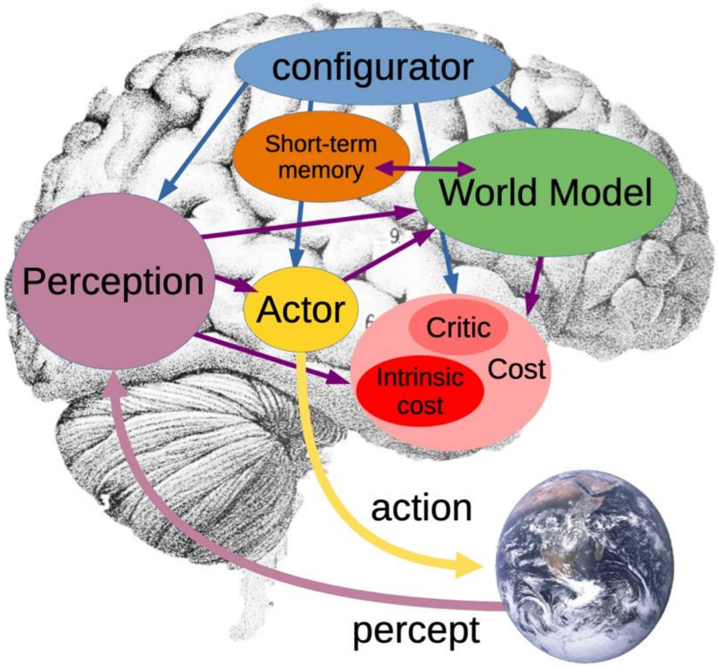
Building Generalizable World Models for Autonomous Driving

Shenyuan Gao

LeCun's AGI System

- Perception
  - Observe and estimate the world state
- Actor
  - Propose actions
- World model (core)
  - Predict plausible future state
- Cost
  - Evaluate the state

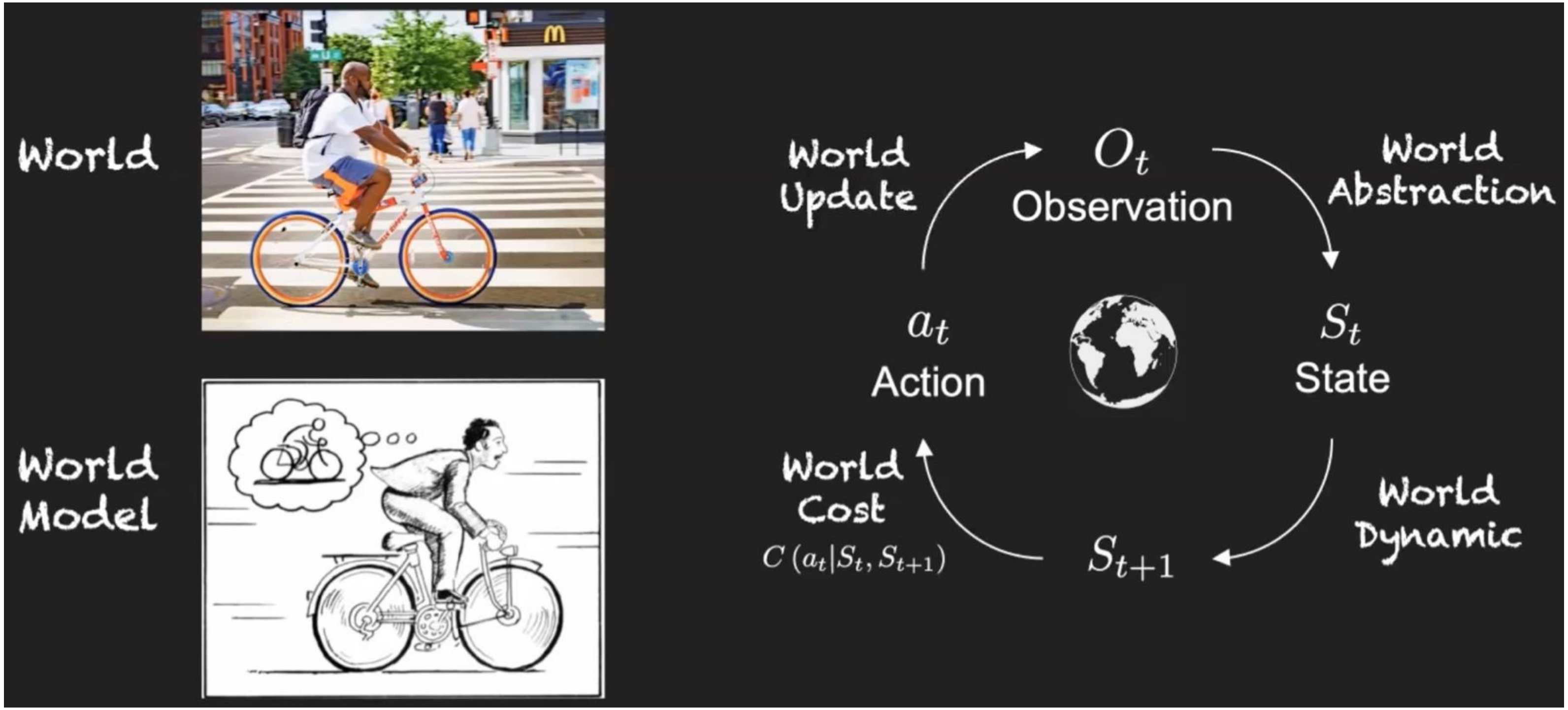


LeCun, Yann. "A Path Towards Autonomous Machine Intelligence version 0.9. 2, 2022-06-27." Open Review 62.1 (2022): 1-62.

