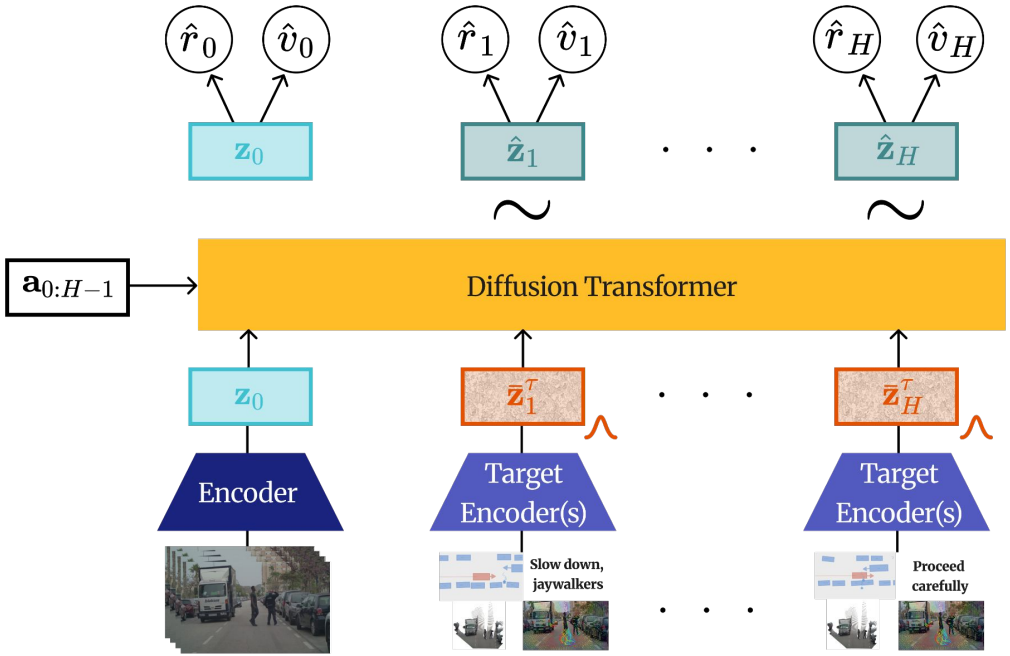
Real Data

Offline Synthetic Data

World Engine

NuRec Simulation

OmniDreams



# Takeaways

We're rapidly pushing the efficiency frontier with **latent world models**

- Video action models
  - Representation learning (making models flexible)
- Action video models
  - Simulation (making data interactive)
  - Enabled 200 kilometers without infraction in complex traffic
- Towards joint representation learning and simulation
  - Combine these and bring in auto-labeling from foundation models

KE:SAI

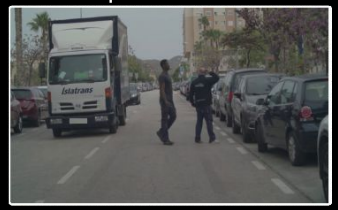
Open Science Autonomy Lab

# Thank You!

kesai.eu/join

# Extra Slides

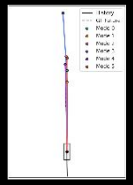
Input Observation



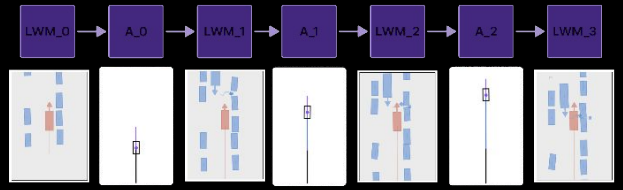
**Text CoT Reasoning**

Ego is keeping lane and maintaining speed, as **there are no immediate obstacles or agents requiring a change in behavior**. The lane lines are clear, and the surrounding vehicles are stationary. No specific agents are currently influencing the ego's behavior.

