OpenDriveLab

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

How to scale up the autonomous driving models?

GenAD: Generalized Predictive Model for Autonomous Driving

CVPR 2024, Highlight

arxiv.2403.09630

Motivation (1/3) | What Makes for Generalized AD Model?

Data Distinction:

+ LLMs pretrained on **trillions of unlabeled text tokens** exhibit strong generalization in a variety of domains and applications

- However, existing AD models are established on **limited labeled data**, which hampers their generalization

LLMs

Unlabeled Text Data

arXiv Common Crawl

✅ **Internet scale**: World knowledge.

✅ **Free of labeling**: Easy to collect and scale up.

✅ ***Strong generalization***

Existing AD Models

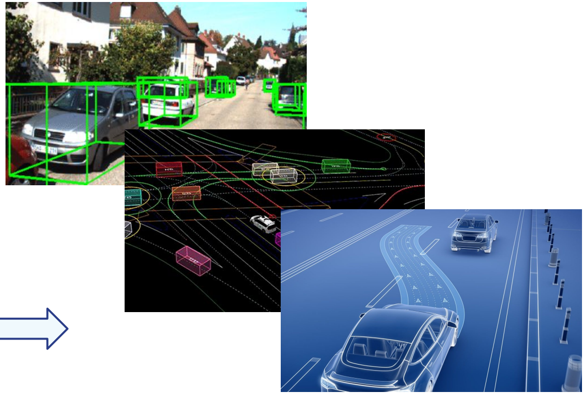
Labeled Driving Data

ARGO AI nuPlan

❌ **Small scale**: Limited domain knowledge.

❌ **Intricate labeling process**: Unscalable.

❌ ***Poor generalization***



- **Bbox, HD map, etc.**

Yang et al., GenAD: Generalized Predictive Model for Autonomous Driving, CVPR 2024

Motivation (2/3) | What Makes for Generalized AD Model?

Learning Objective:

- Supervised by **3D labels**
  ❌ Hard to scale without sufficient labeled data

No accessible labeled data

Model

NUSCENES

Model-XL

- Supervised by **expert features**
  ✅ Scalable with developed expert models (e.g., DINOv2)
  ✅ Focusing on specific objects (e.g., centered or large ones)
  ❌ Ignoring critical details (e.g., small objects)

- Feature map visualization from DINOv2

❌ ***Undesirable** for modeling challenging driving scenes*

Yang et al., GenAD: Generalized Predictive Model for Autonomous Driving, CVPR 2024

Motivation (3/3) | What Makes for Generalized AD Model?

Our Initiative:

Data: Massive online driving videos

Learning Objective:

- Supervised by “**pixels of future frames**” → **Video Prediction**

✅ **Scalable Data** (easy to collect from the web)
✅ No 3D labeling needed
✅ Better detail preservation
✅ Learning **world knowledge** and **how to drive** inherently

