OpenDriveLab

上海人工智能实验室
Shanghai Artificial Intelligence Laboratory

CVPR
SEATTLE, WA JUNE 17-21, 2024

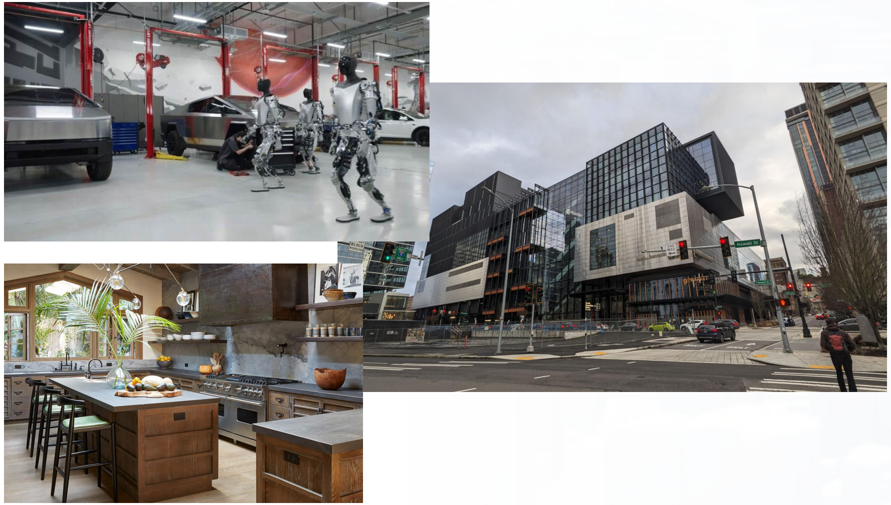
Visual World Models as “Foundation” Models for Autonomous Systems

Li Chen

OpenDriveLab at Shanghai AI Lab

June 17, 2024

Autonomous Systems (Agents)



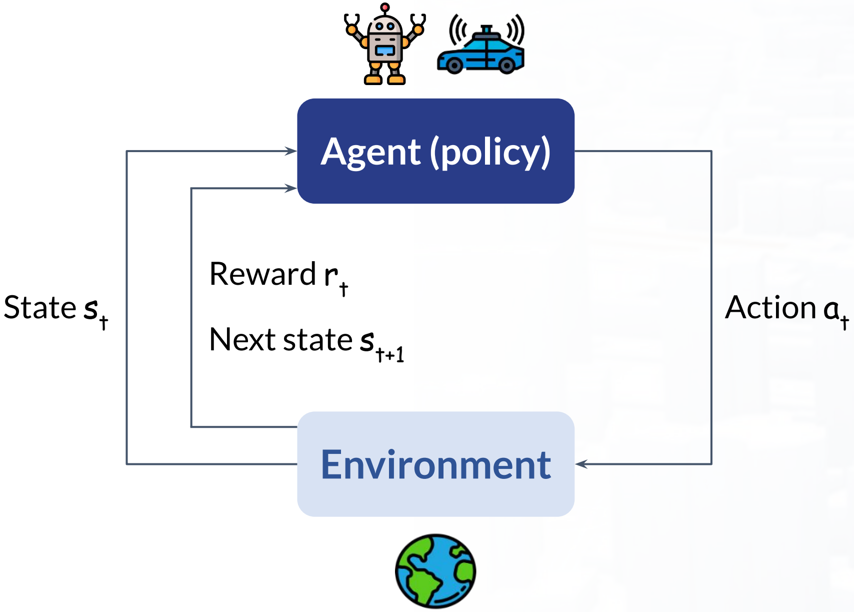
Environment

Multimodal contexts →

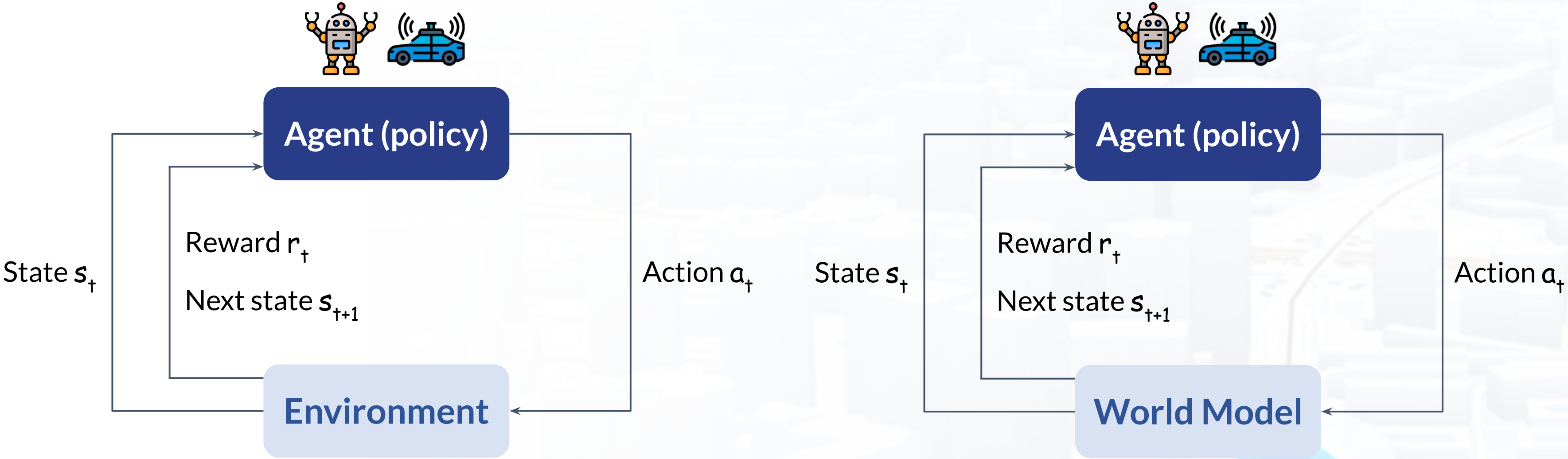
← Reason & Act (Interact)

Autonomous Systems (Agents)

World Model


Agent (policy)
State s_t
Reward r_t
Next state s_{t+1}
Action a_t
Environment

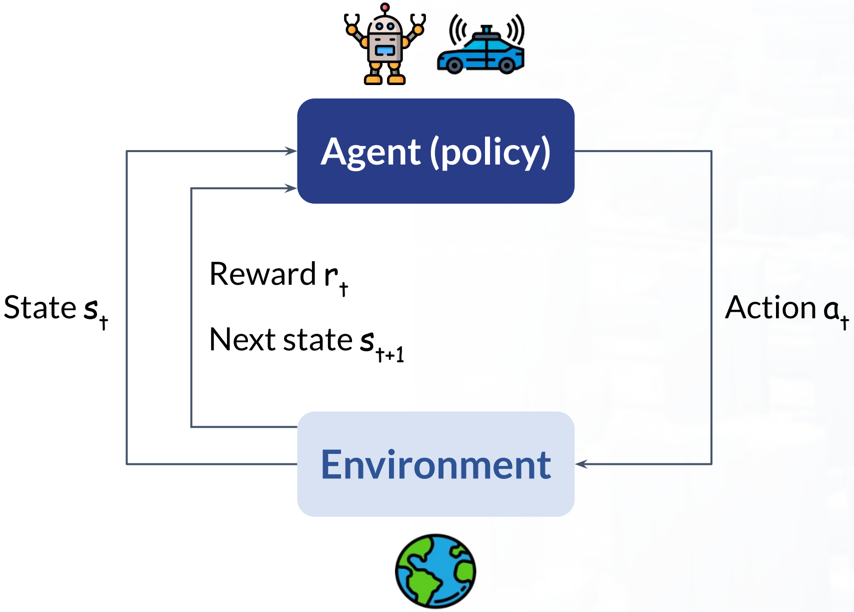
World Model



- **Selected concepts**, and **relationships** between them, to represent the whole system
- A **memory** component that makes predictions about **future** codes based on historical information
- Train a **simple controller** with the internal world model

[1] *D. Ha and J. Schmidhuber. Recurrent World Models Facilitate Policy Evolution. NeurIPS, 2018.*

World Model



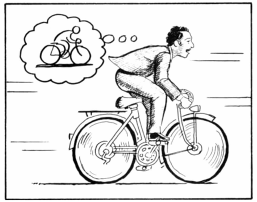
A Path Towards Autonomous Machine Intelligence Version — Yann Lecun

World Model

From simulated agents to real-world driving systems

RL Agents

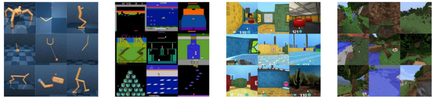
2018



World Models: Training agents inside their dreams

2020

Dreamer V1/2/3: Towards general agents with scalable world models

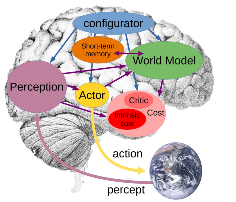


(a) Control Suite (b) Atari (c) DMLab (d) Minecraft

Vision

2022

Position Paper (by LeCun) Positioning the developments of world models



2023

I-JEPA: Capturing visual knowledge in self-supervised manner

Driving Robotics

2024

Scaling up world models on large corpus of videos

General World Model: inhouse data collected around the globe

TESLA

GAIA-1: 4700 hours of driving videos collected in London

Genie / UniPi & UniSim: Internet text-image, videos, human activities, robots, etc.

Foundation Models

Mind-blowing Part

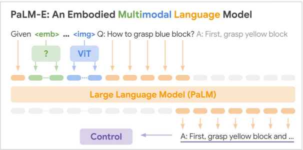
Weakness Samples

Are foundation models like Sora and LLMs world models?

Can Language Models Serve as Text-Based World Simulators?

Ruoyao Wang[†], Graham Todd[‡], Ziang Xiao[♠], Xingdi Yuan[◇]
Marc-Alexandre Côté[◇], Peter Clark[♣], Peter Jansen[†♣]
[†]University of Arizona [◇]Microsoft Research Montréal
[‡]New York University [♠]Johns Hopkins University [♣]Allen Institute for AI
{ruoyaowang,pajansen}@arizona.edu gdrtodd@nyu.edu
ziang.xiao@jhu.edu {eric.yuan,macote}@microsoft.com
PeterC@allenai.org

- Large corpus of data
- Effective generalization
- Diverse range of use cases
- Self-supervision (generally)



“Foundation” Models for Autonomous Systems

Towards Intelligent, Reliable and Generalizable System

– “Foundation” Models for Autonomous Systems

Foundation Model:

- Large corpus of data → **Raw data** | Labeled data
- Effective generalization → **World knowledge** | Task-wise optimization
- Diverse range of use cases → **Self-supervised learning** | Supervised learning
- Self-supervision (generally)

Representation Learning x **Visual World Models** | Specific Task Models

Summary (Questions)

Data

- **Question 1:** How can we find large corpus of data for autonomous driving, which helps effective generalization ability?

Model

- **Question 1:** How can we train a world model with intricate world knowledge, with self-supervised learning?

Application

- **Question 1:** What are the abilities of the world model?

