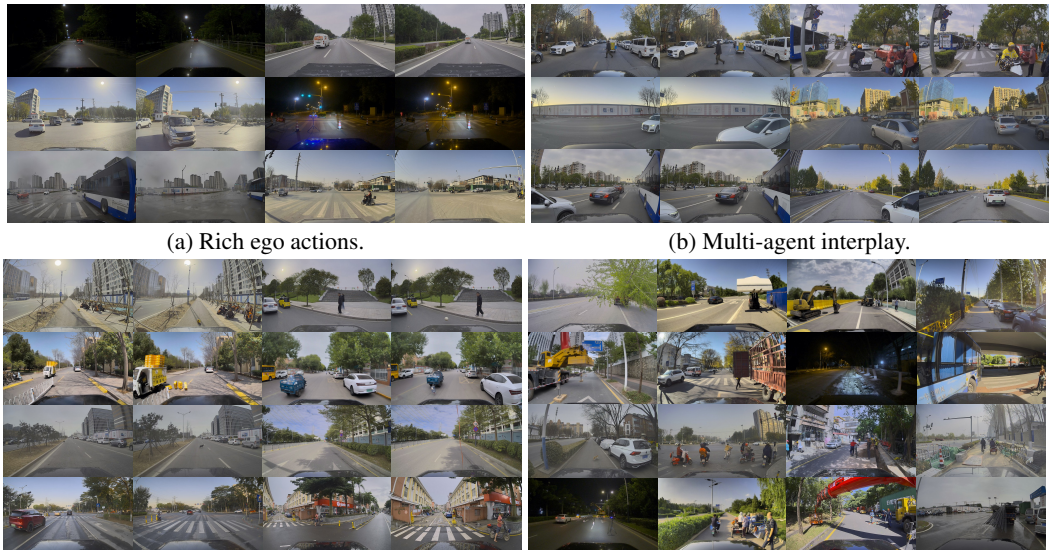
DrivingDojo Dataset: Advancing Interactive and Knowledge-Enriched Driving World Model

Yuqi Wang[1,2†*] **Ke Cheng**[3†] **Jiawei He**[1,2†] **Qitai Wang**[1,2†]
Hengchen Dai[3] **Yuntao Chen**[4✉] **Fei Xia**[3] **Zhaoxiang Zhang**[1,2,4]

[1] New Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences
[3] Meituan Inc. [4] Centre for Artificial Intelligence and Robotics, HKISI, CAS

Project page: https://drivingdojo.github.io



(a) Rich ego actions. (b) Multi-agent interplay.

(c) Rich open-world knowledge. (d) Diverse and rare cases.

Figure 1: **Examples on DrivingDojo.** (a) showcases various driving actions, such as lane changes, abrupt braking at traffic control, and turning at intersections. (b) illustrates the ego-car's interactions with other dynamic agents, including cutting-in and cutting-off maneuvers. (c) displays encounters with rolling or falling objects, moving or floating unknown objects, and interactions with traffic lights and boom barriers. (d) presents diverse cases encountered in real-world driving scenarios.

Abstract

Driving world models have gained increasing attention due to their ability to model complex physical dynamics. However, their superb modeling capability is yet to be fully unleashed due to the limited video diversity in current driving datasets. We introduce DrivingDojo, the first dataset tailor-made for training interactive world models with complex driving dynamics. Our dataset features video clips with a complete set of driving maneuvers, diverse multi-agent interplay, and rich open-world driving knowledge, laying a stepping stone for future world model development. We further define an action instruction following (AIF) benchmark for world models and demonstrate the superiority of the proposed dataset for generating action-controlled future predictions.

*Work done during an internship at Meituan. † equal contributions. ✉ Corresponding author

Submitted to the 38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks. Do not distribute.

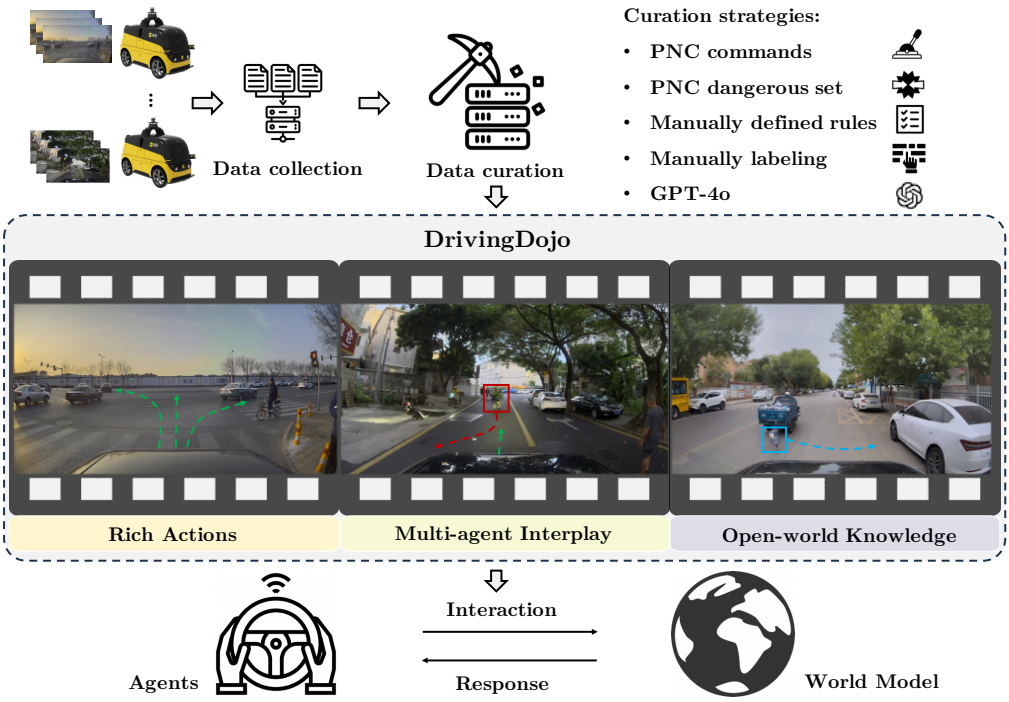


Figure 2: **Enhancing interactive and knowledge-enriched learning of world models.** Data plays a crucial role in modeling the world. DrivingDojo is a large-scale video dataset curated from millions of daily collected videos, designed to investigate real-world visual interactions. DrivingDojo features comprehensive actions, multi-agent interplay, and rich open-world driving knowledge, serving as a superb platform for studying world models.