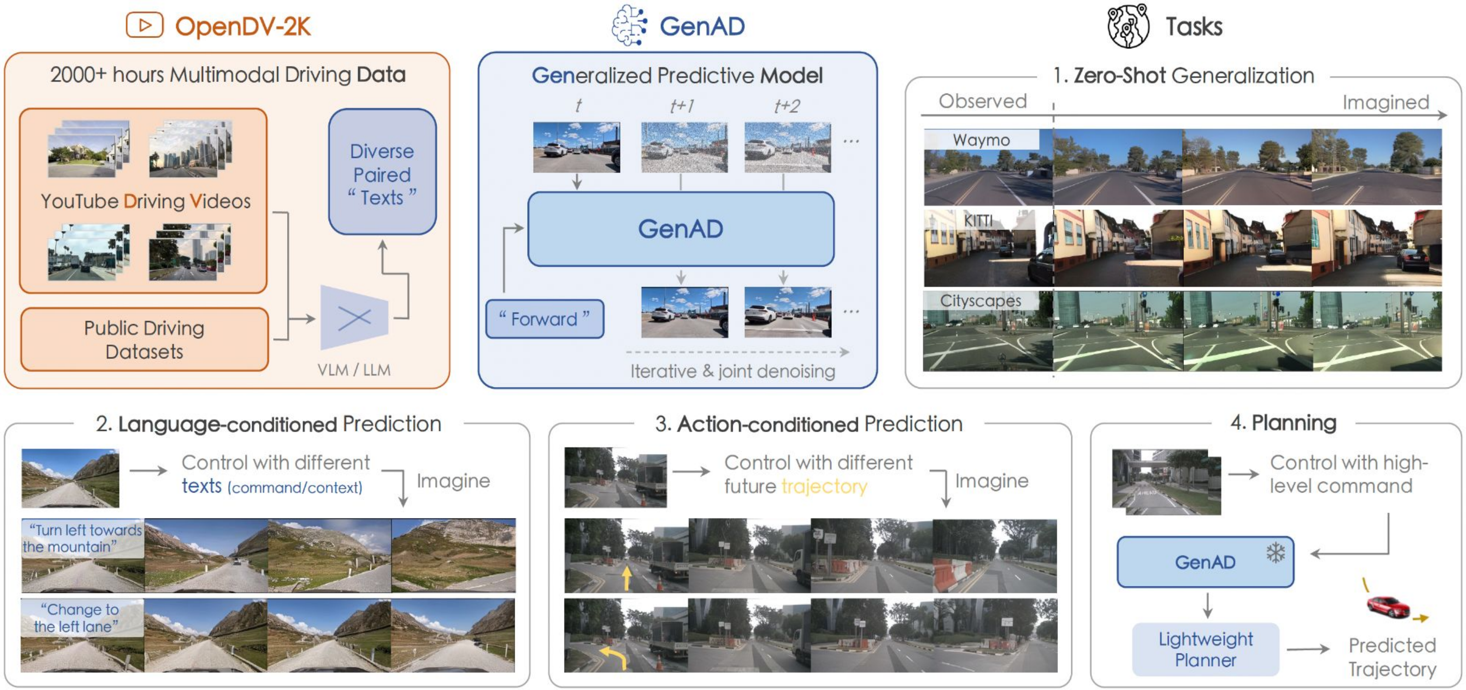
✅ ***Strong** generalization*

Massive YouTube videos, collected worldwide

Yang et al., GenAD: Generalized Predictive Model for Autonomous Driving, CVPR 2024

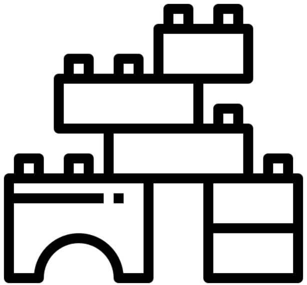
GenAD | At a Glance

Summary: A **billion-scale video prediction model** trained on **web-scale driving videos**, demonstrating **strong generalization across** a wide spectrum of **domains and tasks**.

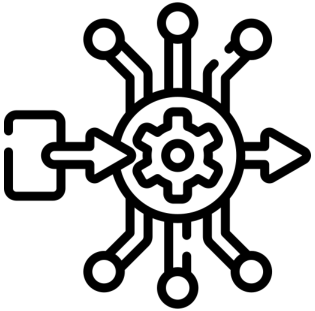


Yang et al., GenAD: Generalized Predictive Model for Autonomous Driving, CVPR 2024

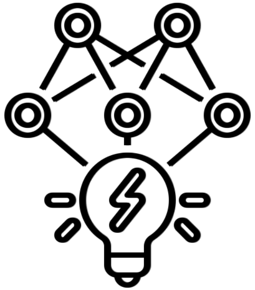
GenAD - Overview



Data



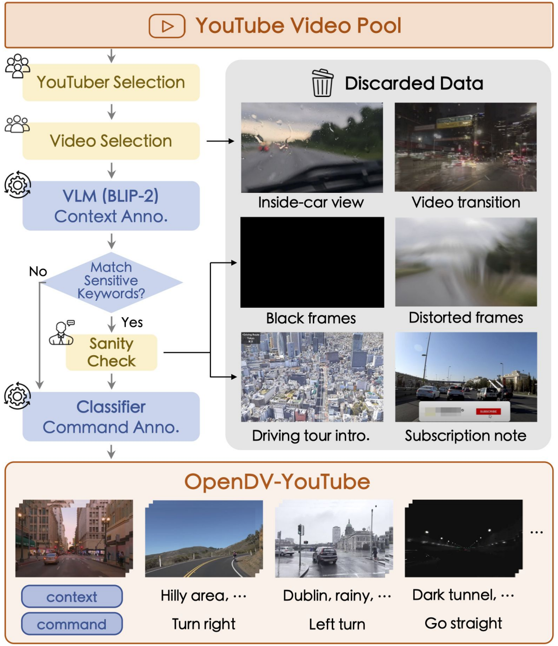
Model



Tasks

Yang et al., GenAD: Generalized Predictive Model for Autonomous Driving, CVPR 2024