Highlight

Thu. 20 Jun 5 p.m – 6:30 p.m

Arch 4A-E Poster #5

Generalized Predictive Model for Autonomous Driving

Jiazhi Yang

Shenyuan Gao

Yihang Qiu

Li Chen

Tianyu Li

Bo Dai

Kashyap Chitta

Penghao Wu

Jia Zeng

Ping Luo

Jun Zhang

Andreas Geiger

Yu Qiao

Hongyang Li

- arXiv: https://arxiv.org/abs/2403.09630
- dataset: https://github.com/OpenDriveLab/DriveAGI

GenAD | At a Glance

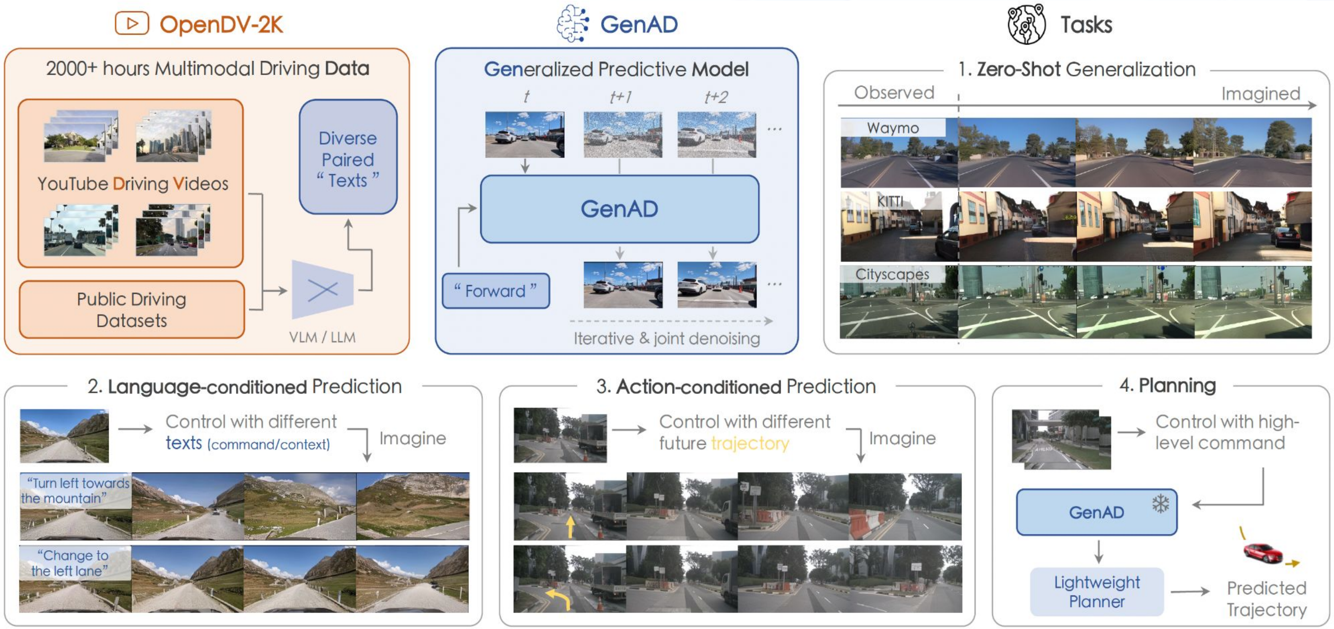
- arXiv: https://arxiv.org/abs/2403.09630
- dataset: https://github.com/OpenDriveLab/DriveAGI

Highlight

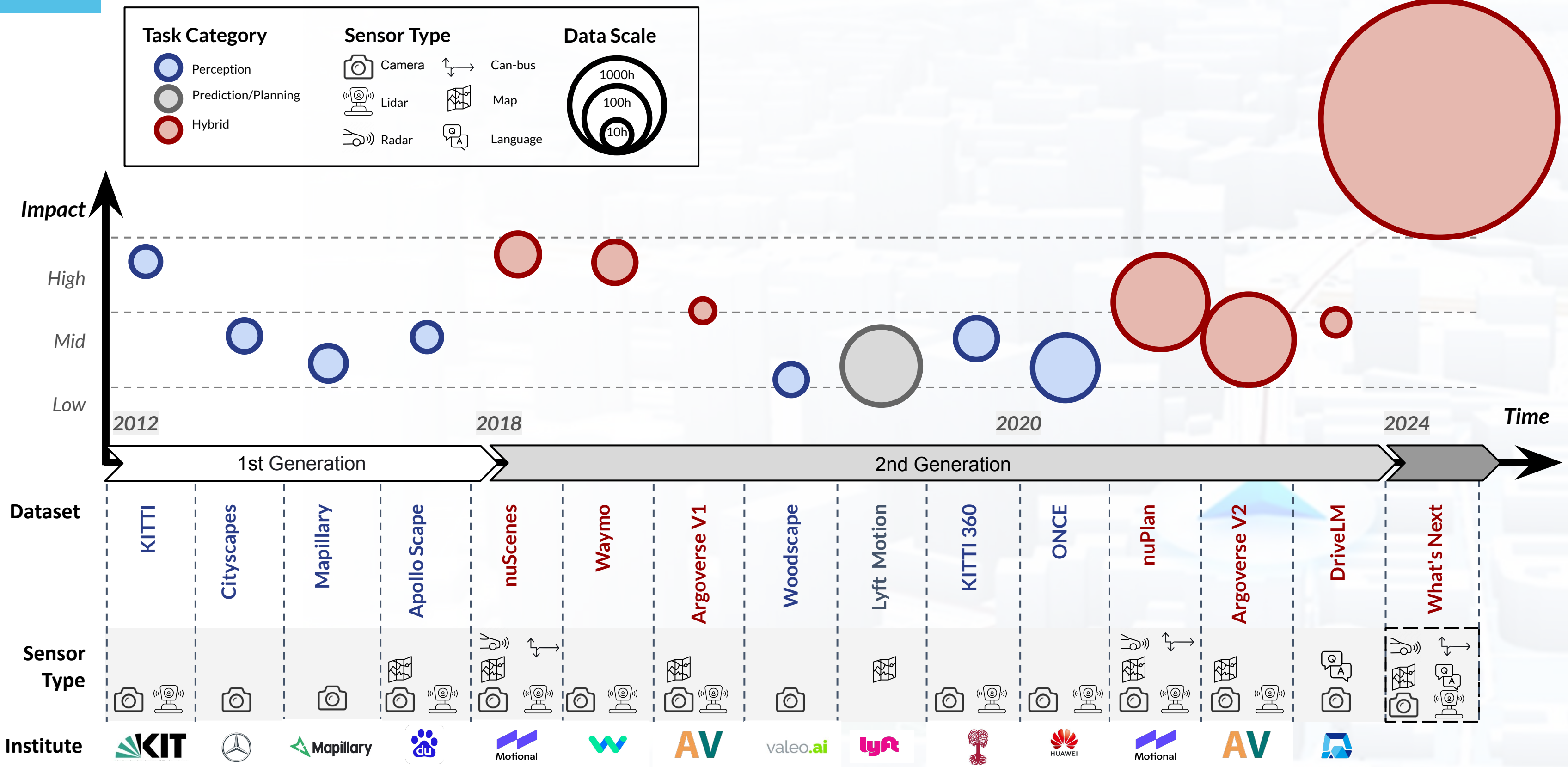
Thu. 20 Jun 5 p.m – 6:30 p.m

Arch 4A-E Poster #5

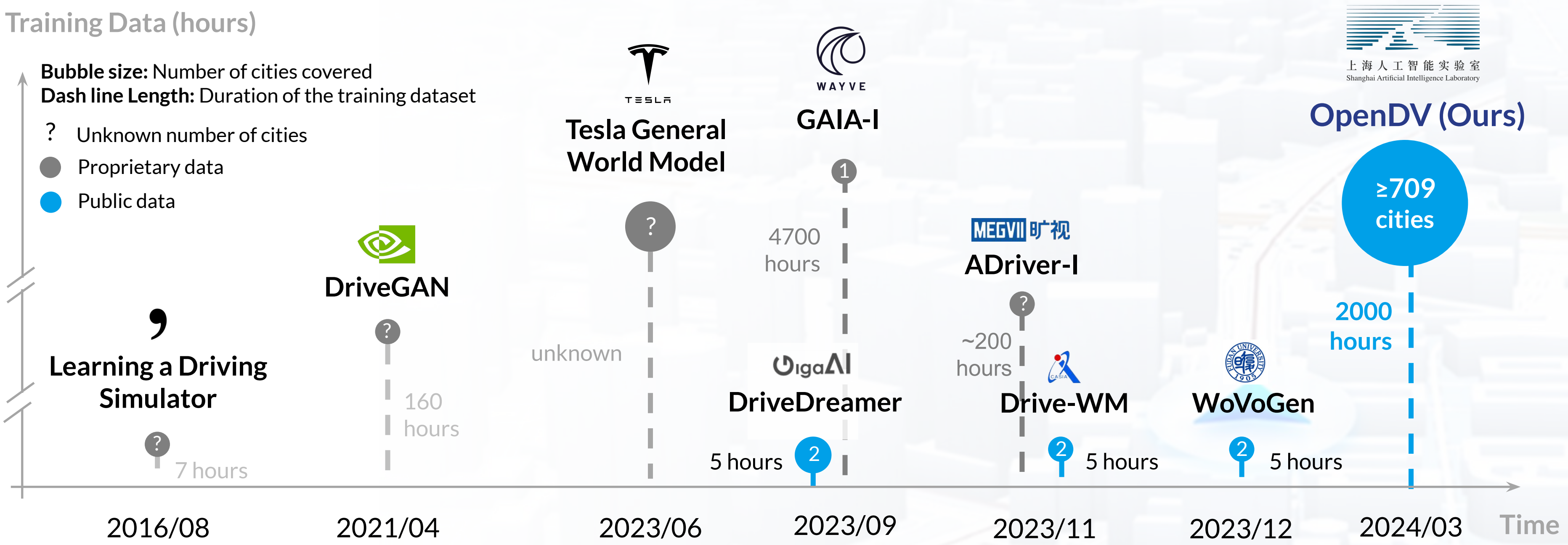
A large-scale video prediction model on **web-scale driving videos**, to enable its **generalization across** a wide spectrum of **domains and tasks**.