1 Introduction

World models [16, 19, 32, 20] has received increasing attention due to its ability to model complex real-world physical dynamics and its potential to serve as a general-purpose real-world simulator which is capable of predicting future states in response to diverse action instructions. Facilitated by advancements in video generation techniques [52, 23, 3, 2], models like Sora have achieved remarkable success in producing high-quality videos, thereby opening up a new avenue that treats video generation as real-world dynamics modeling problem [46, 18]. Generative world models, in particular, hold significant promise as real-world simulators and have garnered extensive research in the field of autonomous driving [27, 47, 29, 48, 53, 58].

However, existing driving world models fall short of meeting the requirements of model-based planning in autonomous driving, which aims to improve driving safety in scenarios with diverse ego maneuvers and intricate interaction between the ego vehicle and other road users. These models perform well for non-interactive in-lane maneuvers but have shown limited capability in following more challenging action instructions like lane change. One significant roadblock to building next-generation driving world models lies in the datasets. Autonomous driving datasets commonly used in current world model literature like nuScenes [6], Waymo [44], and ONCE [36], are primarily designed and curated in a perception-oriented manner. As a result, it contains limited driving patterns and multi-agent interactions, which may not fully capture the complexities of real-world driving scenarios. The scarcity of interaction data limits the ability of models to accurately simulate and predict the complex dynamics of real-world driving environments.

In this paper, we propose **DrivingDojo**, a large-scale driving video dataset designed to simulate real-world visual interaction. As illustrated in Figure 1, DrivingDojo features action completeness, multi-agent interplay, and open-world driving knowledge. Our dataset aims to unleash the full action instruction following the ability of world model by including rich lateral maneuvers like acceleration, emergency braking and stop-and-go as well as longitudinal ones like U-turn, overtaking, and lane

Table 1: **A comparison of driving datasets for world model**.* denotes that the videos are curated from our data pool of around 7500 hours.

Dataset	Videos	Duration (hours)	Ego Trajectory	Complete Actions	Multi-agent Interplay	Open-world Knowledge
nuScenes [6]	1k	5.5	✓			
Waymo [44]	1k	11	✓			
OpenDV-2k [53]	2k	2059		✓		
nuPlan [7]	-	1500	✓	✓	✓	
DrivingDojo (Ours)	18k	150*	✓	✓	✓	✓

change. Besides, we explicitly curate the dataset to include a large volume of trajectories containing multi-agent interplays like cut-in, cut-off, and head-to-head merging. Finally, DrivingDojo taps into the open-world driving knowledge by including videos containing rare events sampled from tens of millions of driving video clips, including crossing animals, falling bottles and debris. As shown in Figure 2, we hope that DrivingDojo could serve as a solid stepping stone for developing next-generation driving world models.

To measure the progress of driving scene modeling, we propose a new action instruction following (AIF) benchmark to assess the ability of world models to perform plausible future rollouts. The AIF benchmark measures the visual and structural fidelity of videos generated by world models in an action-conditioned manner. We propose the AIF errors calculated on the withheld validation data to evaluate the long-term motion controllability for generated videos. The error is defined as the mean error between the actions estimated from the generated video and the given action instructions. Then the baseline world model is evaluated on our DrivingDojo AIF benchmark, for in-domain data and out-of-domain images or action conditions.

Our major contributions are as follows. (1) We design a large-scale driving video dataset to facilitate research in world model for autonomous driving. Compared to previous datasets in Table 1, our dataset features complete driving actions, diverse multi-agent interplay, and rich open-world driving knowledge. (2) We design an action instruction following task for driving world model and provide corresponding video world model baseline methods. (3) Benchmark results on both driving video generation and action instruction following show that there are plenty of new opportunities for future driving world model development on our new dataset.

2 Related Works

2.1 Autonomous Driving Datasets

Datasets for perception. The driving dataset has played a crucial role in advancing computer vision in recent years, aiming to achieve comprehensive perception and understanding surrounding the ego vehicle. Initially, perception in autonomous driving relied on 2D image-based perception. Datasets like Cityscapes [10], Mapillary Vistas [38], and BDD100k [56] provided instance-level masks for learning tasks. With the integration of LiDAR sensors and advancements in 3D perception, datasets like KITTI [13], nuScenes [6], and Waymo [44] have emerged as standard benchmarks for various 3D perception tasks. Additionally, datasets like ONCE [36], Argoverse [8, 49], and others [28, 14, 1] are also utilized for studying various perception tasks.

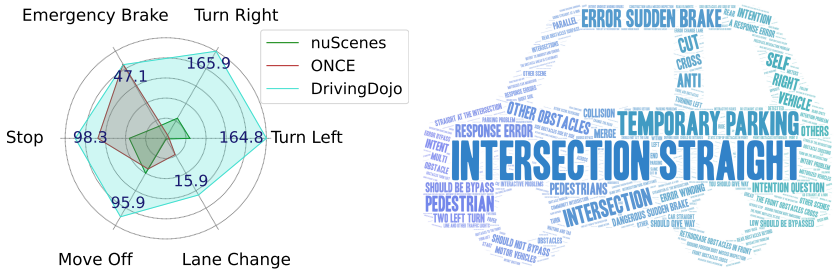
Datasets for prediction and planning. In recent years, there's been increasing attention on prediction and planning in autonomous driving. Prediction involves anticipating the behavior of other agents, while planning relates to the behavior of the ego vehicle. Prediction methods typically rely on semantic maps and dynamic traffic light statuses to anticipate future vehicle motions. Notable datasets in this area include Argoverse Motion Forecasting [8], Waymo Open Motion Dataset [12], Lyft Level 5 Prediction Dataset [25], and nuScenes Prediction [6] challenge. Additionally, the Interaction dataset [57] provides interactive driving scenarios with semantic maps derived from drones and traffic cameras, enriching the understanding of complex driving interactions. Transitioning to planning, CARLA [11] stands out as an open-source simulator designed to simulate real-world traffic scenarios,

providing a platform for testing and validating planning algorithms. Complementing this, nuPlan [7] introduces the first closed-loop planning benchmark for autonomous vehicles, closely mirroring real-world scenarios.