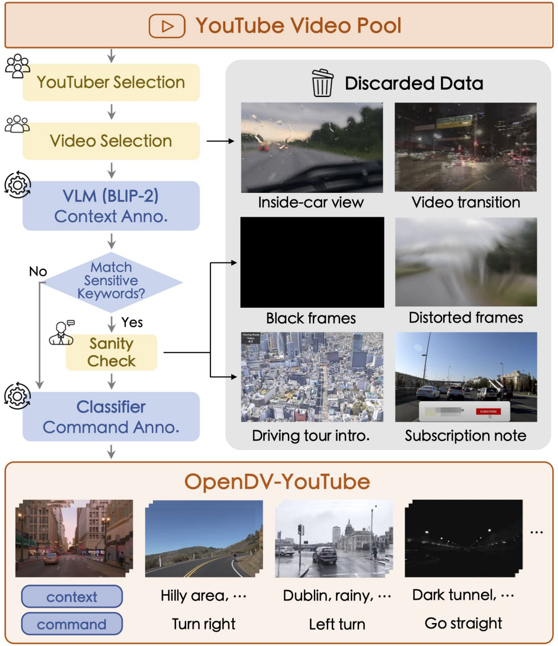
GenAD | Dataset

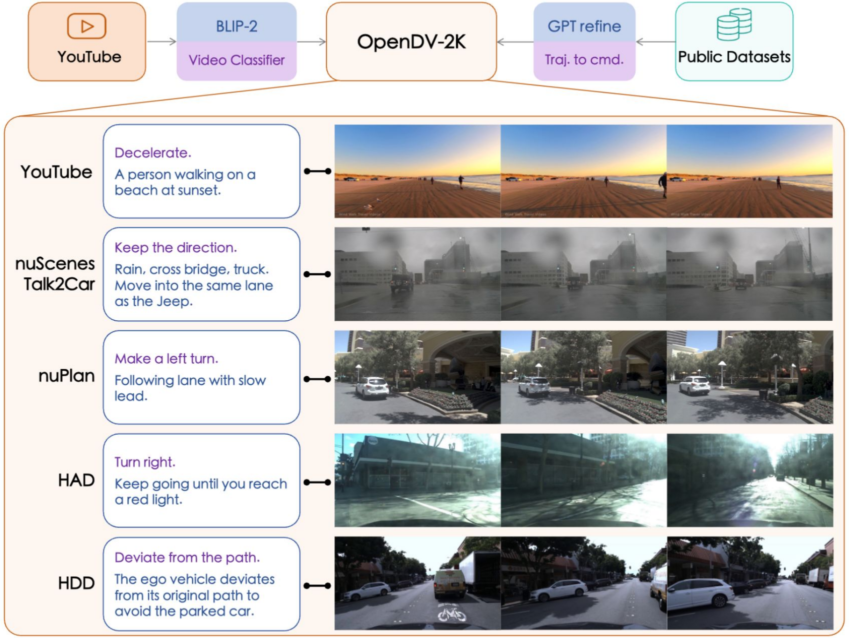


- Rigorous **data collection** and **filtering** strategy

Yang et al., GenAD: Generalized Predictive Model for Autonomous Driving, CVPR 2024

GenAD | Dataset





- Rigorous **data collection** and **filtering** strategy

- **Multi-modal** and **Multi-source Nature**
  - Sourced from both **online videos** and **public datasets** for diversity
  - Paired with textual **context** and **command**

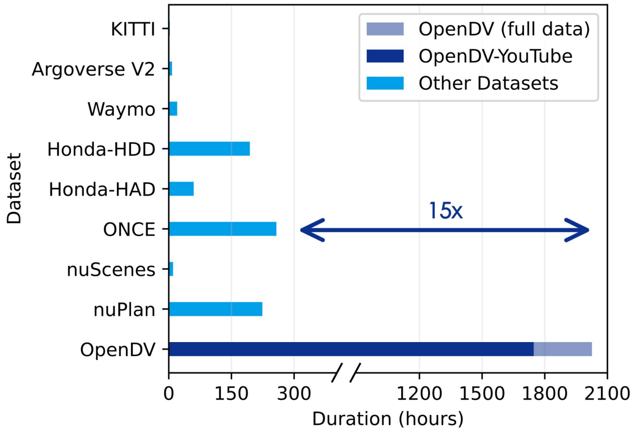
Yang et al., GenAD: Generalized Predictive Model for Autonomous Driving, CVPR 2024

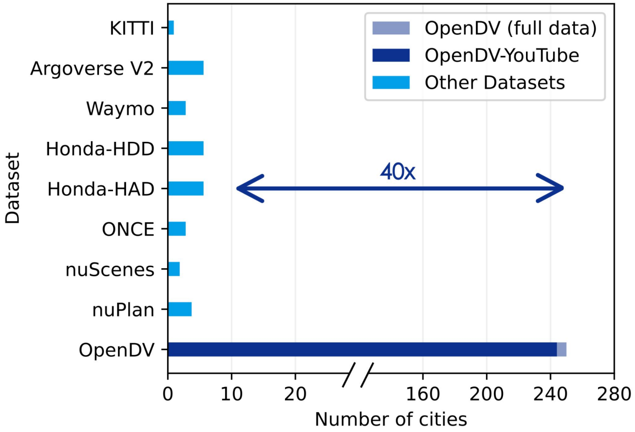
GenAD | Dataset

- ***Largest public dataset*** for autonomous driving
- ≥ **2059 hours**, ≥ **244 cities**

	Dataset	Duration (hours)	Front-view Frames	Geographic Diversity Countries	Cities	Sensor Setup
✗	KITTI [30]	1.4	15k	1	1	fixed
✗	Cityscapes [21]	0.5	25k	3	50	fixed
✗	Waymo Open* [97]	11	390k	1	3	fixed
✗	Argoverse 2* [109]	4.2	300k	1	6	fixed
✓	nuScenes [12]	5.5	241k	2	2	fixed
✓	nuPlan* [13]	120	4.0M	2	4	fixed
✓	Talk2Car [24]	4.7	-	2	2	fixed
✓	ONCE [72]	144	7M	1	-	fixed
✓	Honda-HAD [51]	32	1.2M	1	-	fixed
✓	Honda-HDD-Action [84]	104	1.1M	1	-	fixed
✓	Honda-HDD-Cause [84]	32	-	1	-	fixed
✓	OpenDV-YouTube (Ours)	1747	60.2M	$\geq 40^{\dagger}$	$\geq 244^{\dagger}$	uncalibrated
-	**OpenDV-2K** (Ours)	**2059**	**65.1M**	$\geq \mathbf{40}^{\dagger}$	$\geq \mathbf{244}^{\dagger}$	**uncalibrated**