Dataset in Autonomous Driving



Data | Scale-up Driving Videos

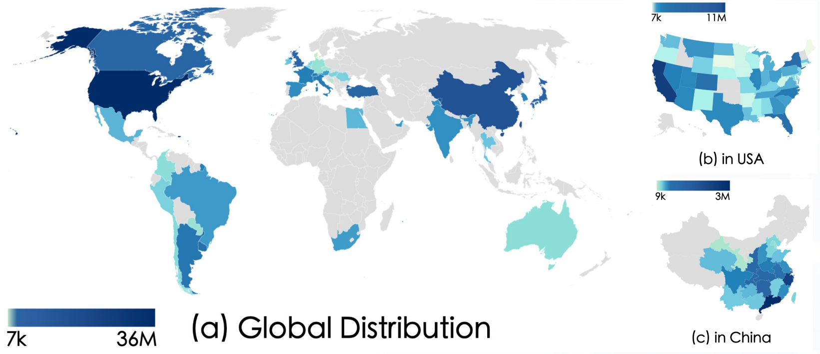


OpenDV: the largest public driving video datasets

Data | OpenDV

Massive YouTube videos, collected worldwide

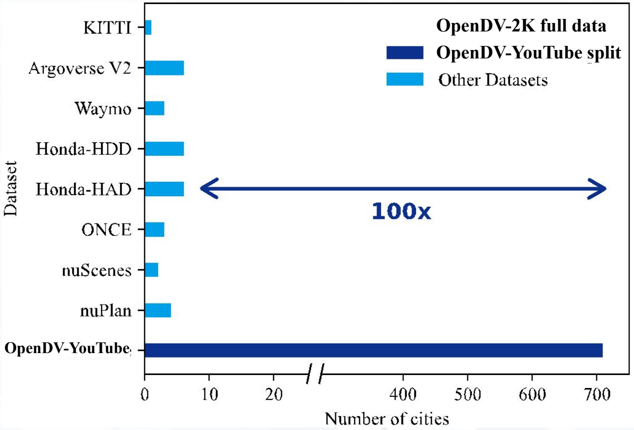
- Diverse, in geography, weather, scenes, traffic, etc.
- No label (vehicle action, 3D boxes, calibrations, etc.)

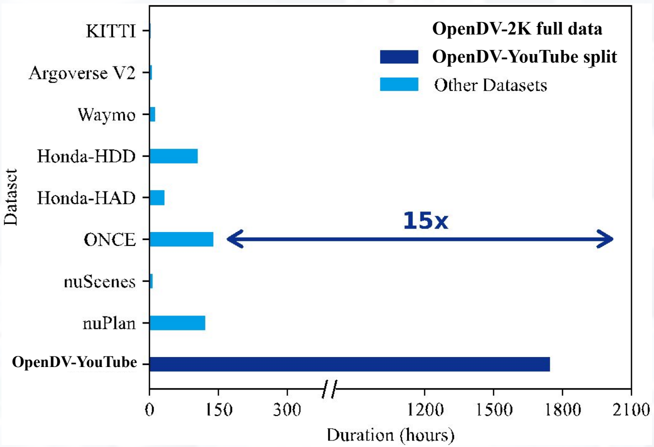


Data | OpenDV

- ***Largest public dataset*** **up-to-date for autonomous driving**
- **2059 hours, 709 areas**

	Dataset	Duration (hours)	Front-view Frames	Geographic Diversity Countries	Geographic Diversity Cities	Sensor Setup
✗	KITTI [14]	1.4	15k	1	1	fixed
✗	Cityscapes [10]	0.5	25k	3	50	fixed
✗	Waymo Open* [41]	11	390k	1	3	fixed
✗	Argoverse 2* [45]	4.2	300k	1	6	fixed
✓	nuScenes [6]	5.5	241k	2	2	fixed
✓	nuPlan [7]	120	4.0M	2	4	fixed
✓	Talk2Car [12]	4.7	-	2	2	fixed
✓	ONCE [32]	144	7M	1	-	fixed
✓	Honda-HAD [23]	32	1.2M	1	-	fixed
✓	Honda-HDD-Action [38]	104	1.1M	1	-	fixed
✓	Honda-HDD-Cause [38]	32	-	1	-	fixed
✓	OpenDV-YouTube (Ours)	1747	60.2M	$\geq 40^{\dagger}$	$\geq 709^{\dagger}$	uncalibrated
-	**OpenDV-2K** (Ours)	**2059**	**65.1M**	$\geq \mathbf{40}^{\dagger}$	$\geq \mathbf{709}^{\dagger}$	**uncalibrated**



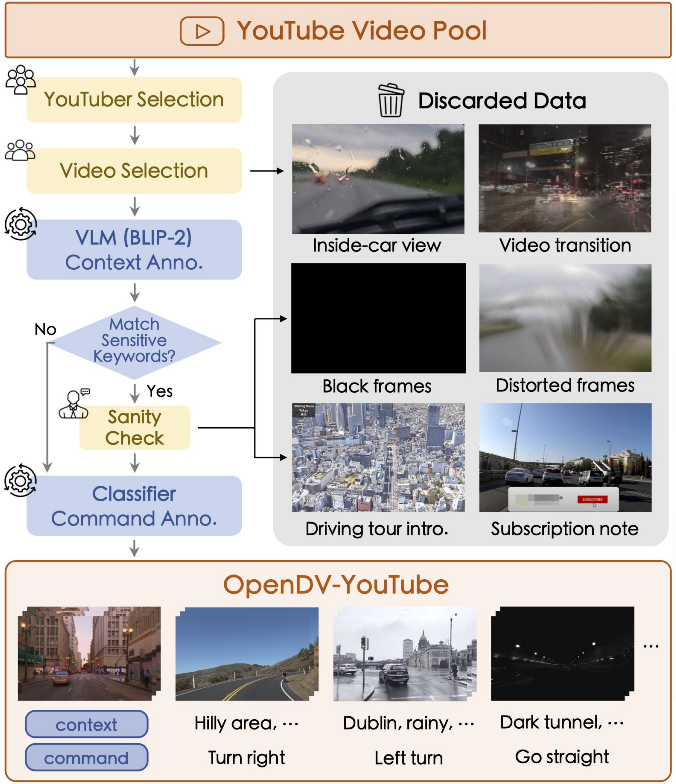


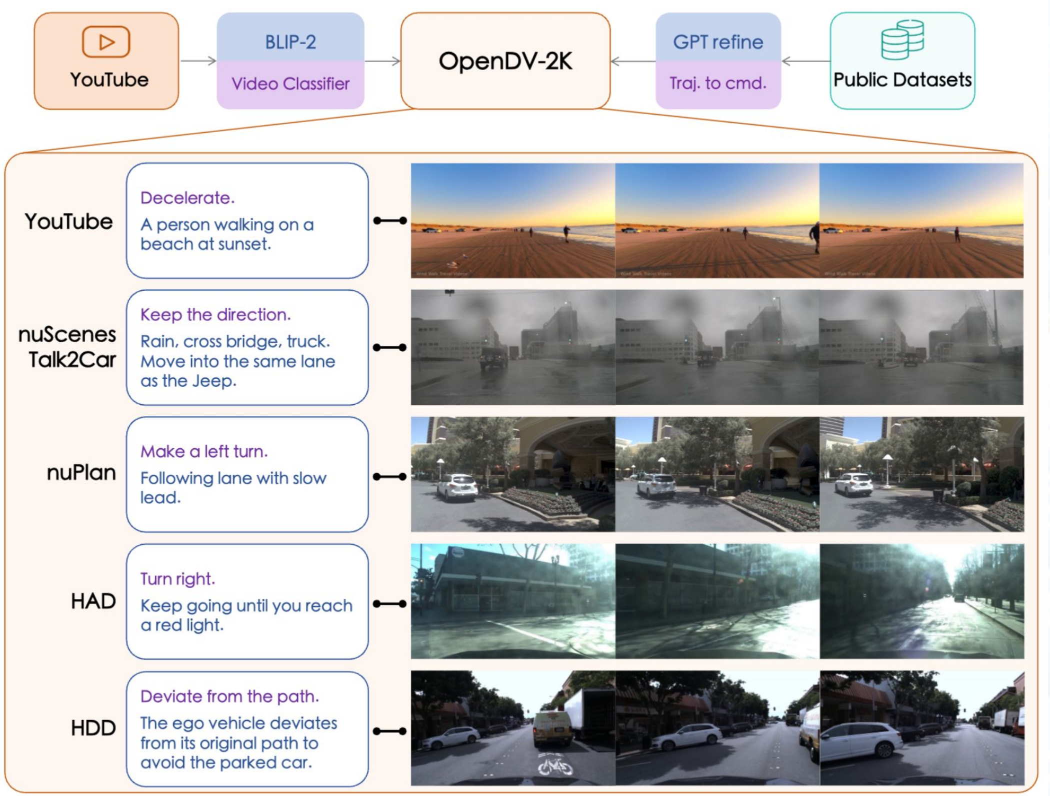
OpenDV-2K (Ours) 🚀

- arXiv: https://arxiv.org/abs/2403.09630
- dataset: https://github.com/OpenDriveLab/DriveAGI

Data | OpenDV

- arXiv: https://arxiv.org/abs/2403.09630
- dataset: https://github.com/OpenDriveLab/DriveAGI