What is "World Model"

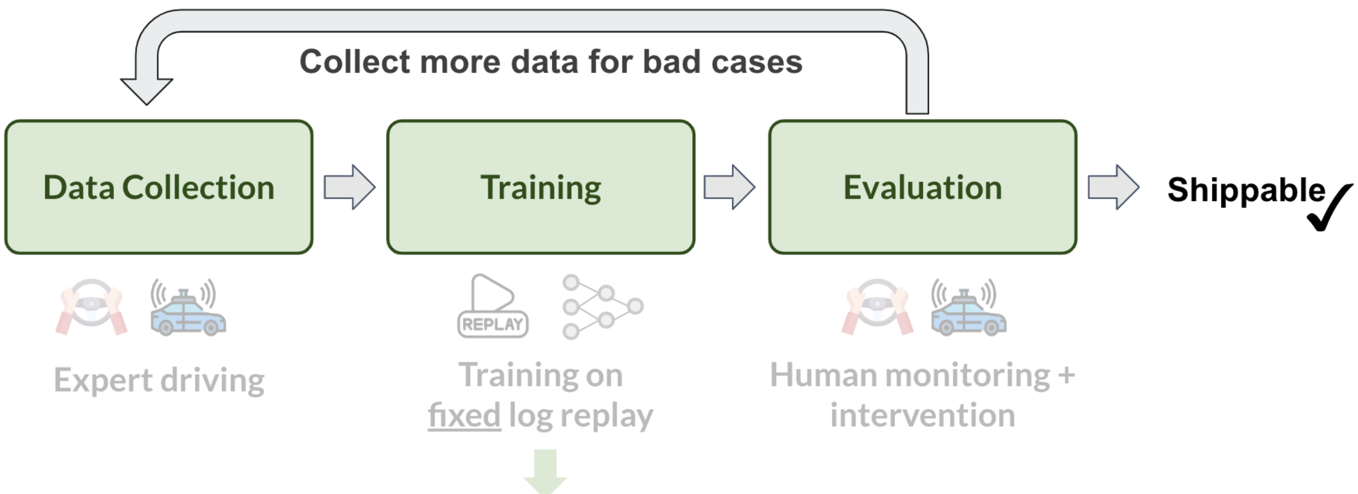
- A substitute that simulates the real world
- Given previous states and actions, predict the future

$$p(s_{t+1}|s_t, a_t)$$



Introducing World Models to Autonomous Driving

- Expectation



\+ Reduced cost

\+ <u>Interactive</u>

\+ Scalable evaluation

Limitations of Existing Driving World Models

Generalization

- Limited data scale and geographic coverage

5 hours only Singapore & Boston

Representation Capacity

- Inferior fidelity
- Low resolution and frame rate

Action Control

- Single action modality
- No zero-shot action controllability

Application

- Underexplored

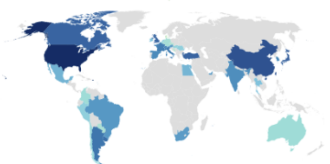
?