OpenDV-2K (Ours) 🚀

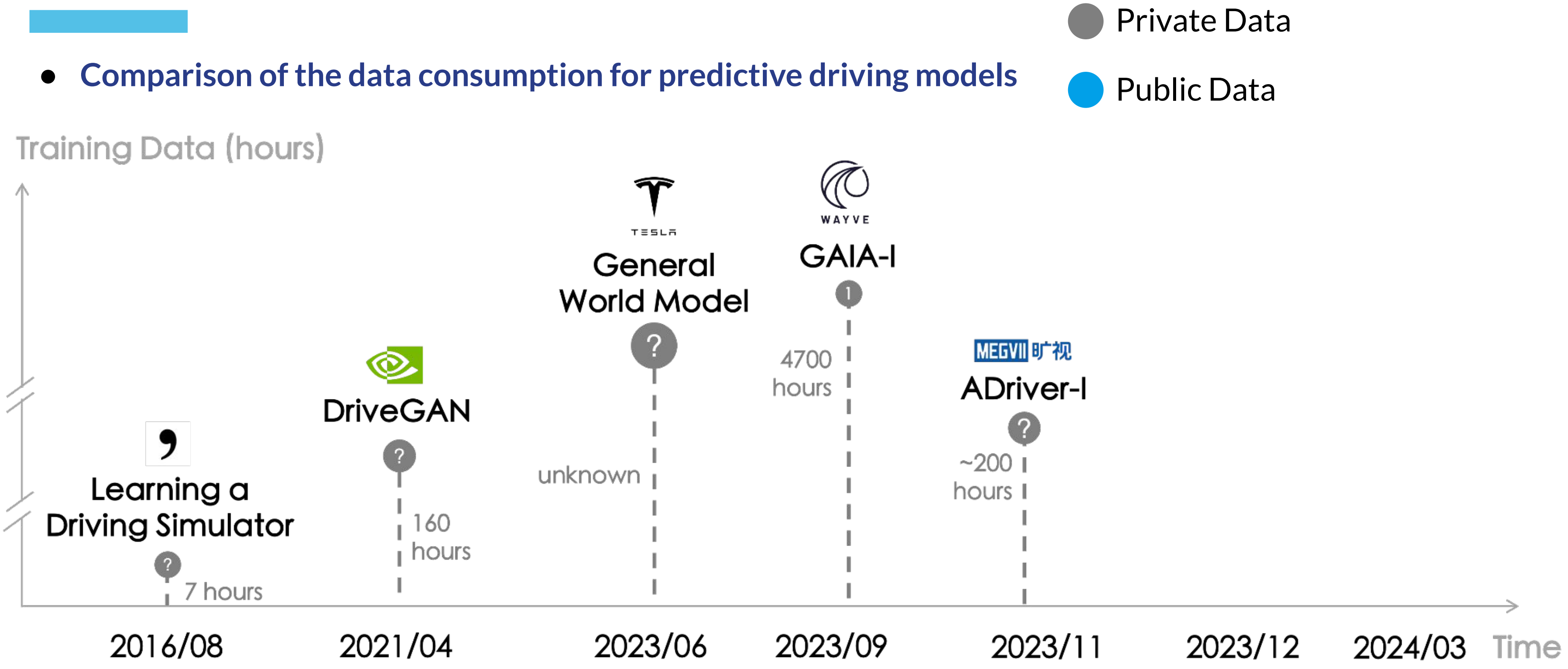




Yang et al., GenAD: Generalized Predictive Model for Autonomous Driving, CVPR 2024