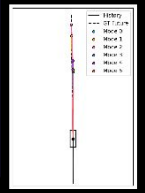
Decoded Trajectory



**Latent CoT Reasoning**




Decoded Trajectory

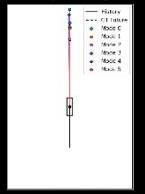


Input Observation

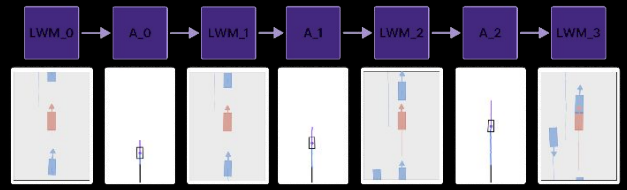
**Text CoT Reasoning**

Ego is keeping its lane and maintaining a steady speed, as there are no immediate obstacles or cut-ins. **The white van in front is moving steadily**, and the lane markings indicate a straight path. **No significant changes in behavior are expected**

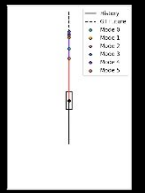
Decoded Trajectory



**Latent CoT Reasoning**




Decoded Trajectory

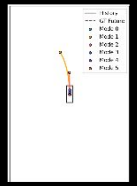


Input Observation

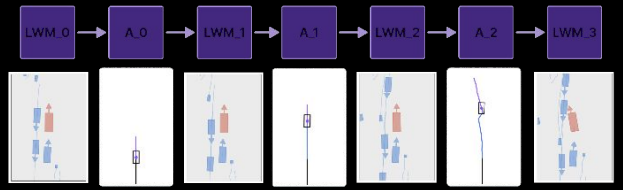
**Text CoT Reasoning**

Ego is waiting at an **intersection due to a red light**. The bus in front is stationary, and the ego vehicle is maintaining a safe distance. After the light turns green, the ego vehicle accelerates and follows the bus while keeping the lane. The ego vehicle is **likely to continue straight**.

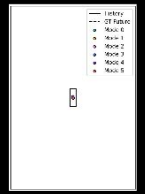
Decoded Trajectory



**Latent CoT Reasoning**




Decoded Trajectory

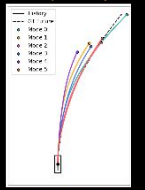


Input Observation

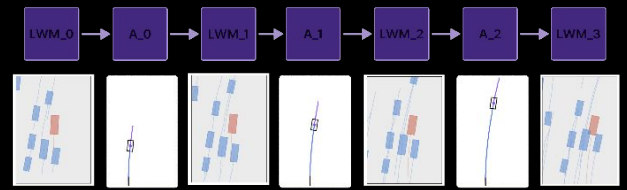
**Text CoT Reasoning**

Ego is accelerating and **changing lanes to the left, likely to overtake the white car in front**. The lane markings indicate a **left turn** lane, suggesting the **ego is preparing for a turn**. The traffic light is green, allowing the ego to proceed.

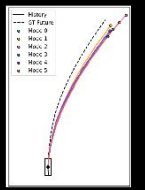
Decoded Trajectory



**Latent CoT Reasoning**




Decoded Trajectory



# Early signs of life for “direct video action” models

- Side experiment, completely infeasible for deployment

| Method | LiDAR | Annotation | PDM Score |
|---|---|---|---|
| “IDM” Network | × | × | 78.4 |
| TransFuser | ✓ | ✓ | 84.0 |
| ReSim + IDM | × | × | **86.6** |
| GT + IDM | × | × | 90.8 |

Yang, Chitta et al., ReSim: Reliable World Simulation for Autonomous Driving. NeurIPS 2025

# Biggest Recent Announcement of Top Driving Orgs

## OmniDreams

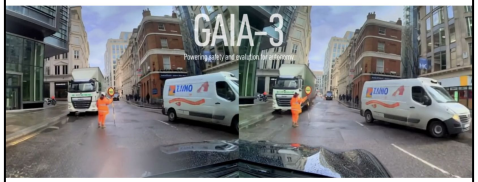
Real-Time Generative Closed-Loop Autonomous Vehicle Simulation Built on NVIDIA Cosmos

Right: Reasoning VLA driving policy, the Alpamayo 1 model, driving closed-loop inside OmniDreams – the Cosmos-generated world. This photorealistic scene is entirely synthesized by NVIDIA Cosmos in real time. Left: Human driver steering in OmniDreams.

## GAIA-3: Scaling World Models to Power Safety and Evaluation

Transforming world modeling from a tool for visual synthesis into a foundation for autonomy evaluation.

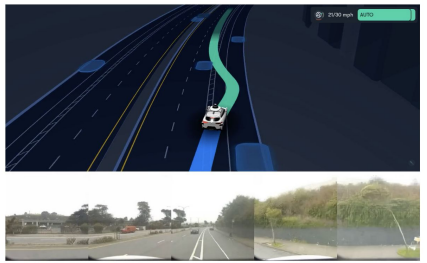
14 minute read




Evaluating autonomous-driving systems at scale remains one of the defining challenges in advancing real-world autonomy. Real-world testing is essential for proving safety, but it is costly, logistically constrained, and increasingly data-inefficient. As driving models improve and make fewer observable mistakes, the number of miles needed for statistically meaningful conclusions rises sharply. Most of those miles are uneventful, offering little signal about rare, safety-critical behavior.