- Not applicable for diverse real-world scenarios

Our Goal

Generalization
- Limited data scale and geographic coverage

5 hours only Singapore & Boston

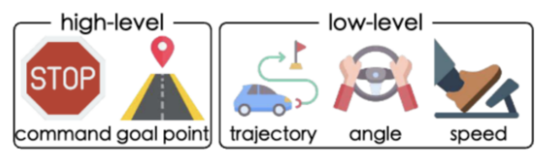


Worldwide Generalization

Representation Capacity
- Inferior fidelity
- Low resolution and frame rate

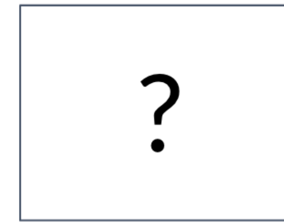
High Fidelity High Resolution

Action Control
- Single action modality
- No zero-shot action controllability



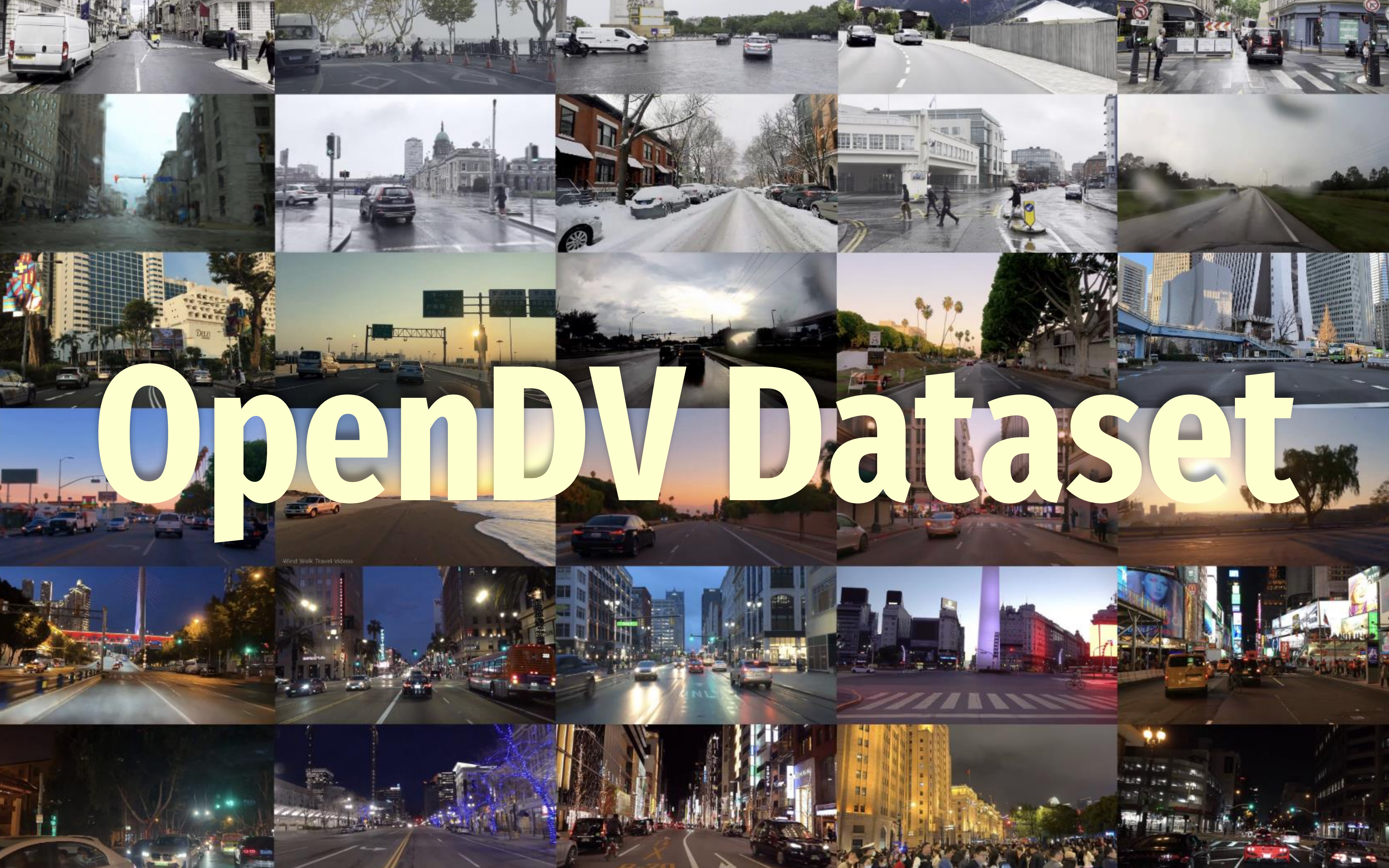
Versatile Action Controllability

Application
- Underexplored



Generalizable Reward Function

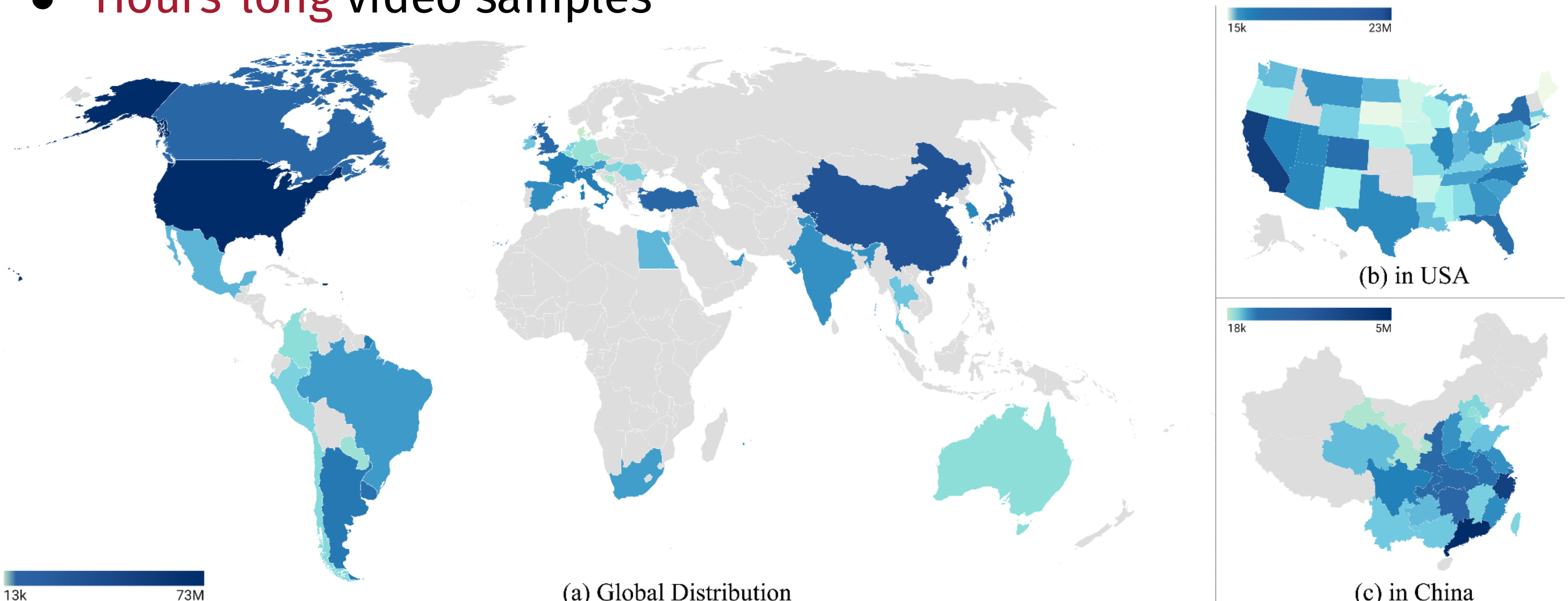
- How to build such a driving world model?