GenAD | Dataset

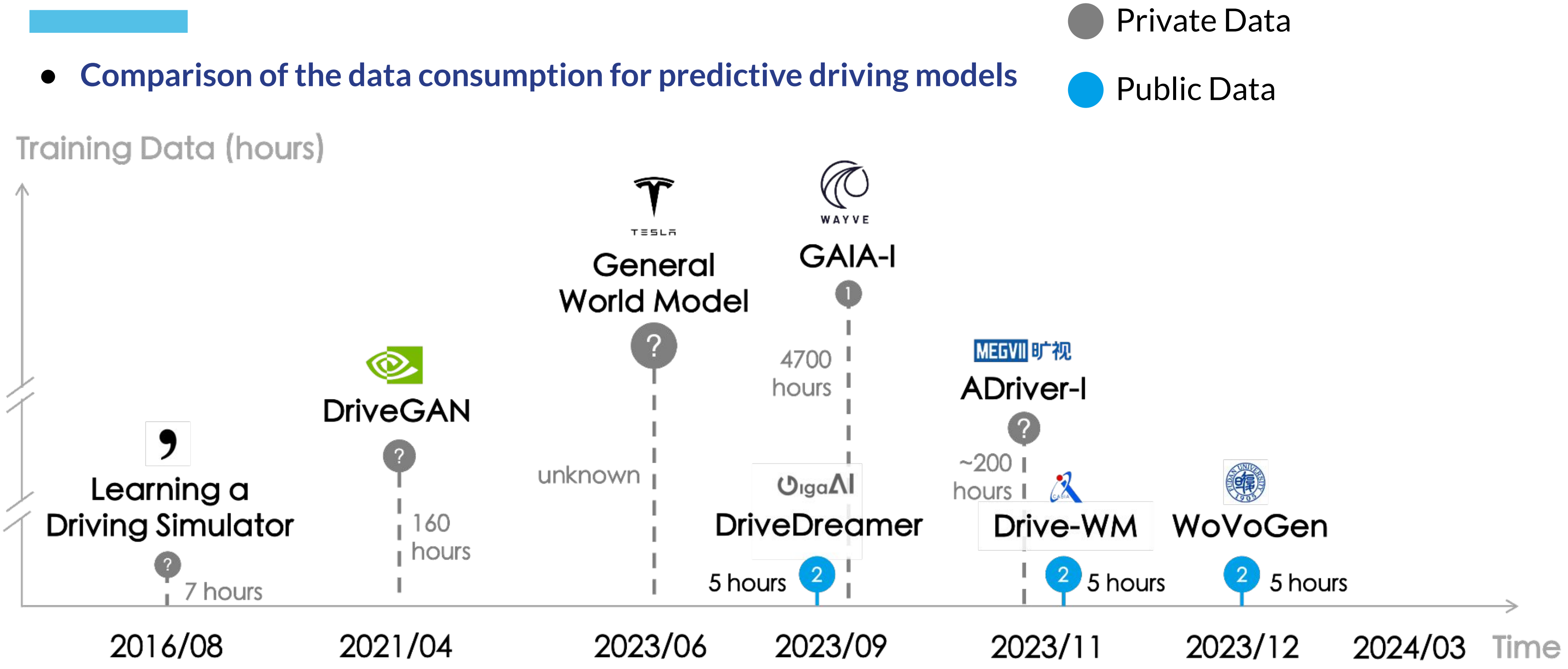
- **Comparison of the data consumption for predictive driving models**



Yang et al., GenAD: Generalized Predictive Model for Autonomous Driving, CVPR 2024

GenAD | Dataset

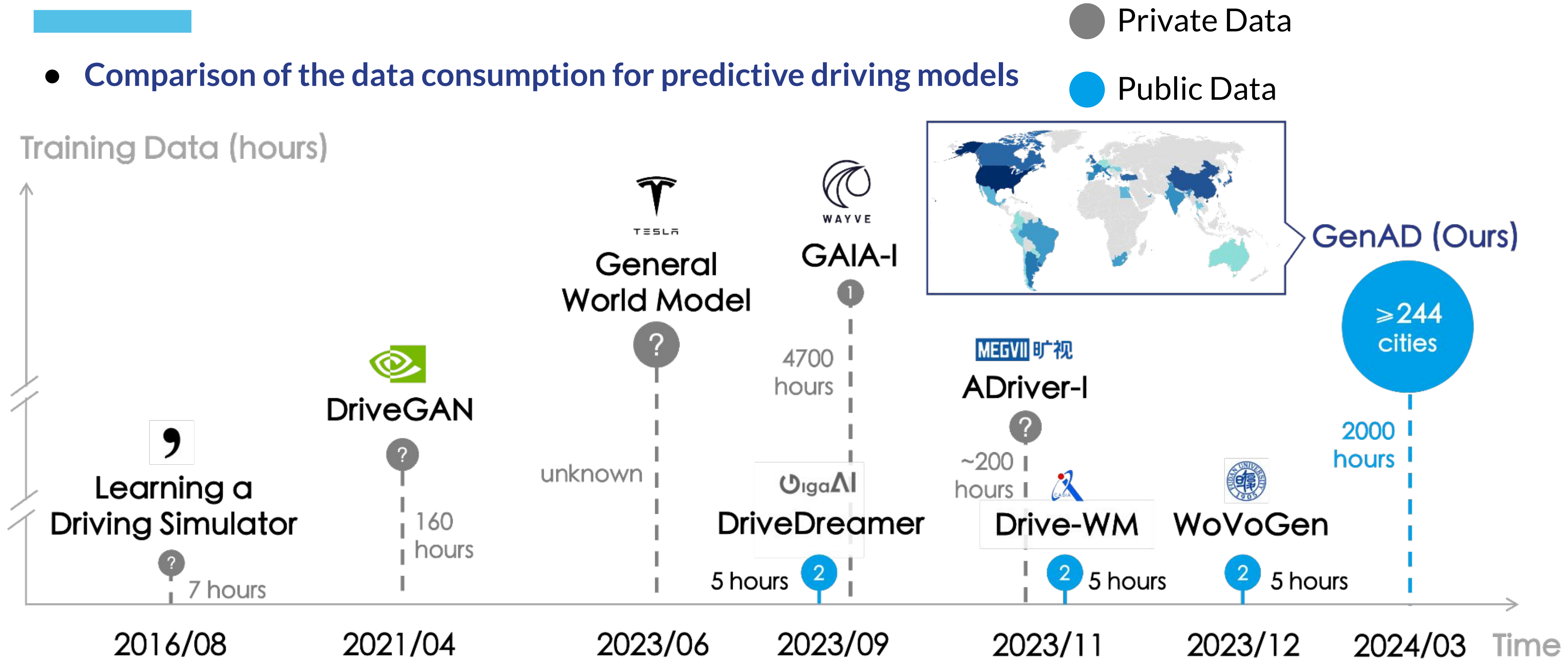
- **Comparison of the data consumption for predictive driving models**