2.2 World Model

Learning world models. World models [16, 32] enable next-frame prediction based on action inputs, aiming to build general simulators of the physical world. However, learning dynamic modeling in pixel space is challenging, leading previous image-based world models to focus on simplistic gaming environments or simulations [17, 19, 9, 51, 43, 42, 20]. With advances in video generation, models like Sora can now produce high-definition videos up to one minute long with natural, coherent dynamics. This progress has encouraged researchers to explore world models in real-world scenarios. DayDreamer [50] applies the Dreamer algorithm to four robots, allowing them to learn online and directly in the real world without simulators, demonstrating that world models can facilitate faster learning on physical robots. Genie [5] demonstrates interactive generation capabilities using vast internet gaming videos and shows potential for robotics applications. UniSim [54] aims to create a universal simulator for real-world interactions using generative modeling, with applications extending to real-robot executions.

World model for autonomous driving. World models serving as real-world simulators have garnered widespread attention [15] and can be categorized into two main branches. The first branch explores agent policies in virtual simulators. MILE [26] employed imitation learning to jointly learn the dynamics model and driving behavior in CARLA [11]. Think2Drive [33] proposed a model-based RL method in CARLA v2, using a world model to learn environment transitions and acting as a neural simulator to train the planner. The second branch focuses on simulating and generating real-world driving scenarios. GAIA-1 [27] introduced a generative world model for autonomous driving, capable of simulating realistic driving videos from inputs like images, texts, and actions. DriveDreamer [47] emphasized scenario generation, leveraging HD maps and 3D boxes to enhance video quality. Drive-WM [48] was the first to propose a multiview world model for generating high-quality, controllable multiview videos, exploring applications in end-to-end planning. ADriver-I [29] constructed a general world model based on MLLM and diffusion models, using vision-action pairs to auto-regressively predict current frame control signals. DriveDreamer2 [58] leveraged LLMs and text prompts to generate diverse driving videos in a user-friendly manner. Unlike previous methods that focused on model design, OpenDV-2K [53] addressed the issue of training data by collecting over 2000 hours of driving videos from the internet. Previous research has predominantly addressed static scene generation, with limited emphasis on multi-agent interplays. Our dataset enables the exploration of world model predictions within dynamic, interactive driving scenarios.



(a) Action distribution (b) Video clip description

Figure 3: **The strengths of the DrivingDojo dataset.** (a) illustrates a comparison of action distributions among nuScenes, ONCE, and our DrivingDojo. We compare the average hourly event counts of driving actions. (b) presents the distribution of text descriptions for the video clips in DrivingDojo.

3 The DrivingDojo Dataset

Our goal is to provide a large and diverse action-instructed driving video dataset DrivingDojo to support the development of driving world models. To accomplish this, we extract highly informative

clips from a video pool collected through fleet data, spanning several years and comprising more than 500 operating vehicles across multiple major Chinese cities. As a result, our DrivingDojo features diverse ego actions, rich interactions with road users, and rare driving knowledge which are crucial for high-quality future forecasting as shown in Table 2.

We begin with the design principles of DrivingDojo and its uniqueness compared with existing datasets in Section 3.1- 3.3. We then describe the data curation procedure and statistics in Section 3.4. Here, we only describe the design principles. More detailed information refer to the Appendix.

Table 2: **DrivingDojo constitution.** We also organize the dataset into three subsets: DrivingDojo-Action, DrivingDojo-Interplay, and DrivingDojo-Open, to facilitate research on specific tasks.

Dataset	Videos	Type	Camera	Ego Trajectory	Text Description
DrivingDojo	18.2k	total	✓	✓	✓
DrivingDojo-Action	7.9k	rich ego-actions	✓	✓	
DrivingDojo-Interplay	6.4k	multi-agent interplay	✓	✓	
DrivingDojo-Open	3.9k	open-world knowledge	✓	✓	✓

3.1 Action Completeness

Using the driving world model as a real-world simulator requires it to follow action prompts accurately. Existing autonomous driving datasets, such as ONCE [36] and nuScenes [6], are generally curated for developing perception algorithms and thus lack diverse driving maneuvering.

To enable the world model to generate an infinite number of high-fidelity, action-controllable virtual driving environments, we create a subset called DrivingDojo-Action that features a balanced distribution of driving maneuvers. This subset includes a diverse range of both longitudinal maneuvers, such as acceleration, deceleration, emergency braking, and stop-and-go driving, as well as lateral maneuvers, including lane-changing and lane-keeping. As demonstrated in Figure 3a, our DrivingDojo-Action subset offers a significantly more balanced and complete set of ego actions compared to existing autonomous driving datasets.

3.2 Multi-agent Interplay

Besides navigating in a static road network environment, modeling the dynamics of multi-agent interplay like merge and yield is also a crucial task for world models. However, current datasets are either built without considering multi-agent interplays, such as nuScenes [6] and Waymo [44], or are constructed from large-scale internet videos that lack proper curation and balancing, like OpenDV-2K [53].

To address this issue, we design the DrivingDojo-Interplay subset focusing on interactions with dynamic agents as a core component of the dataset. As shown in Figure 1b, we curate this subset to include at least one of the following driving scenarios: cutting in/off, meeting, blocked, overtaking, and being overtaken. These scenarios encompass a variety of realistic situations, such as vehicles cutting into lanes, encounters with oncoming traffic, and the necessity for emergency braking. By incorporating these diverse scenarios, our dataset enables world models to better understand and anticipate complex interactions with dynamic agents, thereby improving their performance in real-world driving conditions.

3.3 Rich Open-world Knowledge

In contrast to perception and prediction models, which compress high-dimensional sensor input into low-dimensional vector representations, world models exhibit a superior modeling capacity by operating in the pixel space. This increased capacity enables world models to effectively capture the intricate dynamics of open-world driving scenarios, such as animals unexpectedly crossing the road or parcels falling off the trunks of vehicles.

However, existing datasets, either perception-oriented ONCE [36] or planning-oriented ones like nuPlan [7], do not have adequate data for developing and assessing the long-tail knowledge modeling

ability of world models. Therefore, we place a unique emphasis on including rich open-world knowledge video clips and construct the DrivingDojo-Open subset. As shown in Figure 1c, describing open-world driving knowledge like this is challenging due to its complexity and variability, but these scenarios are crucial for ensuring safe driving.

The DrivingDojo-Open subset consists of 3.9k video clips about the open-world knowledge in driving scenarios. This subset is curated from fleet data that includes unusual weather, foreign objects on the road surface, floating obstacles, falling objects, taking over cases, and interactions with traffic lights and boom barriers. A word cloud of video descriptions for DrivingDojo-Open are shown in Figure 3b. DrivingDojo-Open serves as an invaluable supplementary for driving world modeling by including driving knowledge beyond simply interacting with structured road networks and other regular road users.