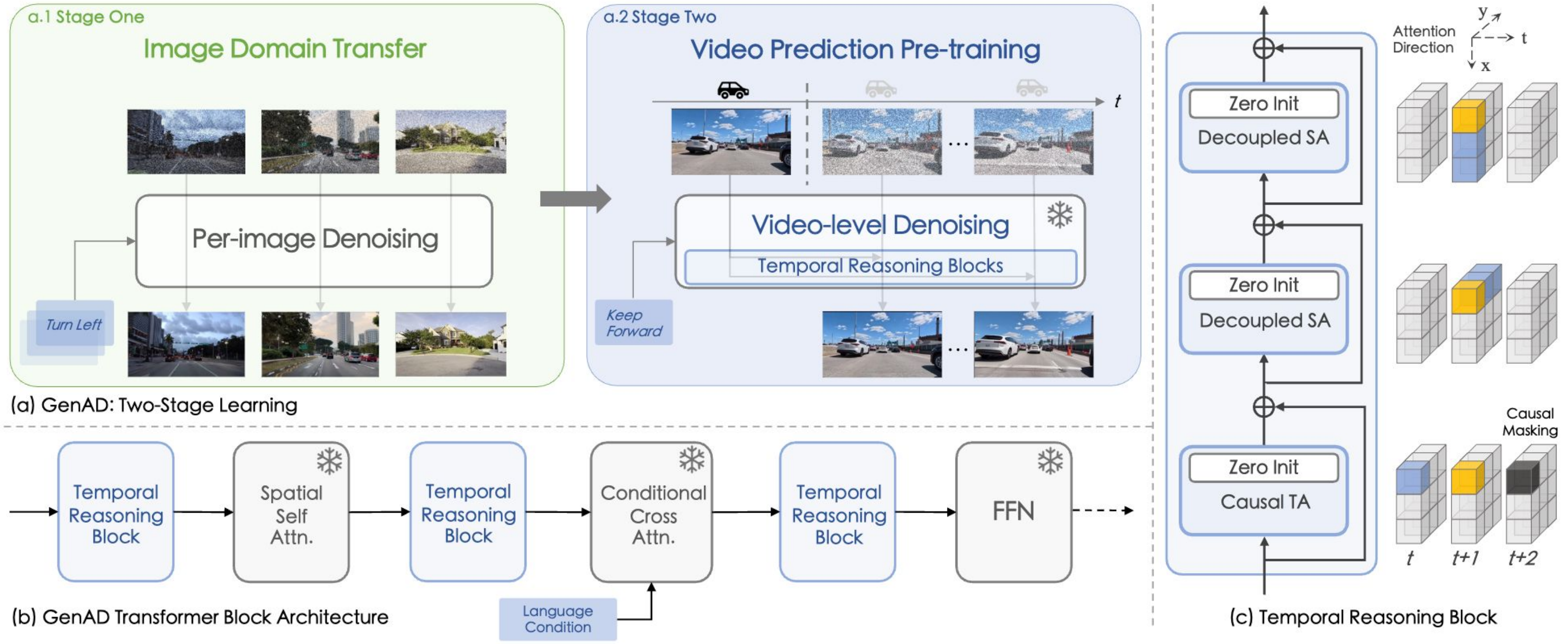




Model | Video Prediction Model for Driving

Keys

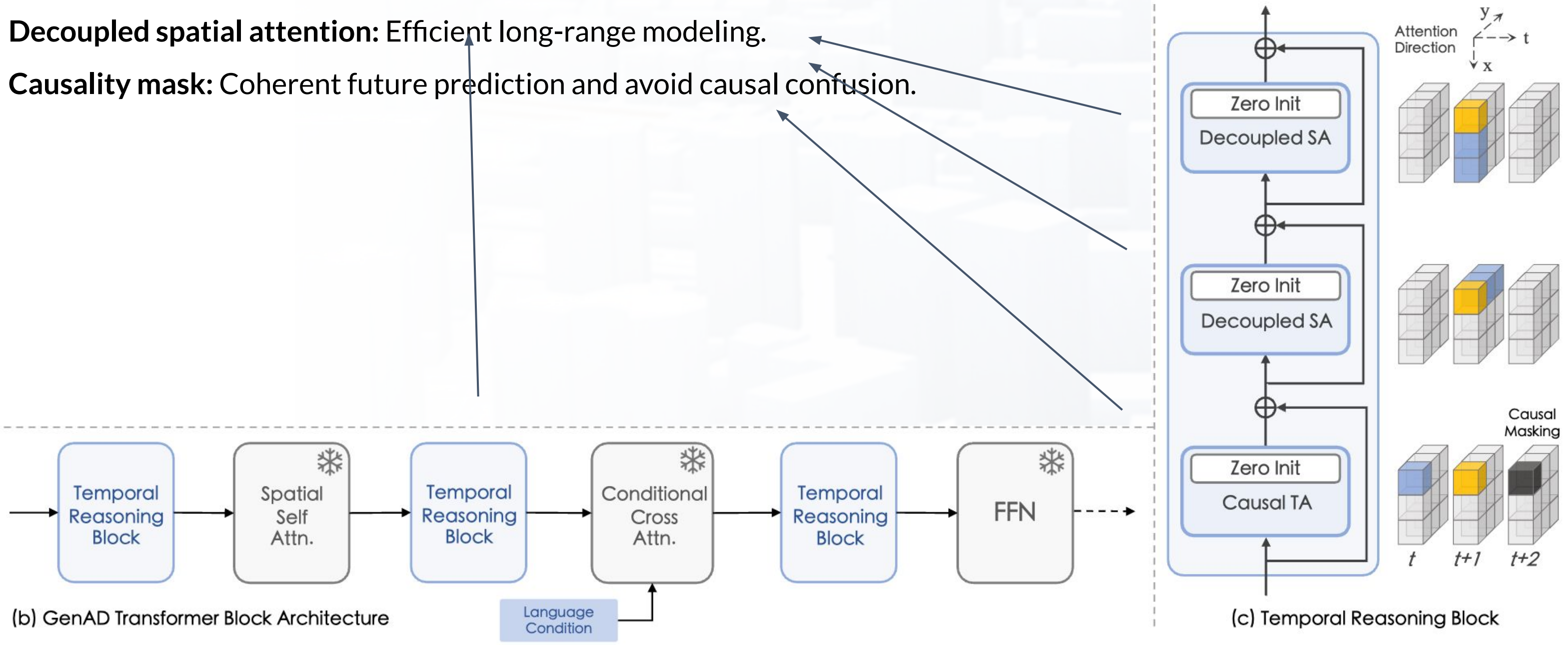
- **GenAD** (5.9B) = SDXL (2.7B) + Temporal Reasoning Blocks (2.5B) + CLIP-Text (0.7B)
- Tuning the **image generation model** into a highly-capable **video prediction model**



Model | Video Prediction Model for Driving

Designs

- **Interleaved temporal blocks:** Sufficient spatiotemporal interaction.
- **Decoupled spatial attention:** Efficient long-range modeling.
- **Causality mask:** Coherent future prediction and avoid causal confusion.



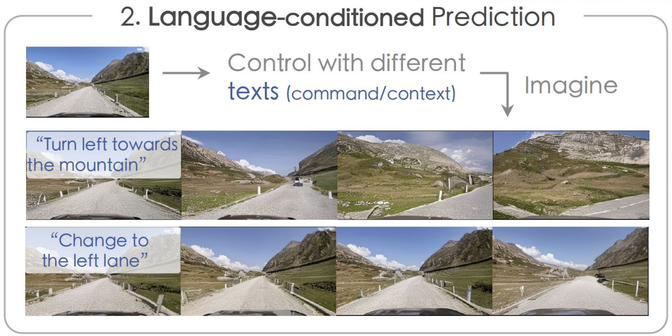
(b) GenAD Transformer Block Architecture

(c) Temporal Reasoning Block

Tasks | Zero-shot Generalization (Video Prediction)

Zero-shot video prediction on unseen datasets including Waymo, KITTI and Cityscapes

Tasks | Language-conditioned Prediction

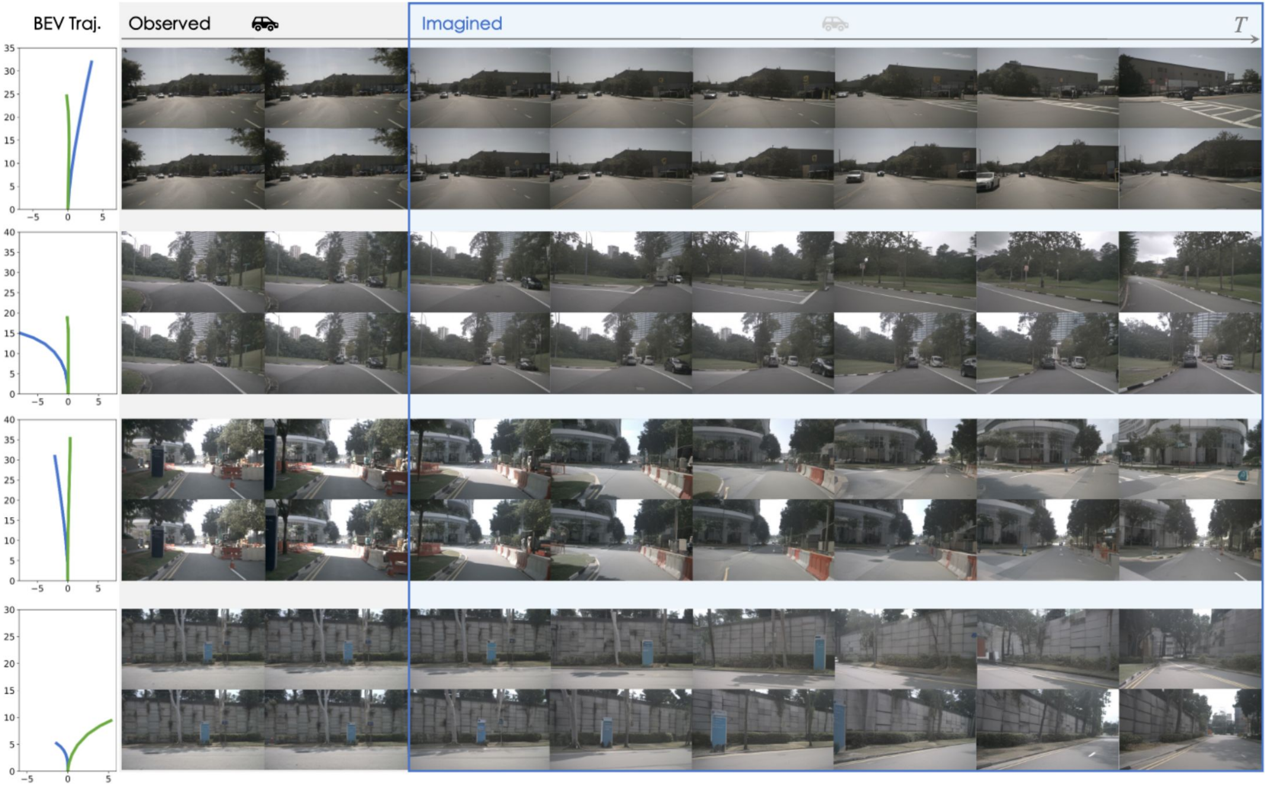


Instruct the future with **free-form texts.**

Tasks | Action-conditioned Prediction (Simulation)

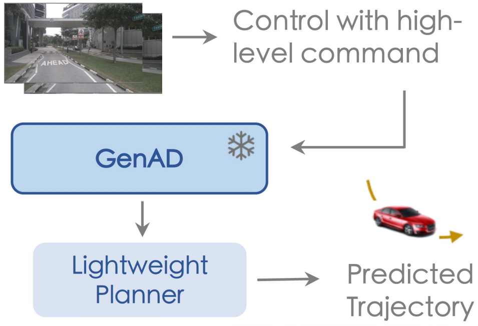
Method	Condition	nuScenes Action Prediction Error (↓)
Ground truth	-	0.9
GenAD	text	2.54
GenAD-act	text + traj.	**2.02**

Table 4. **Task on Action-conditioned prediction**. Compared to GenAD with text conditions only, GenAD-act enables more precise future predictions that follow the action condition.



Simulate the future differently conditioned on **future trajectory.**

Tasks | Planning (Representation Learning)



Method	# Trainable Params.	nuScenes ADE (↓)	nuScenes FDE (↓)
ST-P3* [20]	10.9M	2.11	2.90
UniAD* [22]	58.8M	1.03	1.65
GenAD (Ours)	0.8M	1.23	2.31

- Speeding up training by **3400 times** (*vs. **UniAD***) w/o ego status

Summary

Data

- **Takeaway 1:** Largest available driving video dataset: OpenDV (2000+ hours). The great diversity ensures generalization.

Model

- **Takeaway 1:** Can be a video prediction model conditioned on high-level instructions.

Application

- **Takeaway 1:** Learned representations can be simply trained for policy prediction.

Summary (Question)

Data

- **Takeaway 1:** Largest available driving video dataset: OpenDV (2000+ hours). The great diversity ensures generalization.

Model

- **Takeaway 1:** Can be a video prediction model conditioned on high-level instructions.
- **Question 2**: How about more direct conditions (in the real world)?

Application

- **Takeaway 1:** Learned representations can be simply trained for policy prediction.
- **Question 2**: How about the typical application such as rewarding for model-based RL?