## The Waymo World Model: A New Frontier For Autonomous Driving Simulation

The Waymo Driver has traveled nearly 200 million fully autonomous miles, becoming a vital part of the urban fabric in major U.S. cities and improving road safety. What riders and local communities don't see is our Driver navigating billions of miles in virtual worlds, mastering complex scenarios long before it encounters them on public roads. Today, we are excited to introduce the Waymo World Model, a frontier generative model that sets a new bar for large-scale, hyper-realistic autonomous driving simulation.



*Simulation of the Waymo Driver evading a vehicle going in the wrong direction. The simulation initially follows a real event, and seamlessly transitions to using camera and lidar images automatically generated by an efficient real-time Waymo World Model.*

# What we know about these

**OmniDreams**

- Cosmos-based
- 4 cams
- Layout-conditioned
- 2B params
- 20k hours
- 12FPS on 1 GPU

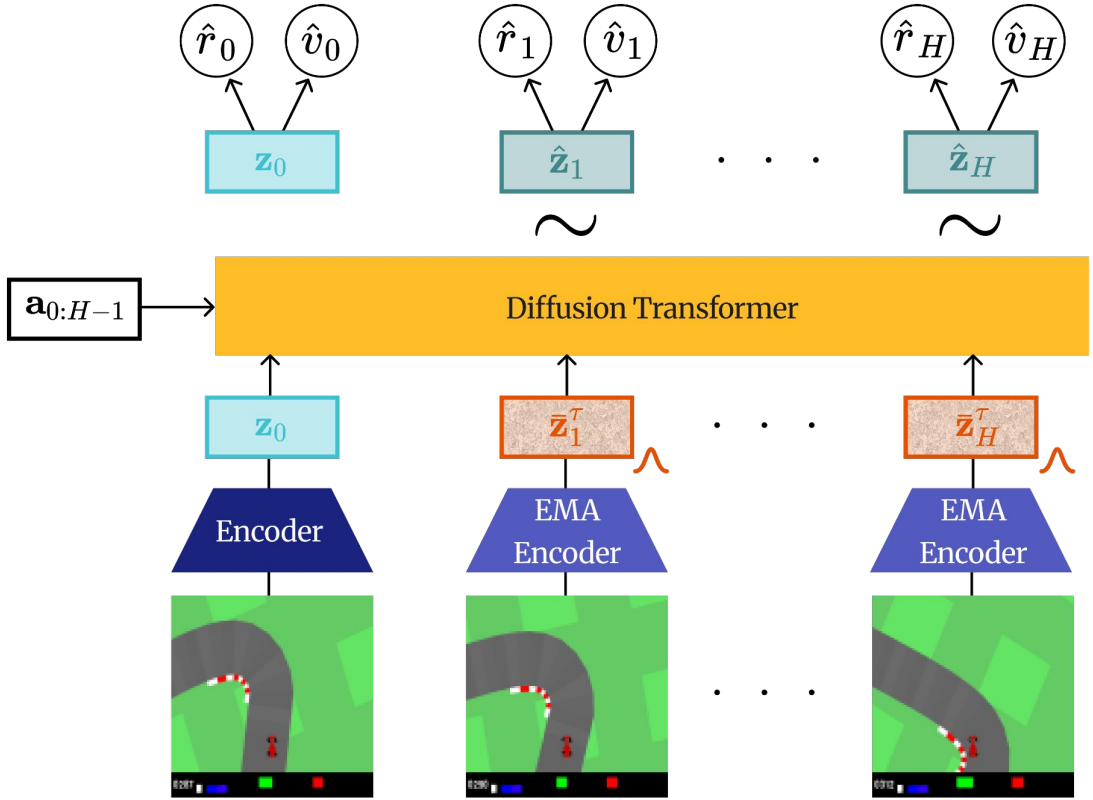
**GAIA-3**

- From scratch?
- 5 cams
- Layout-conditioned
- 15B params
- 140k hours
- ?