OpenDV Dataset
Wind Walk Travel Videos

Largest Public Driving Video Dataset

- Sourced from YouTube
- 370x larger than nuScenes
- 2059 hours, 40+ nations, 244+ cities, 65M+ frames
- Hours-long video samples

https://github.com/OpenDriveLab/DriveAGI



(a) Global Distribution (b) in USA (c) in China

Yang, Jiazhi, et al. "Generalized Predictive Model for Autonomous Driving." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

From Generation Model to Prediction Model

- <u>Base model</u>: SVD (image2video generation, non-predictive)

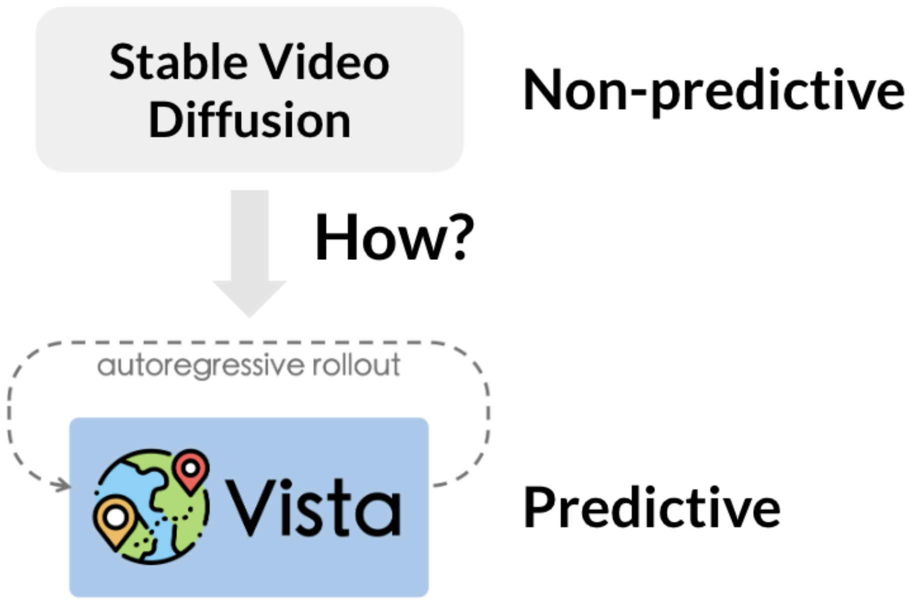
- <u>Our objective</u>

Coherent Long-Horizon Rollout

Blattmann, Andreas, et al. "Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets." arXiv preprint arXiv:2311.15127 (2023).

Gao, Shenyuan, et al. "Vista: A Generalizable Driving World Model with High Fidelity and Versatile Controllability." Advances in Neural Information Processing Systems 38 (2024).

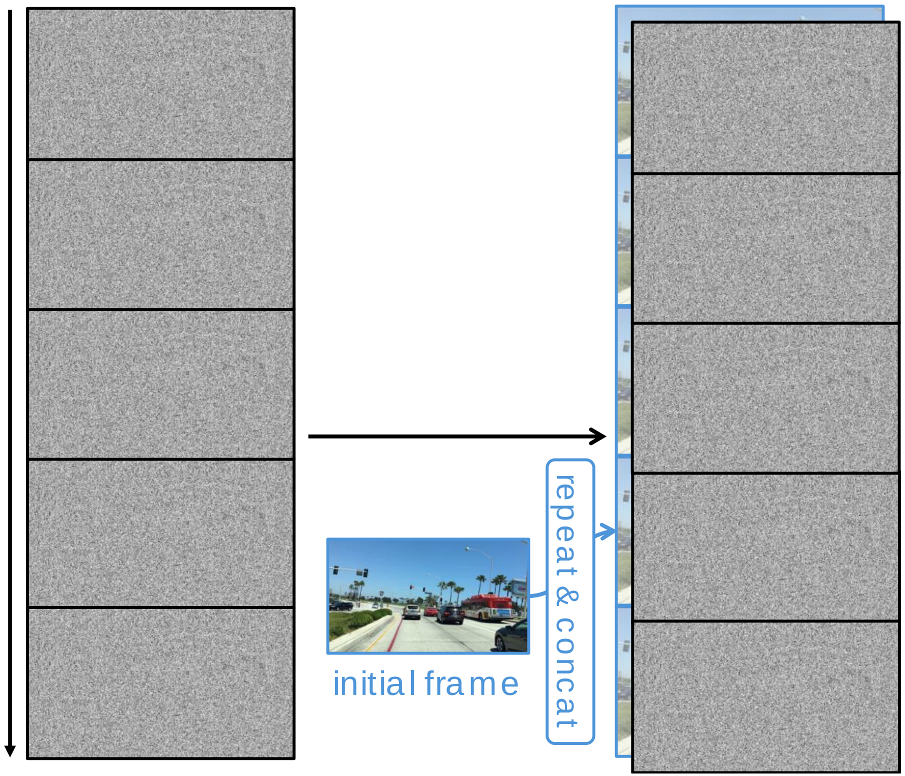
From Generation Model to Prediction Model



Blattmann, Andreas, et al. "Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets." arXiv preprint arXiv:2311.15127 (2023).
Gao, Shenyuan, et al. "Vista: A Generalizable Driving World Model with High Fidelity and Versatile Controllability." Advances in Neural Information Processing Systems 38 (2024).