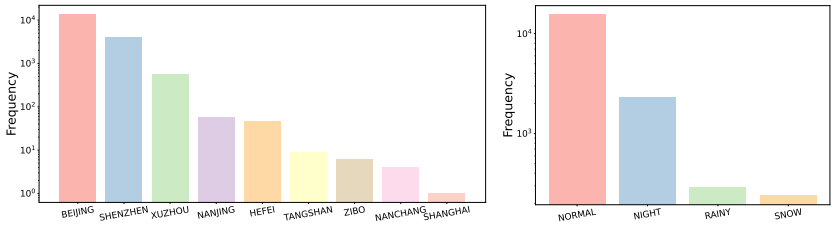


Figure 4: **Descriptive statistics of the DrivingDojo dataset.** The dataset was collected from various regions across China, including nighttime and rainy/snowy conditions.

3.4 Data Curation and Statistics

Dataset statistics. The DrivingDojo dataset contains 18k videos with resolution of 1920×1080 and frame rate at 5 fps. Our video clips are collected from major Chinese cities including Beijing, Shenzhen, Xuzhou, etc., as shown in Figure 4. Furthermore, these videos are recorded in diverse weather conditions at different daylight conditions. All videos are paired with synced camera poses derived from the HD-Map powered high precision localization stack onboard. Videos in the DrivingDojo-Open subset are paired with text descriptions about the rare event happening in each video. More details are in the Appendix.

Data collection. We collected multi-modal fleet data using the platform of Meituan's autonomous delivery vehicles. Our dataset consists of video clips recorded by the front-view camera with a horizontal field of view of 120° to capture comprehensive visual information. The raw data is collected from multiple Chinese cities between May 2022 and May 2024, amassing a total of 900,000 videos and approximately 7,500 hours of driving footage pre-filtered before recording.

Data curation. In order to ensure both the data diversity as well as balanced ego action and multi-agent interplay distribution, we include fleet data with different criteria. The data sources of DrivingDojo include 1) intervention data from safety inspectors during vehicle operation, 2) emergency brake data from automatic emergency braking, 3) randomly sampled 30-second general videos from collected videos, 4) selected distinct scenarios such as traffic light changes, barrier opening, left and right turns, straight crossings, vehicle encounters, lane changes, and pedestrian interactions, 5) manually sorted rare data containing moving and static foreign objects on the road, floating obstacles, falling and rolling objects. The curation details are in the Appendix.

Personal Identification Information (PII) removal. To avoid privacy infringement and obey the regulation laws, we employ a high precision license plate and face detectors [30] to detect and blur these PII for each frame of all videos. An in-house annotation team and the authors have manually double-checked that the PII removal procedure is correctly carried out for all the videos.

4 DrivingDojo for World Model

To facilitate the study of world models in autonomous driving, we define a novel action instruction following (AIF) task. We provide baseline methods (Section 4.2) and evaluation metrics (Section 4.3), enabling further investigations. More details are described in the Appendix.

4.1 Action Instruction Following

Action-controllable video forecasting is the core ability of world models [5]. Instead of solely focusing on predicting high-quality video frames, action instruction following requires world models to take both the initial video frame and ego action prompts into consideration for predicting corresponding world responses. Given the initial image I_t and a sequence of actions $\{A_t, ..., A_{t+k}\}$, the model f_θ predicts future states $\{I_{t+1}, ..., I_{t+k}\}$ as:

$$\{I_t, ..., I_{t+k}\} = f_\theta(I_t, \{A_t, ..., A_{t+k}\}). \tag{1}$$

Here, $\{A_t, ..., A_{t+k}\}$ refers to the action prompts for each frame, with trajectories $A_t = (\Delta x_t, \Delta y_t)$ in our experiment. f_θ represents the world model, and $\{I_{t+1}, ..., I_{t+k}\}$ signifies the visual prediction for subsequent k frames.

4.2 Model Architecture

We propose DrivingDojo baseline, a video generation model based on Stable Video Diffusion (SVD) [2]. While SVD is a latent diffusion model for image-to-video generation, we extend its capability to generate videos conditioned on action. For the AIF task, we encode the value of each action sequence into a 1024-dimensional vector using a Multilayer Perceptron (MLP). Subsequently, the action feature is concatenated with the first-frame image feature and passed into the U-Net [39].

4.3 Evaluation Metrics

Visual quality. To evaluate the quality of the generated video, we utilize FID (Frechet Inception Distance) [22] and FVD (Frechet Video Distance) [45] as the main metrics.

Action instruction following. We propose the action instruction following (AIF) errors E_x^{AIF} and E_y^{AIF} to measure the consistency between the generated video and the input action conditions. Given the generated video sequences $\{I_t, ..., I_{t+k}\}$, we estimate vehicle trajectories in the generated videos with the offline visual structure-from-motion (SfM) implementation like COLMAP [40, 41]: $\{\widetilde{A}_t, ..., \widetilde{A}_{t+k}\} = \text{SfM}(\{I_t, ..., I_{t+k}\})$, where $\{\widetilde{A}_t, ..., \widetilde{A}_{t+k}\}$ are estimated trajectories of unknown scale. We estimated the scale factor $\hat{S}$ for the predicted trajectory by minimizing the error between estimated and input ego-motion in the first N frames. We compare the estimated actions with the ground-truth action instructions $\{A_t, ..., A_{t+k}\}$ and report the mean absolute error for both lateral (E_y^{AIF}) and longitudinal (E_x^{AIF}) actions:

$$(E_x^{\text{AIF}}, E_y^{\text{AIF}}) = \frac{\sum_{i=0}^{k} |A_{t+i} - \widetilde{A}_{t+i} * \hat{S}|}{k+1}, \tag{2}$$

where the scale factor $\hat{S} = \arg\min_S \sum_{i=0}^{N} |A_{t+i} - \widetilde{A}_{t+i} * S|$.

Table 3: **Comparison of visual prediction fine-tuning across different datasets.**, † indicates using camera sweeps data. The performance is zero-shot evaluated on the OpenDV-2K dataset.

Method	Fine-tuning	Evaluation	FID	FVD
SVD [2]	-	OpenDV-2K	24.17	580.94
SVD [2]	nuScenes†	OpenDV-2K	21.05	395.04
SVD [2]	DrivingDojo	OpenDV-2K	**19.20**	**343.91**

5 Experiments

5.1 Results of Visual Prediction

To illustrate the richness of behaviors and dynamics within our dataset, we compare video fine-tuning quality across various datasets. In Table 3, we random selected 256 video segments from the OpenDV-

2K dataset [53] as our test set and evaluated fine-tuning performance of SVD [2] model across various datasets. The results indicate that models trained on our dataset exhibit better visual quality.