- arXiv: https://arxiv.org/abs/2405.17398
- demo page: https://vista-demo.github.io/
- code: https://github.com/OpenDriveLab/Vista

Vista: A Generalized Driving World Model with High Fidelity and Versatile Controllability


Shenyuan Gao


Jiazhi Yang


Li Chen


Kashyap Chitta


Yihang Qiu


Andreas Geiger

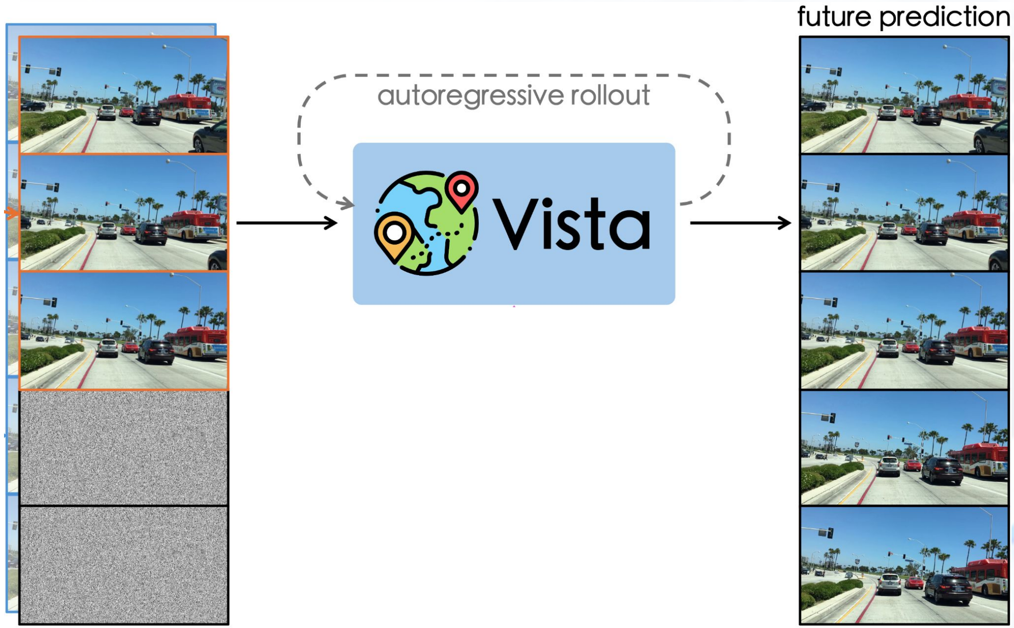

Jun Zhang


Hongyang Li

Driving World Models

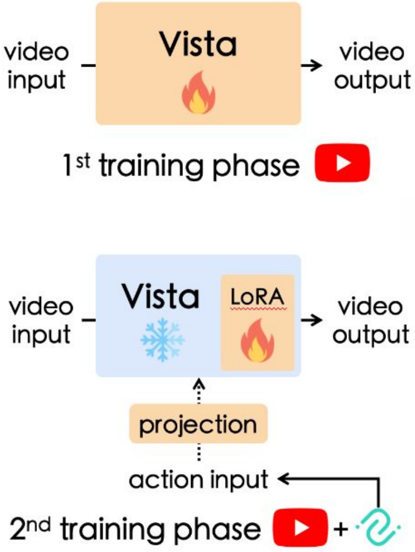
Method	Model Setups			Action Control Modes			
	Data Scale	Frame Rate	Resolution	Angle&Speed	Trajectory	Command	Goal Point
DriveSim [99]	7h	5 Hz	80×160	✓			
DriveGAN [66]	160h	8 Hz	256×256	✓			
DriveDreamer [122]	5h	12 Hz	128×192	✓			
Drive-WM [124]	5h	2 Hz	192×384		✓		
WoVoGen [87]	5h	2 Hz	256×448	✓			
ADriver-I [60]	300h	2 Hz	256×512			✓	
GenAD [133]	2000h	2 Hz	256×448		✓	✓	
GAIA-1 [53]	4700h	25 Hz	288×512	✓			
Vista (Ours)	1740h	10 Hz	576×1024	✓	✓	✓	✓

Vista | Versatile action controllability



From high-level intentions (command, goal point) to low-level maneuvers (trajectory, angle, and speed)

Vista | Model

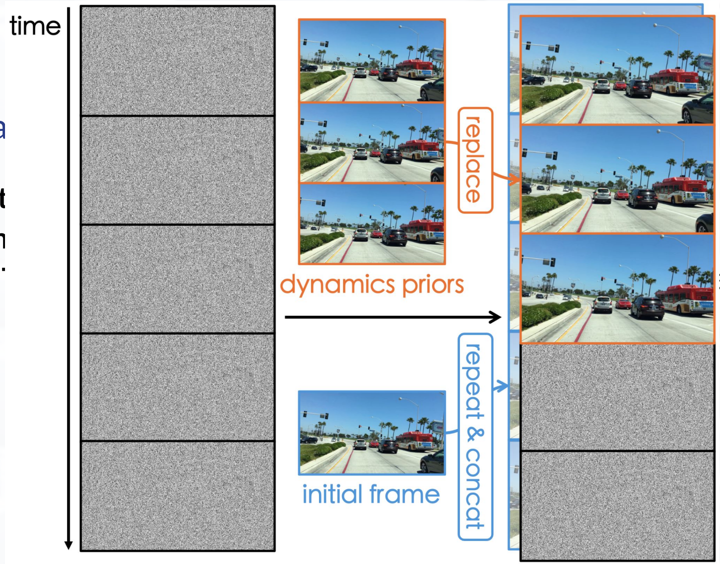


High-fidelity

- **Dynamic Prior Injection:** Replacing the latent to absorb varying numbers of condition frames
- **Dynamics Enhancement** Loss: Dynamics-aware weight to highlight dynamic regions
- **Structure Preservation** Loss: Preserve high-frequency structured features

Versatile Controlla

- **Unified Condit**
- **Efficient Learn**
- **Collaborative** er generalization

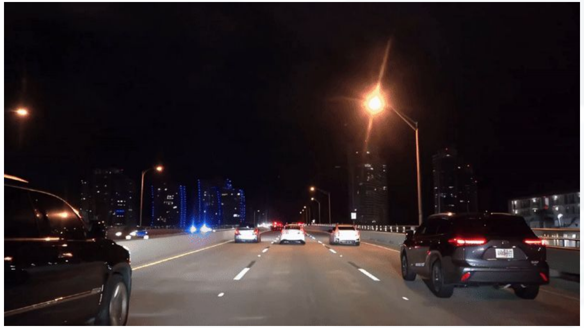


Vista | Video Prediction

- High-fidelity future prediction



- Continuous long-horizon rollout (15 seconds)


Vista | Zero-shot Action Controllability

turn left | go straight | turn right | stop

* The commands above are translated from trajectories, or angles+speed.

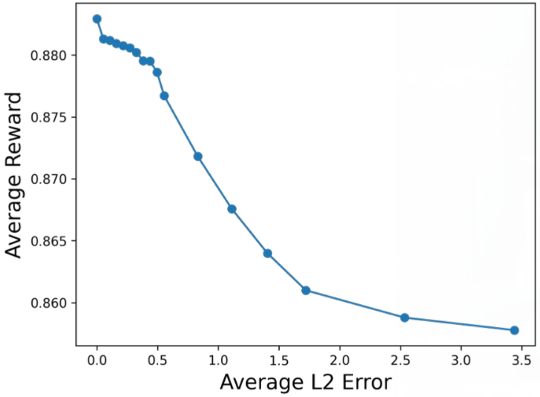
Vista | Reward

- Drive-WM rewards

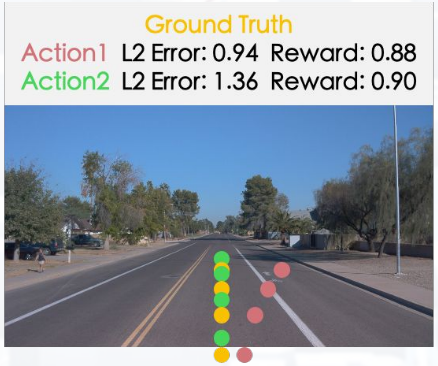
  Drive-WM, CVPR'24

Detection Model → Rewards

- Provide reward without ground truth actions, by uncertainty



Reasonable rewards



More reasonable than ADE

Summary (Question)

Data

- **Takeaway 1:** Largest available driving video dataset: OpenDV (2000+ hours). The great diversity ensures generalization.

Model

- **Takeaway 1:** Can be a video prediction model conditioned on high-level instructions.
- **Takeaway 2**: We can inject kinds of conditions with efficiently to make it a real world model / simulator.

Application

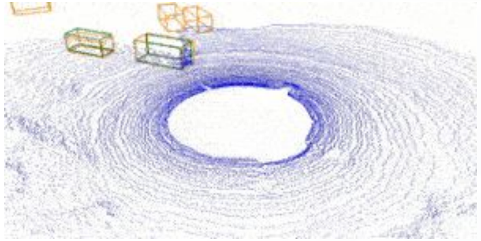
- **Takeaway 1:** Learned representations can be simply trained for policy prediction.
- **Takeaway 2**: The stochastic diffusion process learns inherent rewards.

Can we have more industry-friendly approaches, including data, model, and tasks' application?
Also, evaluations?

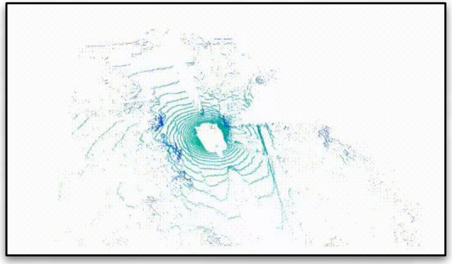
ViDAR | Motivation

+ Action

Future Prediction Model → World Model



S2Net — Point cloud future prediction for planning



4D-Occ — Ego Future Tracjectory

Motivation

- The industry has accumulated huge amount of **image-LiDAR** data with test vehicles
- image-LiDAR naturally has both **semantic** and **geometric** clues

Pre-training with Point Cloud & Visual Image

[1] *Weng et al. S2Net: Stochastic Sequential Pointcloud Forecasting. ECCV, 2022.*
[2] *Khurana et al. Point Cloud Forecasting as a Proxy for 4D Occupancy Forecasting. CVPR, 2023.*

ViDAR | Motivation

VIDAR in multi-view stereo (from Mobileye, CES 2021)

Note:

- Reconstruction purpose
- Lack of exploration in temporal dimension
- More geometric estimation, lack of the reasoning ability of the environment

ViDAR | Motivation

VIDAR in depth estimation (from TRI)

TRI-VIDAR

Installation | Configuration | Datasets | Visualization | Publications | License

TOYOTA RESEARCH INSTITUTE

Official PyTorch repository for some of TRI's latest publications, including self-supervised learning, multi-view geometry, and depth estimation. Our goal is to provide a clean environment to reproduce our results and facilitate further research in this field. This repository is an updated version of PackNet-SfM, our previous monocular depth estimation repository, featuring a different license.