From Generation Model to Prediction Model

- Impose the first frame as the condition image



Gao, Shenyuan, et al. "Vista: A Generalizable Driving World Model with High Fidelity and Versatile Controllability." Advances in Neural Information Processing Systems 38 (2024).

From Generation Model to Prediction Model

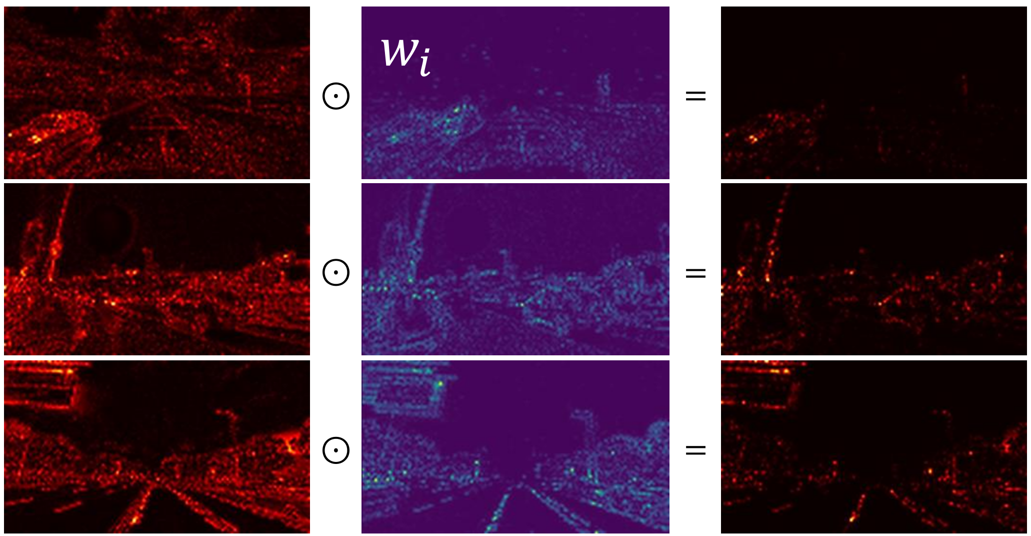
- Inject previous frames as dynamic priors to enable coherent rollout

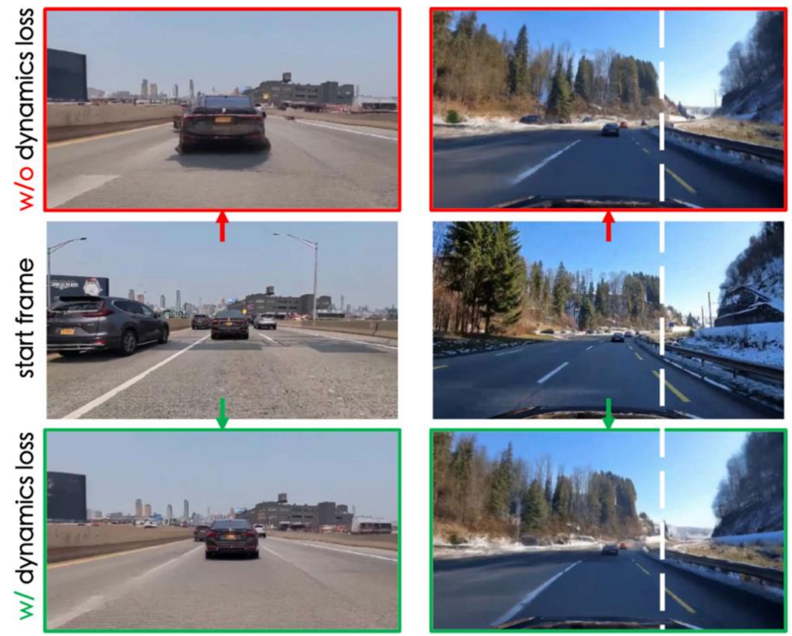
Gao, Shenyuan, et al. "Vista: A Generalizable Driving World Model with High Fidelity and Versatile Controllability." Advances in Neural Information Processing Systems 38 (2024).