5.2 Results of Action Instruction Following

Diverse driving behaviors. Based on different sequences of actions, our model is able to generate multiple possible futures. As shown in Figure 5, we showcase the model's capability to execute forward, left turn, and right turn maneuvers at intersections, as well as lane-changing to the left or right, and maintaining on straight roads.

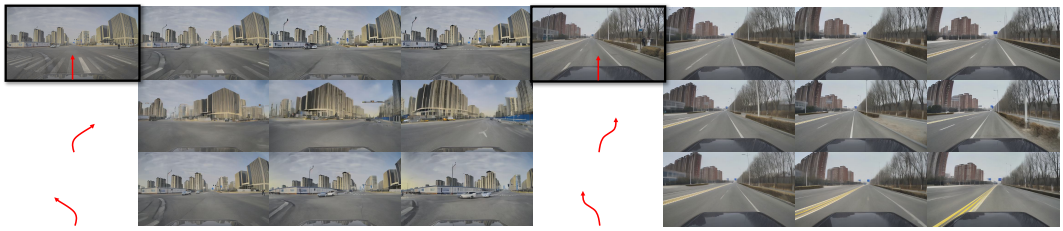


Figure 5: **Predicting multiple futures based on different actions.** Left: going straight, turning left, and turning right at a crossing; Right: changing to the left lane, staying in the current lane, and changing to the right lane.

Action instruction following. Although qualitative evaluations demonstrate the powerful generative ability of our model, we also endeavor to measure the accuracy of action instruction following quantitatively. We seek to evaluate whether the video trajectories generated by the model closely adhere to our expected route paths. This serves as a fundamental assurance for the future application of world model. As shown in Table 4, with the in-domain actions (original action sequences of the test video) as conditions, videos generated by the baseline world model trained on DrivingDojo exhibit strong loyalty towards the action instructions. The mean action error in each video frame is limited to only 6-7 cm in the lateral or longitudinal directions. In row 2, feeding the model with the same initial images and randomly sampled action instructions slightly increases the mean action errors. When the model is applied zero-shot to initial images from OpenDV-2K [53] and fed with randomly sampled action instructions, its generated videos still demonstrate considerable consistency to the action instructions. Note that the proposed action instruction following errors can sensitively reflect the impact of out-of-domain inputs on the performance of the model.

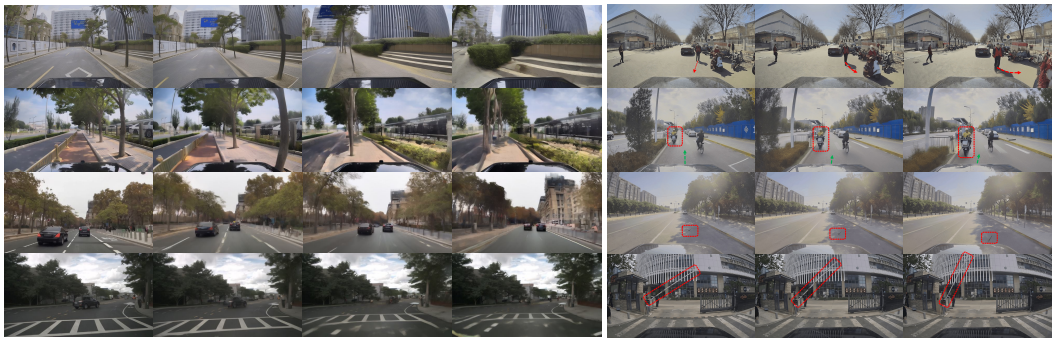
Table 4: **Action instruction following on the DrivingDojo dataset**. * denotes the model is applied zero-shot to this dataset without fine-tuning.

Action Type	Test Dataset	FID	FVD	$E_x^{\text{AIF}}(\downarrow)$	$E_y^{\text{AIF}}(\downarrow)$
In-Domain	DrivingDojo	37.07	658.72	0.068m	0.060m
Out-of-Domain	DrivingDojo	38.30	716.44	0.133m	0.099m
Out-of-Domain	OpenDV-2K*	24.27	442.67	0.175m	0.121m

5.3 Real-world Simulation

Action generalization. Our model demonstrates robust generalization capabilities in two key aspects. As illustrated in Figure 6a, firstly, it effectively generalizes to out-of-domain (OOD) actions, such as forcefully driving on pedestrian walkways, showcasing its adaptability to some unreasonable actions. Secondly, it successfully extends its capabilities to other datasets, executing tasks such as lane changes on the OpenDV-2K [53] dataset and backing-the-car maneuvers on the nuScenes [6] dataset without requiring further fine-tuning. This underscores the model's potential as a real-world simulator, capable of adapting to diverse driving scenarios.

Dynamic agents. We showcase our model's ability to simulate interactions with dynamic agents in Figure 6b. The results indicate that the model can provide reasonable responses based on our actions.



(a) Action generalization (b) Interaction simulation

Figure 6: **Qualitative examples of our model's capability.**

The first scenario depicts a pedestrian opting to yield as our vehicle continues forward, resulting in a change in trajectory. In the second scenario, a delivery person opts to stop and wait at a narrow road.

Open-world dynamics. In Figure 6b, our model showcases the simulations of rare scenarios encountered on the road, including interactions with moving birds and parking lot barriers.

5.4 Limitations and Future Work

This paper primarily delves into the value of the dataset, with the model aspect serving as a baseline without any special design. Although the DrivingDojo dataset significantly improves model capabilities, there are still several limitations that require further investigation in future studies.

Hallucination. As shown in Figure 7, we observed that the model exhibits some hallucinations, such as the sudden disappearance of objects, and when an action is unrealistic given the scene, such as forcefully turning right, the model sometimes imagines a new road.

Figure 7: **Examples of hallucination**. Top: object suddenly disappears. bottom: a non-existed road.

Long-horizon visual prediction. Our baseline model is only capable of generating short videos, which can be used to simulate short-term interaction events. Longer predictions [4, 55, 21] and faster generation [37, 35] are left for future research.

6 Conclusion

In this work, we present DrivingDojo, a large-scale video dataset aimed at advancing the study of driving world models. DrivingDojo offers a testbed for studying diverse real-world interactions. Our findings indicate that simulating interactions and rare dynamics observed in open-world environments remains an unsolved challenge, highlighting significant opportunities for future research.