Note:

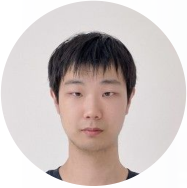
- Reconstruction purpose
- Lack of exploration in temporal dimension
- More geometric estimation, lack of the reasoning ability of the environment

Highlight

Thu. 20 Jun 5 p.m — 6:30 p.m

Arch 4A-E Poster #6

Visual Point Cloud Forecasting enables Scalable Autonomous Driving


Jiazhi Yang


Li Chen


Yanan Sun


Hongyang Li

- arXiv: https://arxiv.org/abs/2312.17655
- code: https://github.com/OpenDriveLab/ViDAR

ViDAR | At a Glance

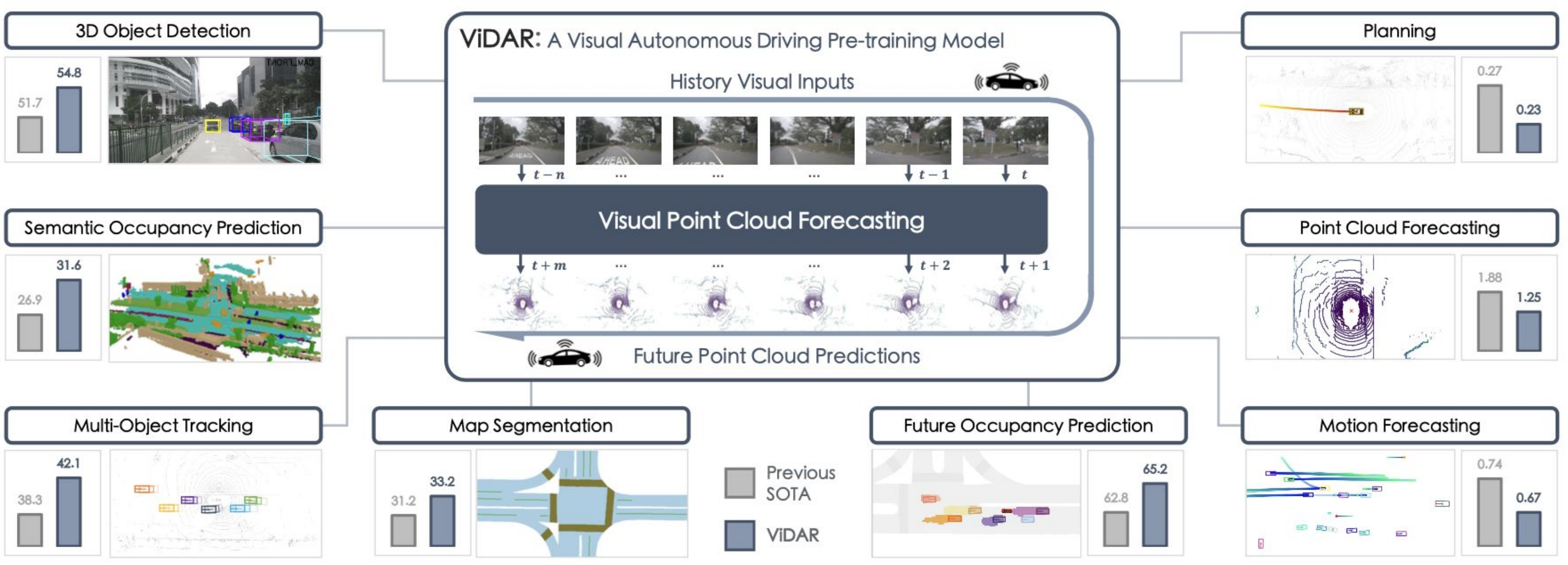
Highlight

Thu. 20 Jun 5 p.m – 6:30 p.m

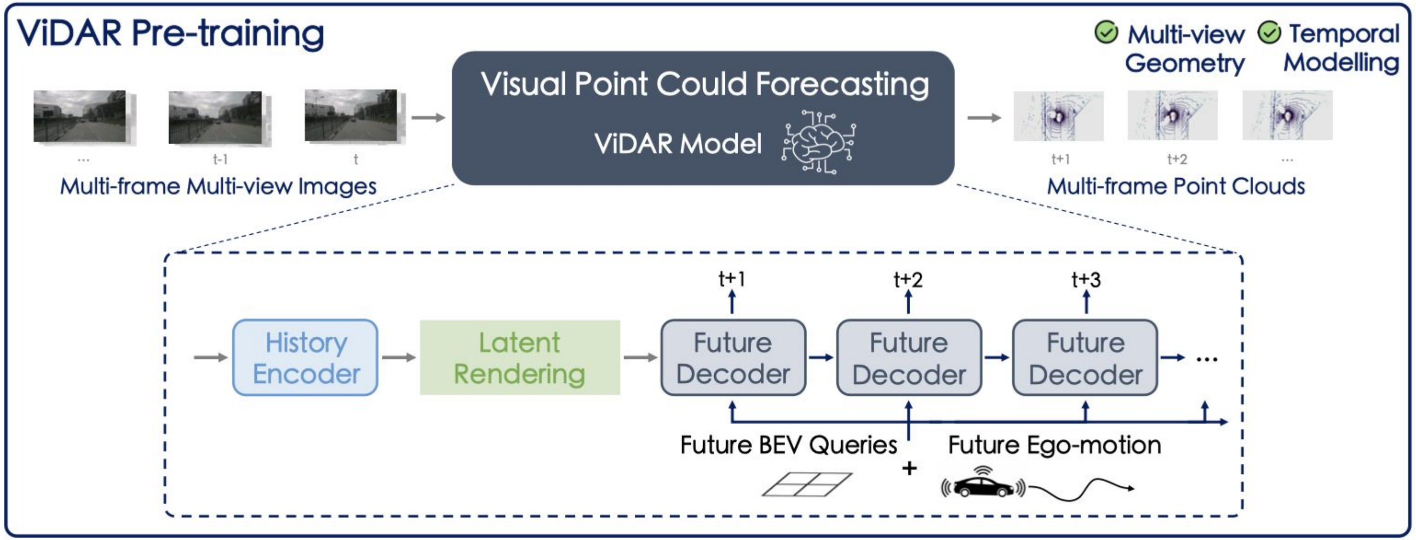
Arch 4A-E Poster #6

Training multimodal world model by **Visual Point Cloud Forecasting** and boosting **End-to-End Autonomous Driving**.

- arXiv: https://arxiv.org/abs/2312.17655
- code: https://github.com/OpenDriveLab/ViDAR



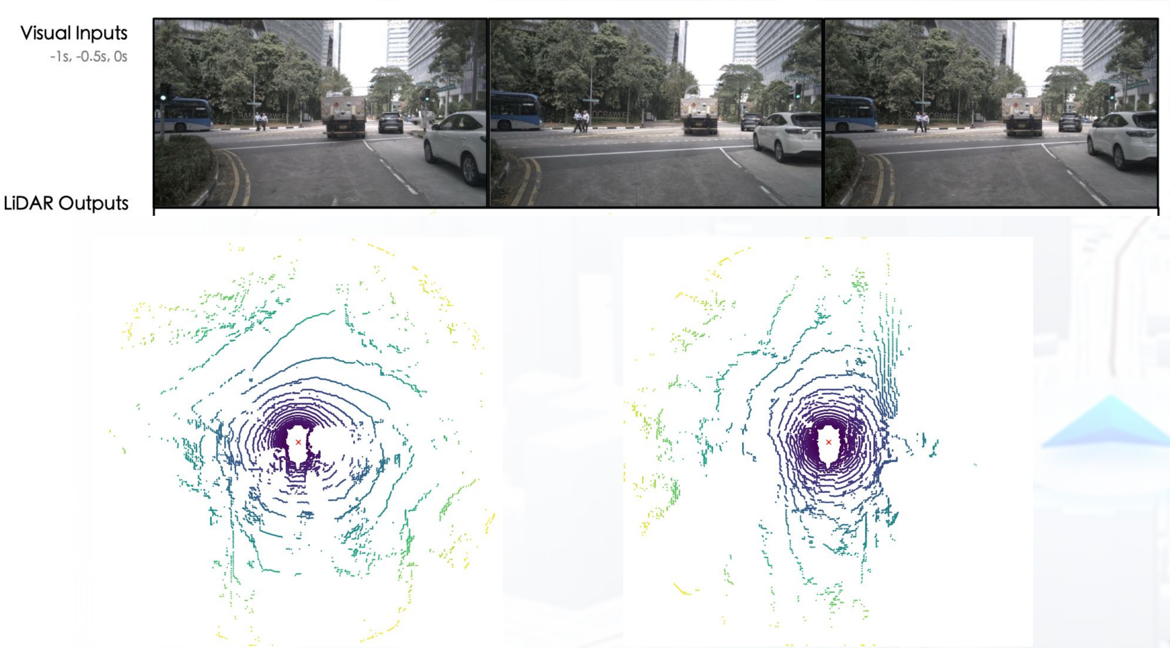
ViDAR | Architecture



- **History Encoder:** Target pre-training structure, extracting BEV embeddings from visual inputs.
- **Latent Rendering:** Extract geometric latent space. Removing ray-shape ambiguities by volume rendering in feature space.
- **Future Decoder:** Iteratively predict future BEV features, conditioned on ego-motion.

ViDAR | World Model in Driving

The First Multimodal World Model

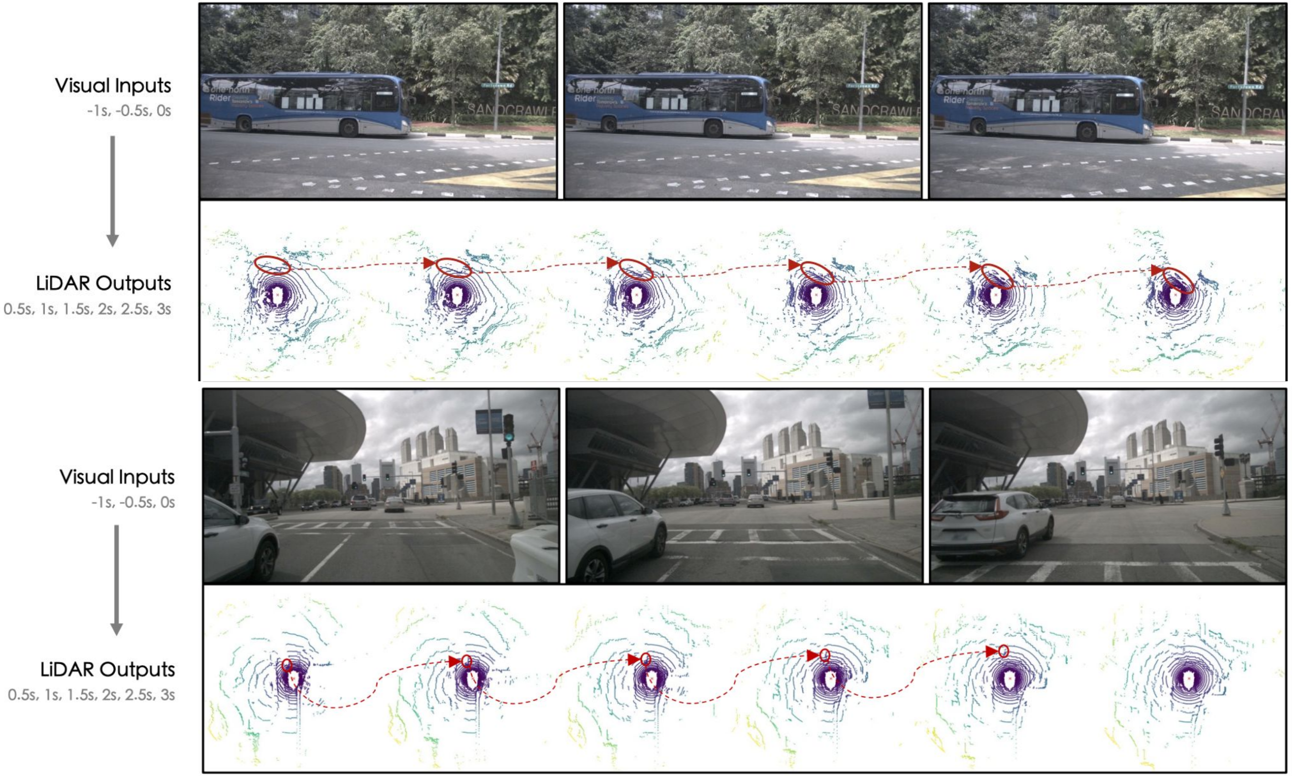


Go straight

Turn left

ViDAR | Future Prediction

ViDAR effectively models relative motion, and motion of other objects.