Yang et al., GenAD: Generalized Predictive Model for Autonomous Driving, CVPR 2024

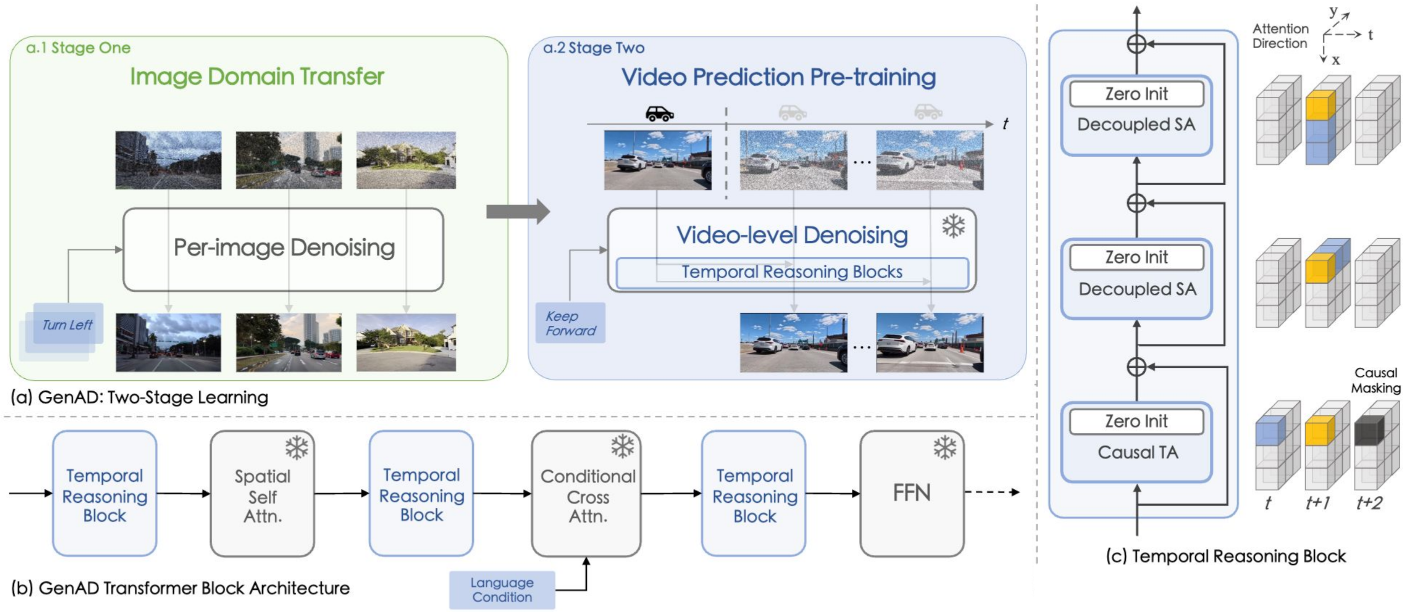
GenAD | Dataset

- **Comparison of the data consumption for predictive driving models**



Yang et al., GenAD: Generalized Predictive Model for Autonomous Driving, CVPR 2024

Algorithm | Video Prediction Model for Driving



- **Two-stage Training**:
  - Tuning the **image generation model** (SDXL) into a highly-capable **video prediction model**
- **Model Specializations for Driving:**
  - Causal Temporal Attention: coherent and consistent future prediction
  - Decoupled Spatial Attention: efficient long-range modeling
  - Interleaved temporal blocks: sufficient spatiotemporal interaction

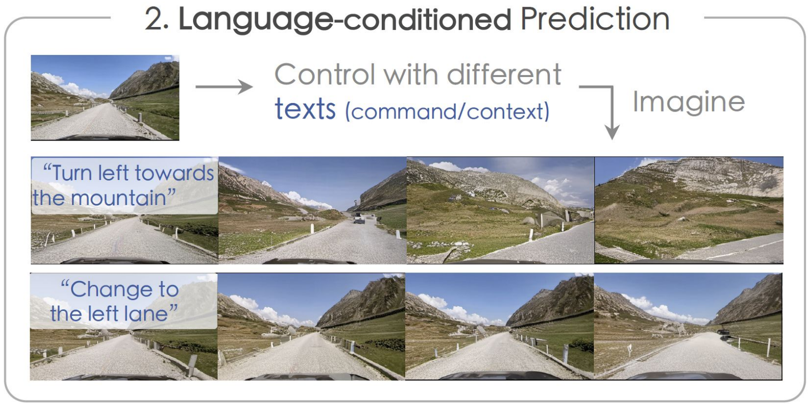
Yang et al., GenAD: Generalized Predictive Model for Autonomous Driving, CVPR 2024