Improving Fidelity (for Driving Scenarios)

- Dynamics-critical regions only occupy a rather small area
- How to emphasize?
  - <u>Highlight motion disparities between prediction and ground truth</u>

$$w_i = \|(D_\theta(\hat{n}_i;\sigma) - D_\theta(\hat{n}_{i-1};\sigma)) - (z_i - z_{i-1})\|^2$$



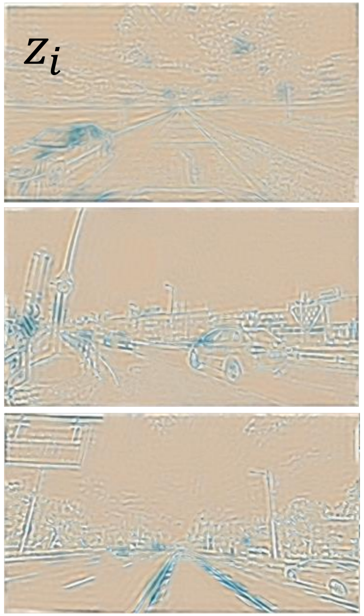


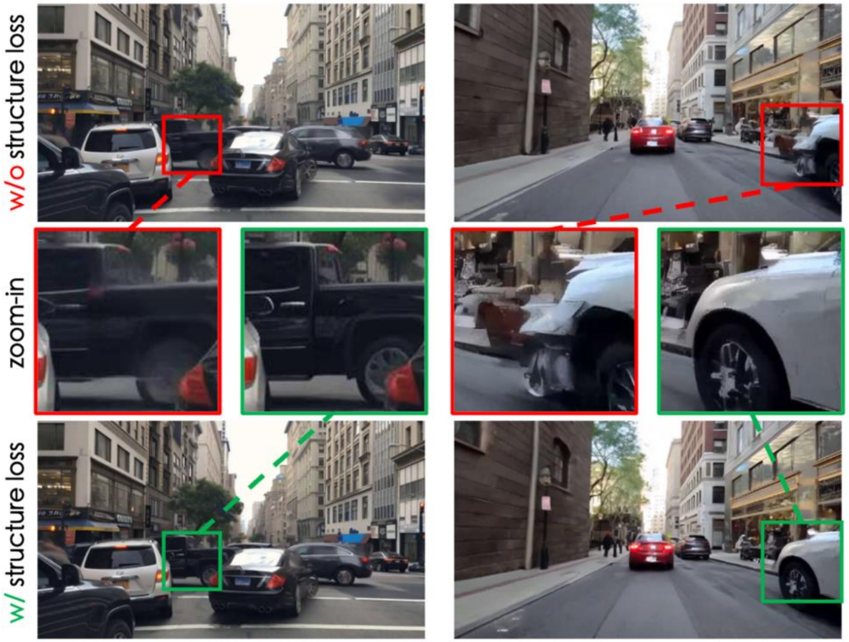
Gao, Shenyuan, et al. "Vista: A Generalizable Driving World Model with High Fidelity and Versatile Controllability." Advances in Neural Information Processing Systems 38 (2024).

Improving Fidelity (for Driving Scenarios)

- Generating fast-moving objects tend to result in corruptions
- How to alleviate?
  - Enhance high-frequency structural information (e.g., edges and lanes)

$$z_i' = \mathcal{F}(z_i) = \mathtt{IFFT}\big(\mathcal{H} \odot \mathtt{FFT}(z_i)\big)$$