ViDAR | Downstream Tasks

Pre-training by visual point cloud forecasting helps **end-to-end autonomous driving**

Method	Detection		Tracking			Mapping		Motion Forecasting			Future Occupancy Prediction				Planning	
	NDS ↑	mAP ↑	AMOTA↑	AMOTP↓	IDS↓	IoU-lane↑	IoU-road↑	minADE↓	minFDE↓	MR↓	IoU-n.↑	IoU-f.↑	VPQ-n.↑	VPQ-f.↑	avg.L2↓	avg.Col.↓
UniAD	49.36	37.96	38.3	1.32	1054	31.3	69.1	0.75	1.08	0.158	62.8	40.1	54.6	33.9	1.12	0.27
ViDAR	**52.57**	**42.33**	**42.0**	**1.25**	**991**	**33.2**	**71.4**	**0.67**	**0.99**	**0.149**	**65.4**	**42.1**	**57.3**	**36.4**	**0.91**	**0.23**

Summary

Data

- **Takeaway 1:** Largest available driving video dataset: OpenDV (2000+ hours). The great diversity ensures generalization.
- **Takeaway 2**: The image and LiDAR pairs are very helpful to capture both semantic and geometric information in the environment.

Model

- **Takeaway 1:** Can be a video prediction model conditioned on high-level instructions.
- **Takeaway 2**: We can inject kinds of conditions with efficiently to make it a real world model / simulator.
- **Takeaway 3**: BEV-based models (c.f. videos) are also effective world models.

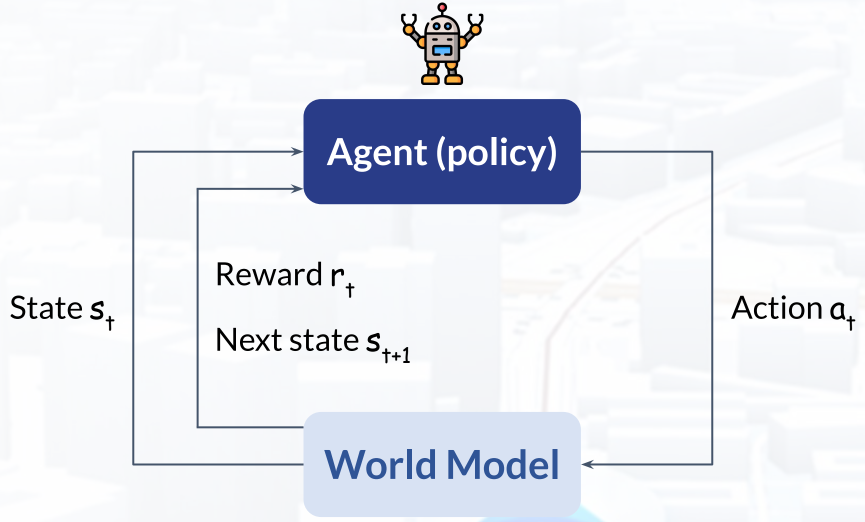
Application

- **Takeaway 1:** Learned representations can be simply trained for policy prediction.
- **Takeaway 2**: The stochastic diffusion process learns inherent rewards.
- **Takeaway 3**: Spatio-temporal pre-training improves all tasks in driving and serves as a foundation model.

How about robotics?

Challenges

- Heavy interactions between robots and environments
- More diverse tasks and environments



Visual data

World knowledge

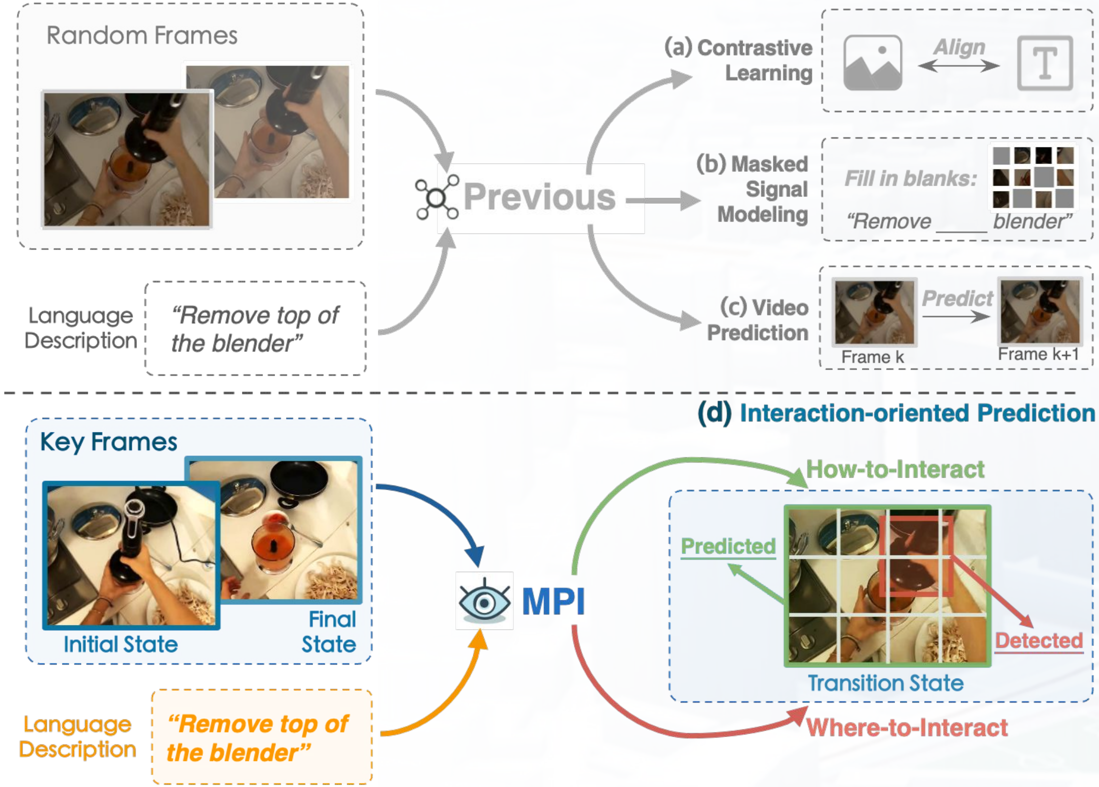
Representation learning

Visual World Models

w/ Highlighted Interaction

Learning Manipulation by Predicting Interaction (MPI)

Robotics: Science and Systems (RSS) 2024, Delft, Netherlands



- arXiv: https://arxiv.org/abs/2406.00439
- project page: https://opendrivelab.com/MPI
- code: https://github.com/OpenDriveLab/MPI

Existing works

- High-level semantics
- Or low-level details

MPI (Ours)

- Interactive dynamics (patterns of behavior and physical interactions)
- w/ both high-level semantics and low-level details

MPI | Interaction Prediction

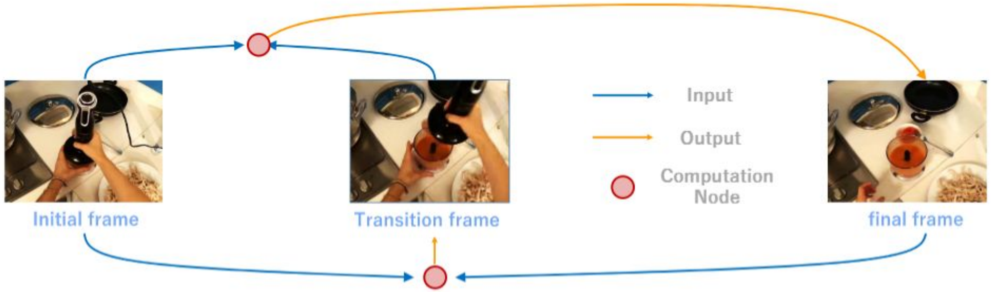
Two Training Objectives

“where to interact”	“how to interact”