Result on Tasks (1/4) | Zero-shot Generalization (Video Prediction)

- **Zero-shot video prediction** on unseen datasets including Waymo, KITTI and Cityscapes
- Outperforming competitive general video generation models

Yang et al., GenAD: Generalized Predictive Model for Autonomous Driving, CVPR 2024

Result on Tasks (2/4) | Language-conditioned Prediction



Controlling the future evolvement with **language**

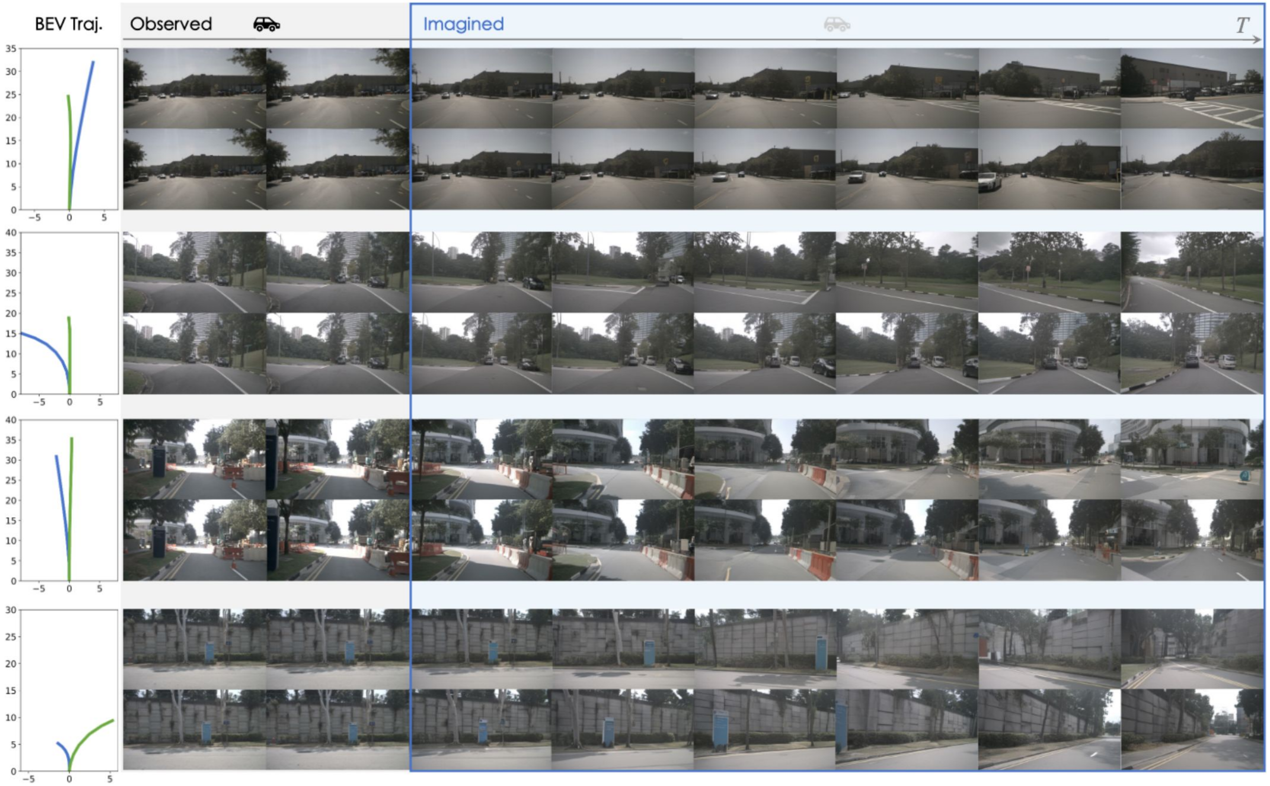
Yang et al., GenAD: Generalized Predictive Model for Autonomous Driving, CVPR 2024

Result on Tasks (3/4) | Action-conditioned Prediction (Simulation)

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

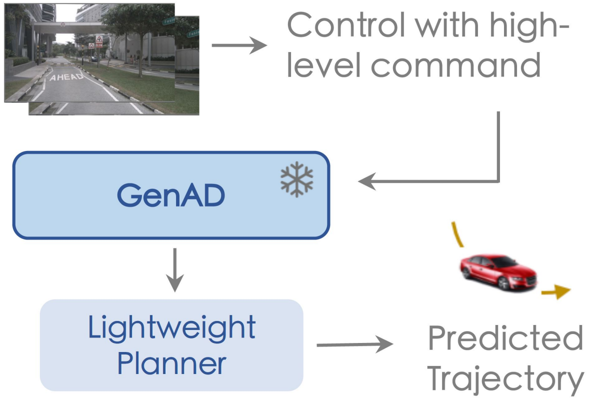
Table 4. **Task on Action-conditioned prediction**. Compared to GenAD with text conditions only, GenAD-act enables more precise future predictions that follow the action condition.