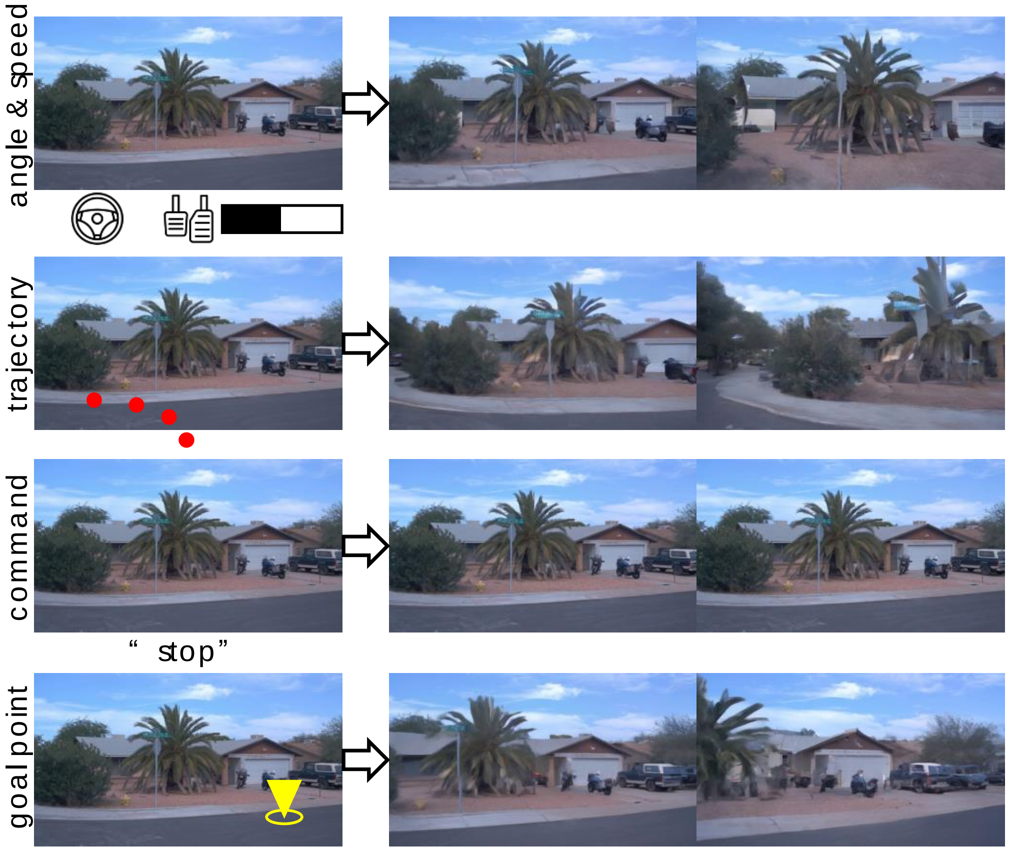


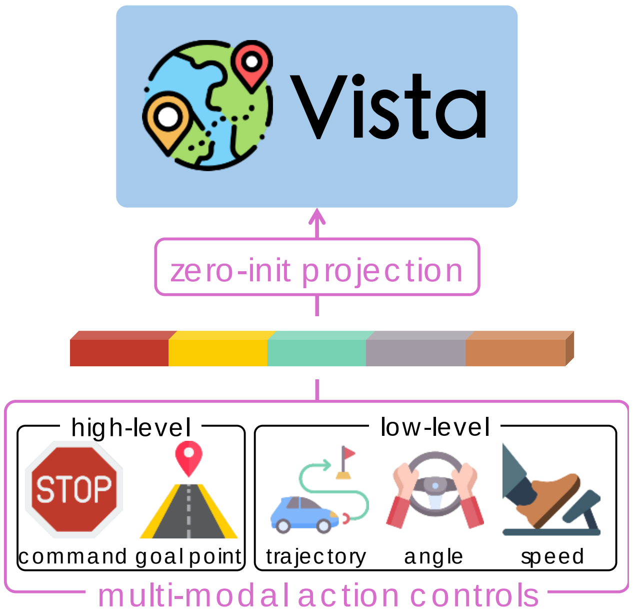


Gao, Shenyuan, et al. "Vista: A Generalizable Driving World Model with High Fidelity and Versatile Controllability." Advances in Neural Information Processing Systems 38 (2024).

Multi-Modal Action Controllability

- Unified control interface for a versatile action suit

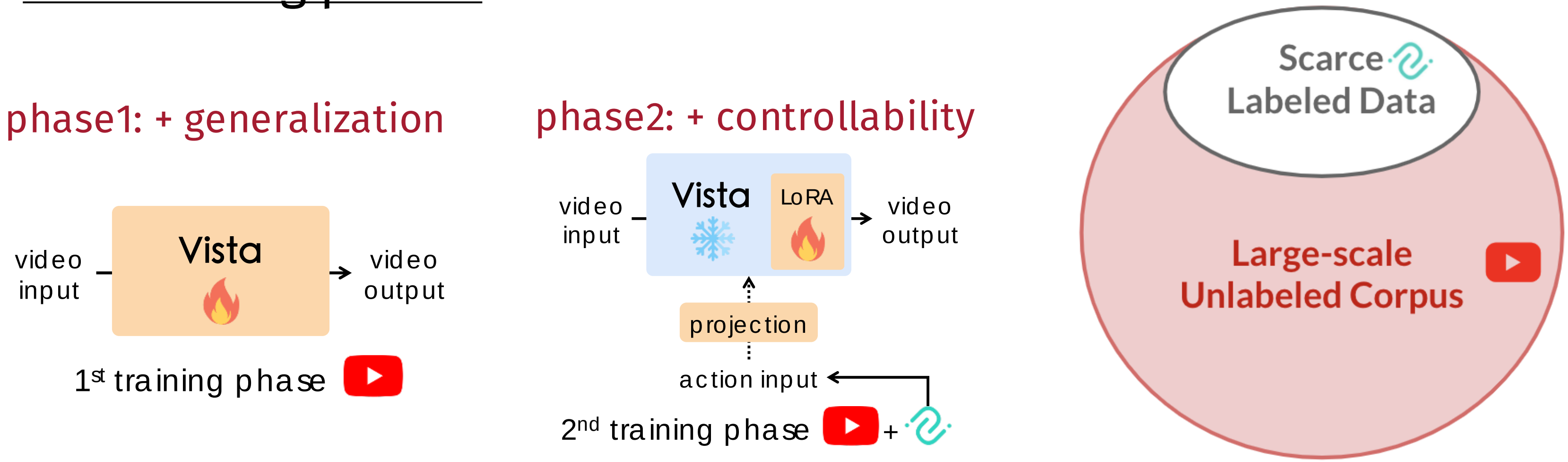




Gao, Shenyuan, et al. "Vista: A Generalizable Driving World Model with High Fidelity and Versatile Controllability." Advances in Neural Information Processing Systems 38 (2024).

Multi-Modal Action Controllability

- Two training phases

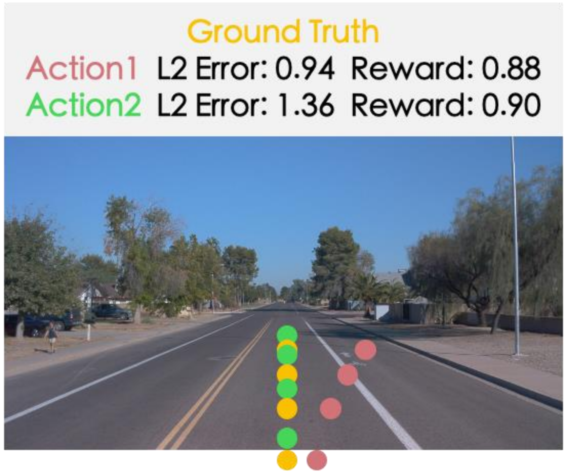


- Effective learning strategies
  - Multi-stage (low-res → high-res)
  - Parameter-efficient LoRA
  - Action independence constraint

Gao, Shenyuan, et al. "Vista: A Generalizable Driving World Model with High Fidelity and Versatile Controllability." Advances in Neural Information Processing Systems 38 (2024).

