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)
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

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

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

Vision 2020
Position Paper (by LeCun) Positioning the developments of world models

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

Driving Robotics 2022
Scaling up world models on large corpus of videos

General World Model: inhouse data collected around the globe

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

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

2023

2024

2018 2020 2022 2023 2024
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 Wang¹, Graham Todd², Ziang Xiao⁎, Xingdi Yuan⁰
Marc-Alexandre Côté⁰, Peter Clark⁎, Peter Jansen¹
¹University of Arizona  ⁰Microsoft Research Montréal
²New York University  ⁿJohns 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
- Effective generalization
- Diverse range of use cases
- Self-supervision (generally)

Raw data → Labeled data
World knowledge → Task-wise optimization
Self-supervised learning → Supervised learning

Representation Learning × 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?
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

- dataset: https://github.com/OpenDriveLab/DriveAGI
A large-scale video prediction model on web-scale driving videos, to enable its generalization across a wide spectrum of domains and tasks.

- dataset: https://github.com/OpenDriveLab/DriveAGI

Highlight
Thu. 20 Jun 5 p.m – 6:30 p.m
Arch 4A-E Poster #5
Data | Scale-up Driving Videos

Training Data (hours)

- Bubble size: Number of cities covered
- Dash line Length: Duration of the training dataset

<table>
<thead>
<tr>
<th>Data Source</th>
<th>Time</th>
<th>Hours</th>
<th>Cities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tesla General World Model</td>
<td>2016/08</td>
<td>7 hours</td>
<td>unknown</td>
</tr>
<tr>
<td>DriveGAN</td>
<td>2021/04</td>
<td>160 hours</td>
<td>unknown</td>
</tr>
<tr>
<td>DriveDreamer</td>
<td>2023/06</td>
<td>5 hours</td>
<td>unknown</td>
</tr>
<tr>
<td>ADriver-I</td>
<td>2023/09</td>
<td>4700 hours</td>
<td>?</td>
</tr>
<tr>
<td>GAIA-I</td>
<td>2023/09</td>
<td>~200 hours</td>
<td>?</td>
</tr>
<tr>
<td>Drive-WM</td>
<td>2023/11</td>
<td>5 hours</td>
<td>2</td>
</tr>
<tr>
<td>WoVoGen</td>
<td>2023/12</td>
<td>5 hours</td>
<td>2</td>
</tr>
<tr>
<td>Drive-WM</td>
<td>2023/12</td>
<td>5 hours</td>
<td>2</td>
</tr>
<tr>
<td>WoVoGen</td>
<td>2024/03</td>
<td>5 hours</td>
<td>2</td>
</tr>
</tbody>
</table>

OpenDV (Ours) | ≥709 cities |

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.)

(a) Global Distribution

(b) in USA

(c) in China
Data | OpenDV

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

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Duration (hours)</th>
<th>Front-view Frames</th>
<th>Geographic Diversity</th>
<th>Sensor Setup</th>
</tr>
</thead>
<tbody>
<tr>
<td>KITTI [14]</td>
<td>1.4</td>
<td>15k</td>
<td>1</td>
<td>fixed</td>
</tr>
<tr>
<td>Cityscapes [10]</td>
<td>0.5</td>
<td>25k</td>
<td>3</td>
<td>fixed</td>
</tr>
<tr>
<td>Waymo Open* [41]</td>
<td>11</td>
<td>390k</td>
<td>1</td>
<td>fixed</td>
</tr>
<tr>
<td>Argoverse 2* [45]</td>
<td>4.2</td>
<td>300k</td>
<td>1</td>
<td>fixed</td>
</tr>
<tr>
<td>nuScenes [6]</td>
<td>5.5</td>
<td>241k</td>
<td>2</td>
<td>fixed</td>
</tr>
<tr>
<td>nuPlan [7]</td>
<td>120</td>
<td>4.0M</td>
<td>2</td>
<td>fixed</td>
</tr>
<tr>
<td>Talk2Car [12]</td>
<td>4.7</td>
<td>-</td>
<td>2</td>
<td>fixed</td>
</tr>
<tr>
<td>ONCE [32]</td>
<td>144</td>
<td>7M</td>
<td>1</td>
<td>fixed</td>
</tr>
<tr>
<td>Honda-HAD [23]</td>
<td>32</td>
<td>1.2M</td>
<td>1</td>
<td>fixed</td>
</tr>
<tr>
<td>Honda-HDD-Action [38]</td>
<td>104</td>
<td>1.1M</td>
<td>1</td>
<td>fixed</td>
</tr>
<tr>
<td>Honda-HDD-Cause [38]</td>
<td>32</td>
<td>-</td>
<td>1</td>
<td>fixed</td>
</tr>
<tr>
<td>OpenDV-YouTube (Ours)</td>
<td>1747</td>
<td>60.2M</td>
<td>≥40†</td>
<td>uncalibrated</td>
</tr>
<tr>
<td>OpenDV-2K (Ours)</td>
<td>2059</td>
<td>65.1M</td>
<td>≥40†</td>
<td>uncalibrated</td>
</tr>
</tbody>
</table>

OpenDV-2K (Ours)

- dataset: [https://github.com/OpenDriveLab/DriveAGI](https://github.com/OpenDriveLab/DriveAGI)
Data | OpenDV

- 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.
Tasks | Zero-shot Generalization (Video Prediction)

Zero-shot video prediction on unseen datasets including Waymo, KITTI and Cityscapes
Tasks | Language-conditioned Prediction

2. Language-conditioned Prediction

Control with different texts (command/context) Imagine

“Turn left towards the mountain”
“Change to the left lane”