Transition / Future states

Visual World Models
w/ Highlighted Interaction

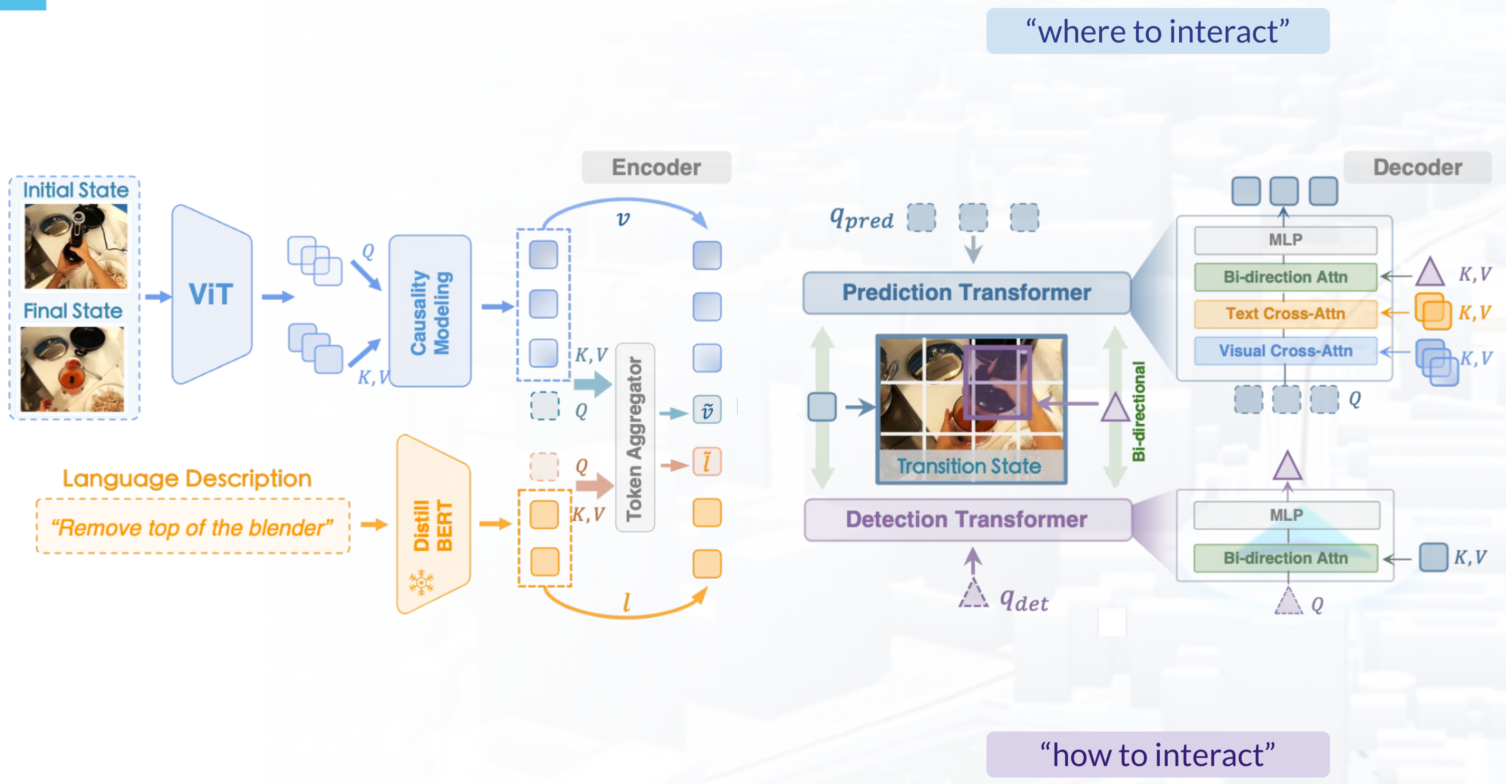
Ego4D
Hand-and-Object subset



State-change: Plant removed from ground

State-change: Wood smoothed

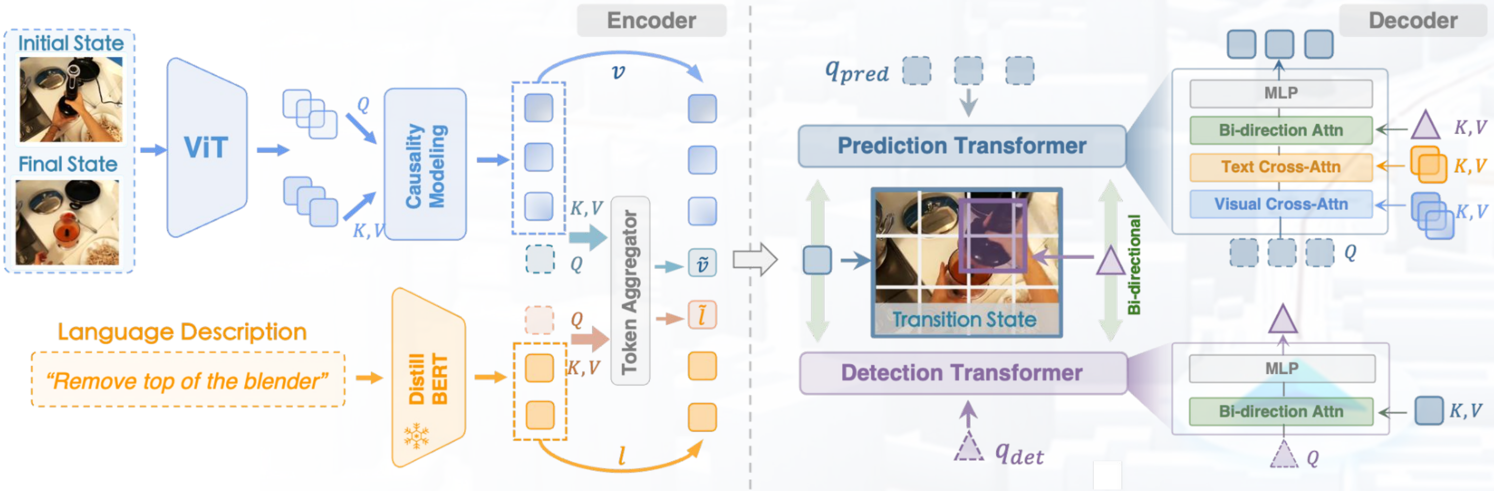
MPI | Model



“where to interact”
Initial State
Final State
ViT
Q
K,V
Causality Modeling
Encoder
v
K,V
Q
Token Aggregator
$\tilde{v}$
$\tilde{l}$
Q
K,V
l
Language Description
“Remove top of the blender”
Distill BERT
q_{pred}
Prediction Transformer
Transition State
Bi-directional
Detection Transformer
q_{det}
Decoder
MLP
Bi-direction Attn
Text Cross-Attn
Visual Cross-Attn
K,V
K,V
K,V
Q
MLP
Bi-direction Attn
K,V
Q
“how to interact”

MPI | Model

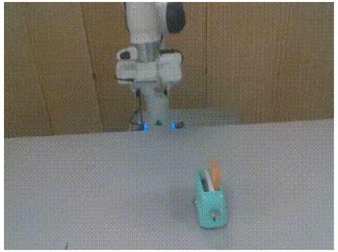
"where to interact"

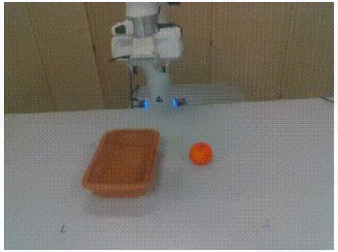
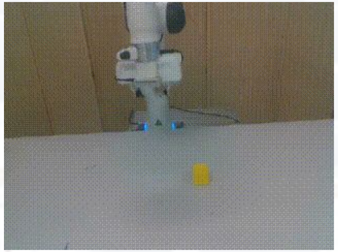


"how to interact"

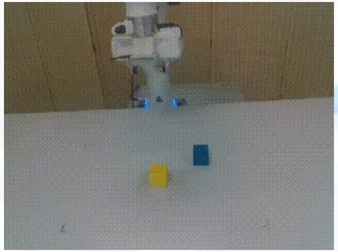
MPI | Results

Demos in clean background with varied positions/angles/etc

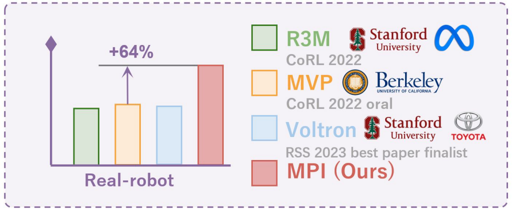



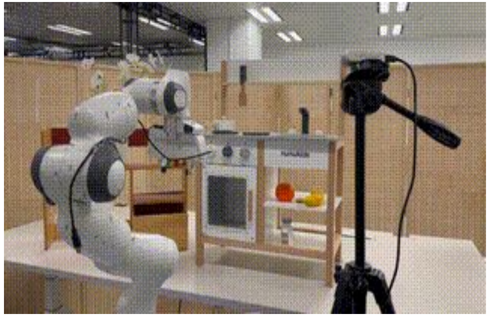


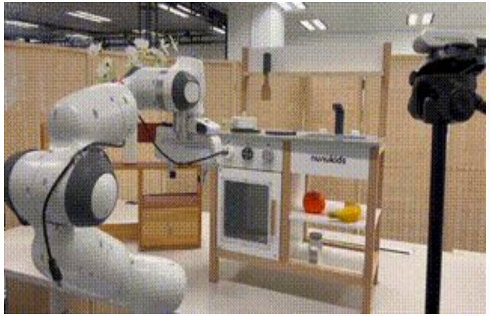


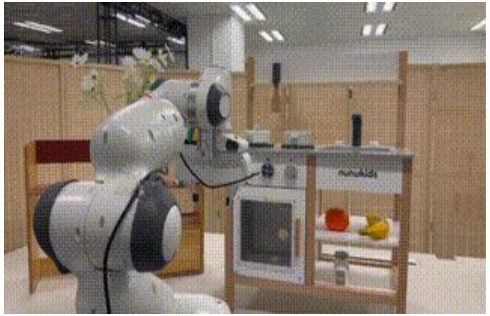

MPI | Results



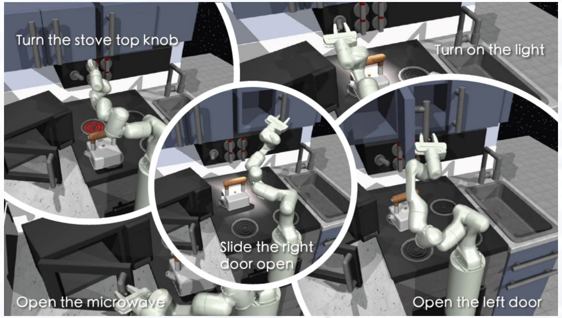
Real-robot Experiments



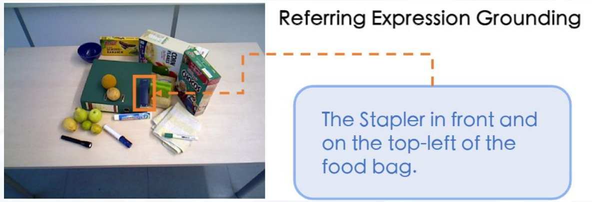




Visuomotor Control in Simulation

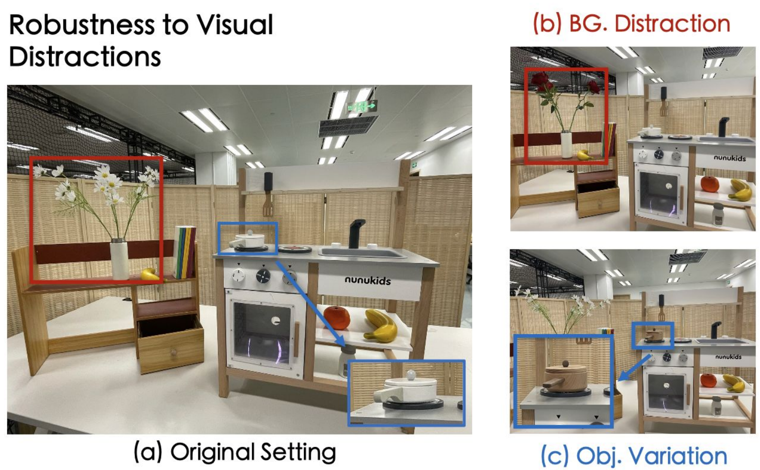


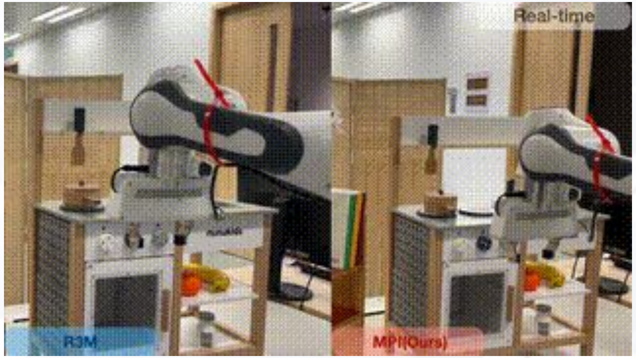
Referring Expression Grounding



MPI | Generalization Results

Generalization Validation





Object Variation

White plastic pot → Wooden pot

Background Distraction

Daisies → Roses

Conclusion

- **Data: Visual data**, like **large-scale** videos and image-LiDAR pairs, are valuable to train a **generalized** world model by **self-supervised learning**.
- **Model:** World models have **different forms**, like videos and BEVs, and **different conditions**; all serving as effective environmental abstractions.
- **Application**: Learning representations by learning world models are helpful for multiple applications, including **policy learning**, reward evaluation, and diverse driving tasks.

Visual World Models as Foundation Models for Autonomous Agents

? **Visual World Models with LLM/VLMs as Foundation Models for Autonomous Agents**

Thank you