Societal impacts. By providing a comprehensive dataset covering diverse driving scenarios and behaviors, researchers can develop and refine algorithms that increase the safety, reliability, and efficiency of autonomous vehicles. However, the development of driving world model requires large and diverse driving videos, introducing privacy issues.

References

[1] Mina Alibeigi, William Ljungbergh, Adam Tonderski, Georg Hess, Adam Lilja, Carl Lindström, Daria Motorniuk, Junsheng Fu, Jenny Widahl, and Christoffer Petersson. Zenseact open dataset: A large-scale and diverse multimodal dataset for autonomous driving. In *ICCV*, 2023. 3

[2] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 2, 7, 8, 25

[3] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *CVPR*, 2023. 2

[4] Tim Brooks, Janne Hellsten, Miika Aittala, Ting-Chun Wang, Timo Aila, Jaakko Lehtinen, Ming-Yu Liu, Alexei Efros, and Tero Karras. Generating long videos of dynamic scenes. *NeurIPS*, 35, 2022. 9

[5] Jake Bruce, Michael Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. *arXiv preprint arXiv:2402.15391*, 2024. 4, 7

[6] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020. 2, 3, 5, 8, 25

[7] Holger Caesar, Juraj Kabzan, Kok Seang Tan, Whye Kit Fong, Eric Wolff, Alex Lang, Luke Fletcher, Oscar Beijbom, and Sammy Omari. nuplan: A closed-loop ml-based planning benchmark for autonomous vehicles. *arXiv preprint arXiv:2106.11810*, 2021. 3, 4, 5

[8] Ming-Fang Chang, John Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, et al. Argoverse: 3d tracking and forecasting with rich maps. In *CVPR*, 2019. 3

[9] Chang Chen, Jaesik Yoon, Yi-Fu Wu, and Sungjin Ahn. Transdreamer: Reinforcement learning with transformer world models. In *Deep RL Workshop NeurIPS 2021*, 2021. 4

[10] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 3

[11] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *CoRL*, 2017. 3, 4

[12] Scott Ettinger, Shuyang Cheng, Benjamin Caine, Chenxi Liu, Hang Zhao, Sabeek Pradhan, Yuning Chai, Ben Sapp, Charles R Qi, Yin Zhou, et al. Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset. In *ICCV*, 2021. 3

[13] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012. 3

[14] Jakob Geyer, Yohannes Kassahun, Mentar Mahmudi, Xavier Ricou, Rupesh Durgesh, Andrew S Chung, Lorenz Hauswald, Viet Hoang Pham, Maximilian Mühlegg, Sebastian Dorn, et al. A2d2: Audi autonomous driving dataset. *arXiv preprint arXiv:2004.06320*, 2020. 3

[15] Yanchen Guan, Haicheng Liao, Zhenning Li, Jia Hu, Runze Yuan, Yunjian Li, Guohui Zhang, and Chengzhong Xu. World models for autonomous driving: An initial survey. *IEEE Transactions on Intelligent Vehicles*, 2024. 4

[16] David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018. 2, 4

[17] Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *ICLR*, 2019. 4

[18] Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *ICML*, 2019. 2

[19] Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with discrete world models. *arXiv preprint arXiv:2010.02193*, 2020. 2, 4

[20] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023. 2, 4

[21] Roberto Henschel, Levon Khachatryan, Daniil Hayrapetyan, Hayk Poghosyan, Vahram Tadevosyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Streamingt2v: Consistent, dynamic, and extendable long video generation from text. *arXiv preprint arXiv:2403.14773*, 2024. 9

[22] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NeurIPS*, 30, 2017. 7, 22

[23] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 2

[24] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 22

[25] John Houston, Guido Zuidhof, Luca Bergamini, Yawei Ye, Long Chen, Ashesh Jain, Sammy Omari, Vladimir Iglovikov, and Peter Ondruska. One thousand and one hours: Self-driving motion prediction dataset. In *CoRL*, 2021. 3

[26] Anthony Hu, Gianluca Corrado, Nicolas Griffiths, Zachary Murez, Corina Gurau, Hudson Yeo, Alex Kendall, Roberto Cipolla, and Jamie Shotton. Model-based imitation learning for urban driving. *NeurIPS*, 2022. 4

[27] Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023. 2, 4

[28] Xinyu Huang, Peng Wang, Xinjing Cheng, Dingfu Zhou, Qichuan Geng, and Ruigang Yang. The apolloscape open dataset for autonomous driving and its application. *TPAMI*, 42(10):2702–2719, 2019. 3

[29] Fan Jia, Weixin Mao, Yingfei Liu, Yucheng Zhao, Yuqing Wen, Chi Zhang, Xiangyu Zhang, and Tiancai Wang. Adriver-i: A general world model for autonomous driving. *arXiv preprint arXiv:2311.13549*, 2023. 2, 4

[30] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics YOLO, January 2023. 6

[31] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *NeurIPS*, 35, 2022. 22

[32] Yann LeCun. A path towards autonomous machine intelligence. 2022. 2, 4

[33] Qifeng Li, Xiaosong Jia, Shaobo Wang, and Junchi Yan. Think2drive: Efficient reinforcement learning by thinking in latent world model for quasi-realistic autonomous driving (in carla-v2). *arXiv preprint arXiv:2402.16720*, 2024. 4

[34] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2018. 22

[35] Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023. 9

[36] Jiageng Mao, Minzhe Niu, Chenhan Jiang, Hanxue Liang, Jingheng Chen, Xiaodan Liang, Yamin Li, Chaoqiang Ye, Wei Zhang, Zhenguo Li, et al. One million scenes for autonomous driving: Once dataset. *arXiv preprint arXiv:2106.11037*, 2021. 2, 3, 5, 25

[37] Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *CVPR*, 2023. 9

[38] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *ICCV*, 2017. 3

[39] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241, 2015. 7

[40] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 7, 22

[41] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016. 7, 22

[42] Younggyo Seo, Danijar Hafner, Hao Liu, Fangchen Liu, Stephen James, Kimin Lee, and Pieter Abbeel. Masked world models for visual control. In *CoRL*, 2023. 4

[43] Younggyo Seo, Kimin Lee, Stephen L James, and Pieter Abbeel. Reinforcement learning with action-free pre-training from videos. In *ICML*, 2022. 4

[44] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *CVPR*, 2020. 2, 3, 5

[45] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 7, 22

[46] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. *NeurIPS*, 2016. 2

[47] Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, and Jiwen Lu. Drivedreamer: Towards real-world-driven world models for autonomous driving. *arXiv preprint arXiv:2309.09777*, 2023. 2, 4