Simulating the future with **user-specified trajectory**



Yang et al., GenAD: Generalized Predictive Model for Autonomous Driving, CVPR 2024

Result on Tasks (4/4) | Planning



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

Table 5. **Task on Planning**. A lightweight MLP with *frozen* GenAD gets competitive planning results with 73× fewer trainable parameters and front-view image alone. *: multi-view inputs.

- Speeding up training by **3400 times** *(vs. **UniAD**)*
- Demonstrating the **effectiveness of** the learned spatiotemporal **representations**

Yang et al., GenAD: Generalized Predictive Model for Autonomous Driving, CVPR 2024

Summary

- **Largest Public Driving Dataset:**
  - **OpenDV-2K** provides ***2059 hours*** of ***worldwide*** driving videos.
- **Generalized Predictive Model for Autonomous Driving:**
  - **GenAD** can predict plausible futures with ***language*** conditions and generalize to ***unseen*** datasets in a ***zero-shot*** manner.
- **Broad Applications:**
  - **GenAD** can readily adapt to ***planning*** and ***simulation***.

Yang et al., GenAD: Generalized Predictive Model for Autonomous Driving, CVPR 2024

OpenDriveLab

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

(Follow-up work)

How to build a generally applicable driving world model?

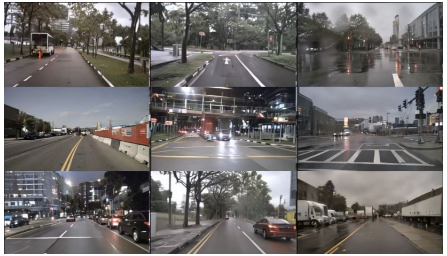
Vista: A Generalizable Driving World Model with High Fidelity and Versatile Controllability

arxiv.2405.17398

Limitations of Existing Driving World Models

- **Generalization:** limited data scale and geographical coverage

5h
within Singapore & Boston
nuScenes



- **Representation capacity:** low resolution and low frame rate

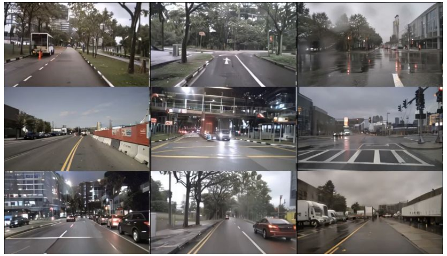
- **Control flexibility:** single modality, incompatible with planning algorithms

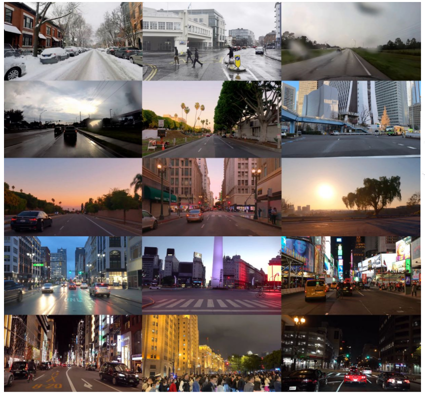
Gao et al., Vista: A Generalizable Driving World Model with High Fidelity and Versatile Controllability

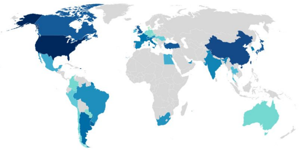
Our Investigation: A Generalizable Driving World Model

- **Generalization:** largest driving video dataset

5h
within Singapore & Boston
nuScenes







1740h
worldwide

- **Representation capacity:** high spatiotemporal resolution

576×1024

- **Control flexibility:** multi-modal action inputs

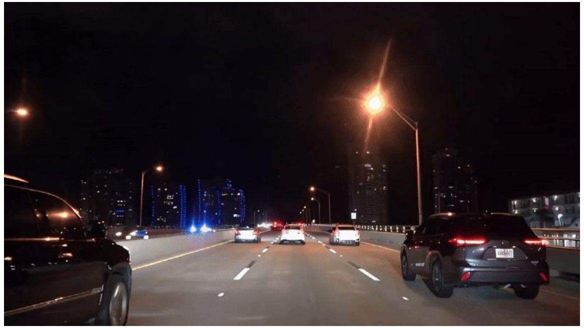
Gao et al., Vista: A Generalizable Driving World Model with High Fidelity and Versatile Controllability

Capability of Vista

- **High-fidelity future prediction**



- **Continuous long-horizon rollout (15 seconds)**


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

Capability of Vista

- **Zero-shot action controllability**

turn left | go straight | turn right | stop

- **Provide reward without ground truth actions**

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

Summary

- **Vista is a generalizable driving world model that can:**
  - *Predict high-fidelity futures in open-world scenarios.*
  - *Extend its predictions to continuous and long horizons.*
  - *Execute multi-modal actions (steering angles, speeds, commands, trajectories, goal points).*
  - *Provide rewards for different actions without accessing ground truth actions.*

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

OpenDriveLab

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

Thanks

https://opendrivelab.com/