“Rain, Wait at crossroad”
“Drive slowly down at intersection, several barriers beside the road”
“Turn right, some parked cars, a parking lot”

Instruct the future with free-form texts.
Tasks | Action-conditioned Prediction (Simulation)

<table>
<thead>
<tr>
<th>Method</th>
<th>Condition</th>
<th>nuScenes Action Prediction Error ($\downarrow$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ground truth</td>
<td>-</td>
<td>0.9</td>
</tr>
<tr>
<td>GenAD</td>
<td>text</td>
<td>2.54</td>
</tr>
<tr>
<td>GenAD-act</td>
<td>text + traj.</td>
<td>2.02</td>
</tr>
</tbody>
</table>

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.
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?
Vista: A Generalized Driving World Model with High Fidelity and Versatile Controllability
## Driving World Models

<table>
<thead>
<tr>
<th>Method</th>
<th>Data Scale</th>
<th>Model Setups</th>
<th>Resolution</th>
<th>Action Control Modes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Frame Rate</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DriveSim [99]</td>
<td>7h</td>
<td>5 Hz</td>
<td>80×160</td>
<td>✓</td>
</tr>
<tr>
<td>DriveGAN [66]</td>
<td>160h</td>
<td>8 Hz</td>
<td>256×256</td>
<td>✓</td>
</tr>
<tr>
<td>DriveDreamer [122]</td>
<td>5h</td>
<td>12 Hz</td>
<td>128×192</td>
<td>✓</td>
</tr>
<tr>
<td>Drive-WM [124]</td>
<td>5h</td>
<td>2 Hz</td>
<td>192×384</td>
<td>✓</td>
</tr>
<tr>
<td>WoVoGen [87]</td>
<td>5h</td>
<td>2 Hz</td>
<td>256×448</td>
<td>✓</td>
</tr>
<tr>
<td>ADriver-I [60]</td>
<td>300h</td>
<td>2 Hz</td>
<td>256×512</td>
<td>✓</td>
</tr>
<tr>
<td>GenAD [133]</td>
<td>2000h</td>
<td>2 Hz</td>
<td>256×448</td>
<td>✓ ✓</td>
</tr>
<tr>
<td>GAIA-1 [53]</td>
<td>4700h</td>
<td>25 Hz</td>
<td>288×512</td>
<td>✓</td>
</tr>
<tr>
<td><strong>Vista (Ours)</strong></td>
<td>1740h</td>
<td>10 Hz</td>
<td>576×1024</td>
<td>✓ ✓ ✓ ✓ ✓</td>
</tr>
</tbody>
</table>
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 Conditioning**: A unified concatenation of Fourier embeddings
- **Efficient Learning**: Parameter-efficient LoRA adapters to finetune
- **Collaborative Learning**: Learning with unlabeled YouTube data for better generalization
Vista | Video Prediction

- High-fidelity future prediction

- Continuous long-horizon rollout (15 seconds)
Vista | Zero-shot Action Controllability

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

- Drive-WM rewards
  
  Drive-WM, CVPR’24

- Provide reward without ground truth actions, by uncertainty

![Graph showing the relationship between Average Reward and Average L2 Error](image1)

**Reasonable rewards**

- Reward: 0.872 0.815
- Reward: 0.870 0.849
- Reward: 0.872 0.832
- Reward: 0.888 0.860

**More reasonable than ADE**

- Ground Truth
  
  Action1: L2 Error: 0.94, Reward: 0.88
  
  Action2: L2 Error: 1.36, Reward: 0.90
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

Future Prediction Model

World Model

S2Net — Point cloud future prediction for planning

4D-Occ — Ego Future Trajectory

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

ViDAR | Motivation

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

**ViDAR**

“Visual Lidar”: DNN-based Multi-view Stereo

- Redundant to the appearance and measurement engines
- handling “rear protruding” objects – which hover above the object’s ground plane.

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)

Note:

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

TRI-VIDAR

Installation | Configuration | Datasets | Visualization | Publications | License

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.
Visual Point Cloud Forecasting enables Scalable Autonomous Driving

Jiazhi Yang
Li Chen
Yanan Sun
Hongyang Li

- code: [https://github.com/OpenDriveLab/ViDAR](https://github.com/OpenDriveLab/ViDAR)
ViDAR | At a Glance

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

- code: https://github.com/OpenDriveLab/ViDAR

Highlight
Thu. 20 Jun 5 p.m – 6:30 p.m
Arch 4A-E Poster #6
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

Visual Inputs:
-1s, 0s, 0.5s

LiDAR Outputs

Go straight

Turn left
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**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>UniAD</td>
<td>49.36 37.96</td>
<td>38.3 1.32</td>
<td>1054</td>
<td>31.3 69.1</td>
<td>0.75 1.08 0.158</td>
<td>62.8 40.1 54.6 33.9</td>
</tr>
<tr>
<td>ViDAR</td>
<td>52.57 42.33</td>
<td>42.0 1.25</td>
<td>991</td>
<td>33.2 71.4</td>
<td>0.67 0.99 0.149</td>
<td>65.4 42.1 57.3 36.4</td>
</tr>
</tbody>
</table>
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)

**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

**References**
- project page: [https://opendrivelab.com/MPI](https://opendrivelab.com/MPI)
- code: [https://github.com/OpenDriveLab/MPI](https://github.com/OpenDriveLab/MPI)
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"

"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

The Stapler in front and on the top-left of the food bag.
MPI | Generalization Results

**Generalization Validation**

- **Robustness to Visual Distractions**
  - (a) Original Setting
  - (b) BG. Distraction
  - (c) Obj. Variation

- **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