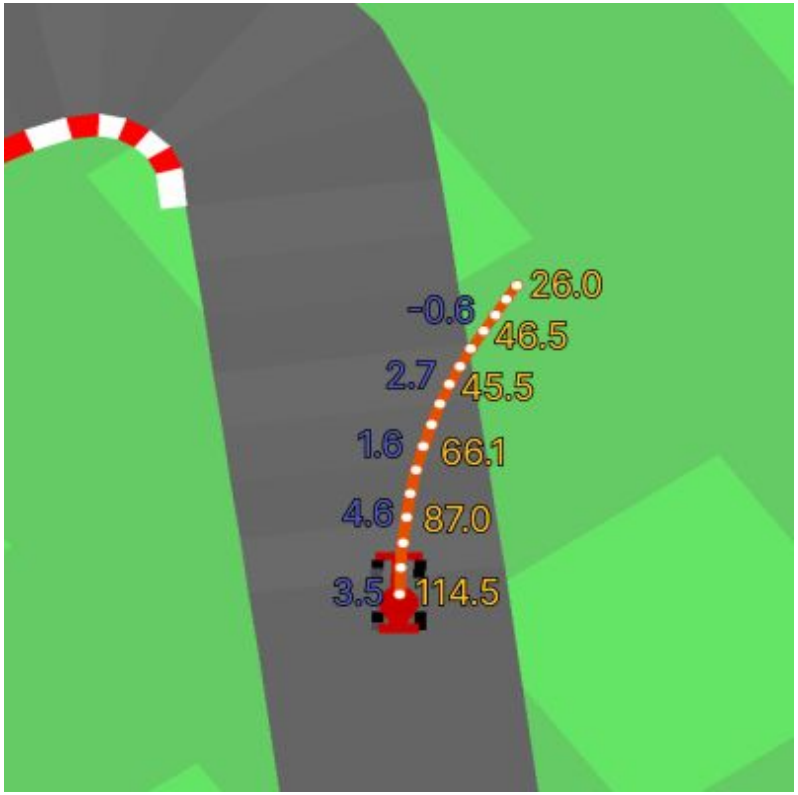
**WaymoWM**

- Genie 3-based
- 8 cams + LiDAR
- Optional conditions
- ?
- ?
- ?

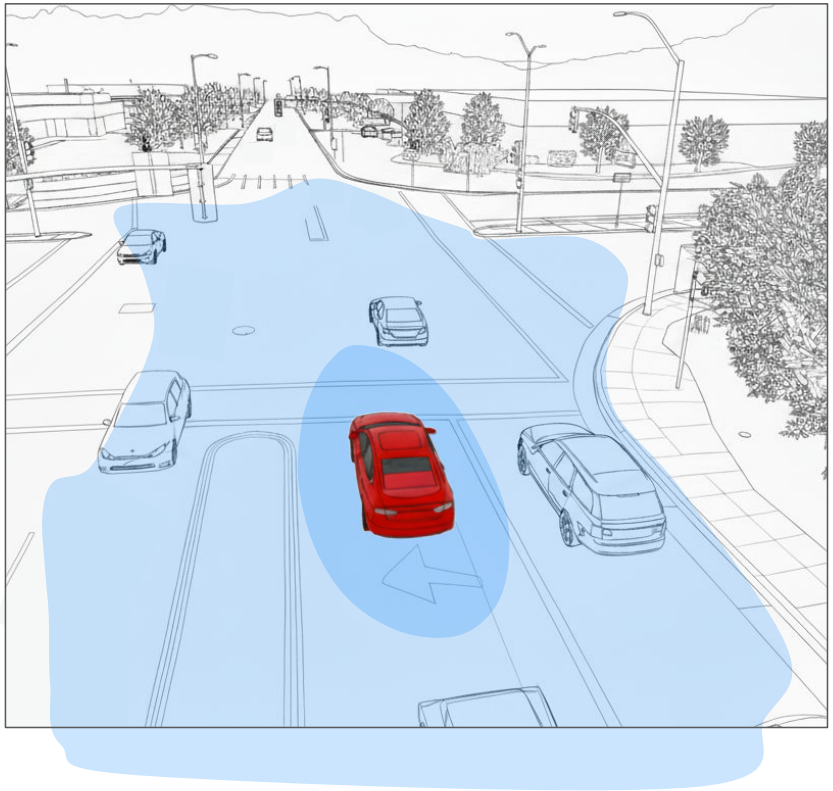
# Value Diffusion: First Demo



# Value Diffusion: First Results




# Generative World Models: Beyond Neural Rendering



Real-world drive

Geometric transformations

Neural reconstruction

Game engine-based rendering
Generative world models