[48] Yuqi Wang, Jiawei He, Lue Fan, Hongxin Li, Yuntao Chen, and Zhaoxiang Zhang. Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving. *CVPR*, 2024. 2, 4

[49] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, et al. Argoverse 2: Next generation datasets for self-driving perception and forecasting. *arXiv preprint arXiv:2301.00493*, 2023. 3

[50] Philipp Wu, Alejandro Escontrela, Danijar Hafner, Pieter Abbeel, and Ken Goldberg. Daydreamer: World models for physical robot learning. In *CoRL*, 2023. 4

[51] Ziyi Wu, Nikita Dvornik, Klaus Greff, Thomas Kipf, and Animesh Garg. Slotformer: Unsupervised visual dynamics simulation with object-centric models. In *ICLR*, 2023. 4

[52] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*, 2021. 2

[53] Jiazhi Yang, Shenyuan Gao, Yihang Qiu, Li Chen, Tianyu Li, Bo Dai, Kashyap Chitta, Penghao Wu, Jia Zeng, Ping Luo, et al. Generalized predictive model for autonomous driving. *arXiv preprint arXiv:2403.09630*, 2024. 2, 3, 4, 5, 8, 25

[54] Mengjiao Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. *arXiv preprint arXiv:2310.06114*, 2023. 4

[55] Shengming Yin, Chenfei Wu, Huan Yang, Jianfeng Wang, Xiaodong Wang, Minheng Ni, Zhengyuan Yang, Linjie Li, Shuguang Liu, Fan Yang, et al. Nuwa-xl: Diffusion over diffusion for extremely long video generation. *arXiv preprint arXiv:2303.12346*, 2023. 9

[56] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *CVPR*, 2020. 3

[57] Wei Zhan, Liting Sun, Di Wang, Haojie Shi, Aubrey Clausse, Maximilian Naumann, Julius Kummerle, Hendrik Konigshof, Christoph Stiller, Arnaud de La Fortelle, et al. Interaction dataset: An international, adversarial and cooperative motion dataset in interactive driving scenarios with semantic maps. *arXiv preprint arXiv:1910.03088*, 2019. 3

[58] Guosheng Zhao, Xiaofeng Wang, Zheng Zhu, Xinze Chen, Guan Huang, Xiaoyi Bao, and Xingang Wang. Drivedreamer-2: Llm-enhanced world models for diverse driving video generation. *arXiv preprint arXiv:2403.06845*, 2024. 2, 4

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] See section 1
   (b) Did you describe the limitations of your work? [Yes] See section 5.4
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See section 6
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A] No theoretical results
   (b) Did you include complete proofs of all theoretical results? [N/A] No theoretical results
3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] In the supplemental material
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] In the supplemental material
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] We repeat evaluation multiple times and report the mean performance.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes] See supplemental material
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] See https://drivingdojo.github.io
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] The public release of the data has been approved and authorized by Meituan Inc.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] See Personal Identification Information Removal in Section 3.4
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Appendix

A Dataset

A.1 Overview

We will publish the DrivingDojo dataset, data format and annotation instructions, AIF benchmark, and code for the baseline method on our project page: https://drivingdojo.github.io.

Terms of use and License. Our dataset is released under the **CC BY-NC 4.0** license, allowing everyone to use it for non-commercial research purposes.

Data maintenance. The data is stored on Google Drive for global accessibility, and we will supply various links (e.g., Baidu Cloud Drive and Hugging Face) for researchers' convenience. We will maintain the data long-term and periodically verify its accessibility.

In the following, we showcase more video examples in our DrivingDojo dataset, the corresponding videos are better illustrated on our project page.

Figure 8: **Examples of rich ego-actions on the DrivingDojo dataset.**

Action completeness. We include more dataset visualizations depicting various ego-actions in Figure 8. From top to bottom, the images show the ego vehicle performing left turns, right turns, going straight, lane-changing, and making emergency brakes during the driving.

Figure 9: **Examples of multi-agent interplay on the DrivingDojo dataset.**

Multi-agent interplay. Interaction plays a crucial role in driving scenarios. It usually means that the ego vehicle has engaged with other road users, leading to changes in the behavior of either the ego vehicle or the other road users. As shown in Figure 9, we present a series of interaction examples in our dataset. In the first scenario, the car suddenly encounters another vehicle crossing its path while moving forward, prompting an abrupt braking maneuver. The second scenario portrays the car encountering an electric scooter unexpectedly crossing its path. Illustrating the third scenario, the car comes across a vehicle in front opening its door, forcing an abrupt brake. In the fourth scenario, the challenge involves encountering a bicycle approaching from the opposite direction, while the fifth scenario involves navigating around a stroller. The sixth scenario showcases encountering road construction ahead, followed by encountering a street sweeper in the seventh scenario. The

eighth scenario presents a situation where a car suddenly makes a U-turn from the opposite direction, prompting an urgent braking response from our vehicle. Subsequent scenarios involve interactions with pedestrians. These diverse interaction scenarios provide a crucial foundation for studying the interaction of real-world simulators.