World Model as A Reward Function

- Our insights
  - OOD action condition will increase prediction uncertainty
- Generalizable reward estimation
  - **Conditional variance** as a source of reward
    - Without referring to ground truth trajectories
    - Inherit the strong generalization ability of Vista
    - Could be used to evaluate actions in the wild



Gao, Shenyuan, et al. "Vista: A Generalizable Driving World Model with High Fidelity and Versatile Controllability." Advances in Neural Information Processing Systems 38 (2024).

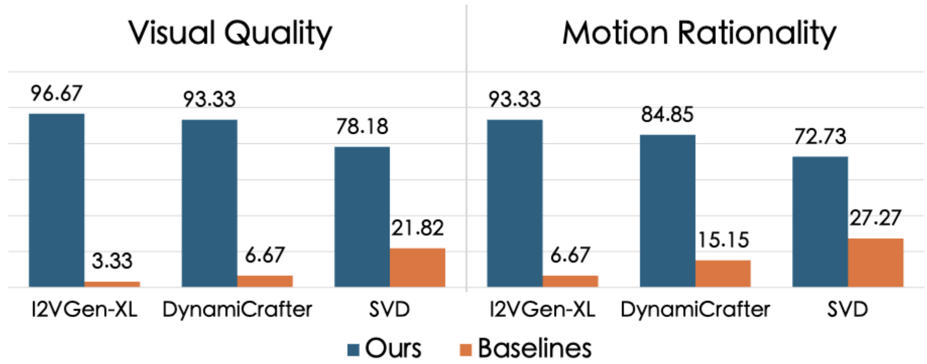
Quantitative Results

ours

- Compared to existing driving world models

Metric	DriveGAN [100]	DriveDreamer [123]	WoVoGen [88]	Drive-WM [125]	GenAD [134]	**Vista (Ours)**
FID ↓	73.4	52.6	27.6	15.8	15.4	**6.9**
FVD ↓	502.3	452.0	417.7	122.7	184.0	**89.4**

- Compared to prominent video generators



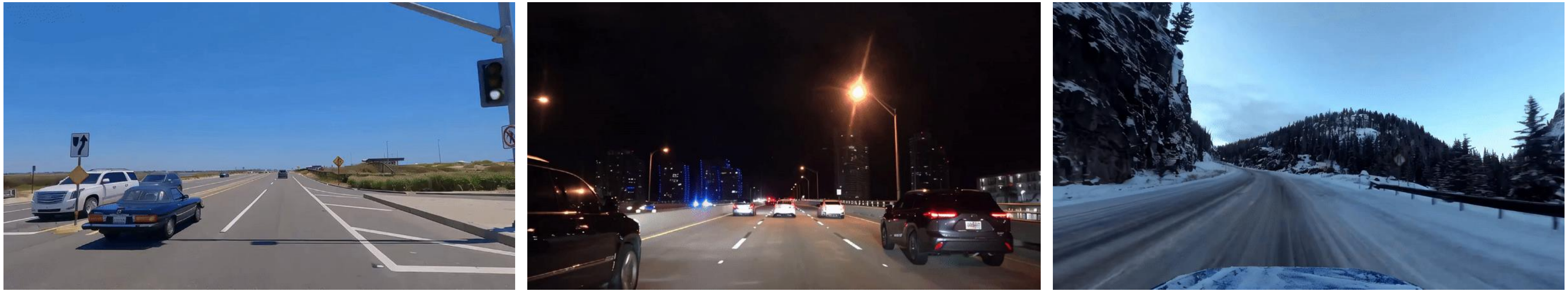
Yang, Jiazhi, et al. "Generalized Predictive Model for Autonomous Driving." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

Gao, Shenyuan, et al. "Vista: A Generalizable Driving World Model with High Fidelity and Versatile Controllability." Advances in Neural Information Processing Systems 38 (2024).

Visualization: Open-World Generalization

- Generalization to novel scenarios

- Coherent long-horizon rollout (up to 15s)



Gao, Shenyuan, et al. "Vista: A Generalizable Driving World Model with High Fidelity and Versatile Controllability." Advances in Neural Information Processing Systems 38 (2024).

Visualization: Action Controllability

turn left | go straight | turn right | stop

Gao, Shenyuan, et al. "Vista: A Generalizable Driving World Model with High Fidelity and Versatile Controllability." Advances in Neural Information Processing Systems 38 (2024).

Visualization: Counterfactual Reasoning Ability

- Could be used for close-loop simulation

Gao, Shenyuan, et al. "Vista: A Generalizable Driving World Model with High Fidelity and Versatile Controllability." Advances in Neural Information Processing Systems 38 (2024).

Our Code and Model are Open-Sourced!

- OpenDV Dataset
  https://github.com/OpenDriveLab/DriveAGI

- Vista Code & Model
  https://github.com/OpenDriveLab/Vista

- Video Demo
  https://opendrivelab.com/Vista

Yang, Jiazhi, et al. "Generalized Predictive Model for Autonomous Driving." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.
Gao, Shenyuan, et al. "Vista: A Generalizable Driving World Model with High Fidelity and Versatile Controllability." Advances in Neural Information Processing Systems 38 (2024).