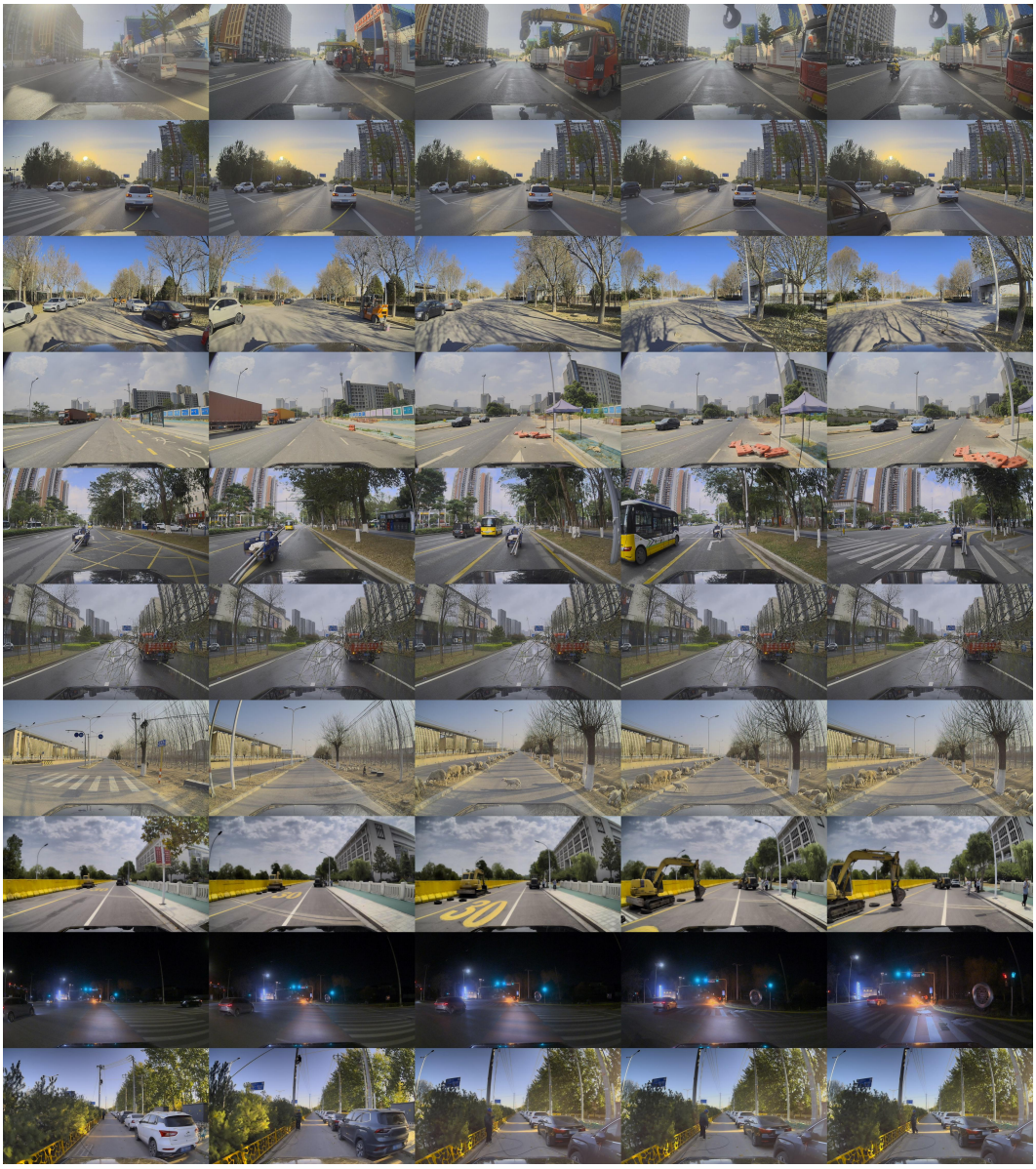


Figure 10: **Examples of diverse open-world objects on the DrivingDojo dataset.**

Open-world knowledge. In complex driving environments, we often encounter a wide variety of open-world situations. These scenarios can include sudden appearances of unexpected obstacles such as fallen trees, construction barriers, or abandoned vehicles. Typically belonging to the tail end of a long-tail distribution, these scenarios are rare yet crucial for ensuring safe driving. In Figure 10, we showcase a series of examples from the dataset, which fully demonstrate the richness of our dataset in capturing long-tail scenarios. From top to bottom, the examples illustrate encounters with a crane, a towing rope, construction barriers, a fallen roadblock, a vehicle transporting iron pipes, a vehicle transporting tree branches, a herd of sheep, an excavator, a bonfire, and power lines.

A.2 Curation

In this section, we provide the details of the curation procedure of each subset of DrivingDojo dataset and the descriptions of curated actions and interactions. This section supplements the details for Section 3.4 in the main paper.

Action Completeness Ego maneuvers for a car, particularly in the context of autonomous driving, refer to the actions and decisions the vehicle makes to navigate its environment safely and efficiently. Here is an exhaustive list of common ego maneuvers, and some examples in our datasets are shown in Figure 1a in the main paper:

- **Acceleration**: Increasing speed to match traffic flow.
- **Deceleration**: Gradual slowing down for stop signs, traffic lights, or traffic congestion.
- **Lane Keeping**: Maintaining the current lane.
- **Lane Changing**: Changing lanes to overtake slower vehicles or merge into traffic.
- **Turning**: Left/right or U-turns at intersections or roundabouts.
- **Stop and Move on**: Stopping/proceeding at traffic signals or stop signs.
- **Emergency brake**: Abrupt and sudden braking maneuver to avoid a collision or mitigate the impact of a potential hazard.

So, in the DrivingDojo-Action set, the videos follow different action commands, and the actions are mainly from the planning and control (PNC) signals, such as left and right turns, straight crossings, and lane changes. Each curated video clip begins with the PNC issuing a specific command and ends when the command is completed.

Multi-agent Interplay The examples of multi-agent interplay are shown in Figure 1b in the main paper. Then we describe the detailed cases of the interactions with dynamic agents.

- **Cutting in/off**: Another vehicle abruptly changes lanes and enters the path of the autonomous vehicle. Ego vehicle changes lanes and enters the path of the other vehicles.
- **Meeting**: Ego vehicle encounters other vehicles traveling in the opposite direction.
- **Blocked**: Ego vehicle is stopped by other agents, such as vehicles, motorcycles, and pedestrians.
- **Overtaking and being overtaken**: Ego vehicle attempting to pass another vehicle and being passed by another vehicle.

In the DrivingDojo-Interplay set, the core data curation strategy is to find the interaction with other agents. The interaction is determined using PNC signals and manually defined rules. The main interaction videos are from PNC dangerous interaction data. PNC conducts a deduction between the ego vehicle and obstacles. When the ego vehicle cannot avoid collision by turning the steering wheel or slowing down slightly, it is a PNC interaction case.

Open-world Knowledge Here, we select some representative and interesting examples from these rare cases and show them in Figure 1c in the main paper. Based on the provided image and the given descriptions, here are the detailed descriptions of each rare case in autonomous driving:

(a) A worker's helmet rolls on the sidewalk next to the road. (b) A soccer ball is seen flying across the road. (c) A water bucket is depicted falling onto the road. (d) Parcel boxes have fallen onto the road. (e) A dog is crossing the road. (f) A rope is floating over the road. (g) The traffic light turns red. (h) A boom barrier blocks the vehicle from moving forward.

As mentioned above, we curated DrivingDojo-Open set in which the videos are more carefully categorized and labeled with text descriptions. The sources are unusual weather, foreign objects on the road surface, floating obstacles, falling objects, taking over cases, and interactions with traffic lights and boom barriers. For curating the foreign objects/obstacles, we manually check and label them by a large number of data annotators.

Dataset Format DrivingDojo dataset provides a file named 'dataset_info.json' that stores information corresponding to each video segment, including the information shown in Table 5. The 'type' represents the major category, 'tag' represents the minor category, and 'remark' provides detailed descriptions of the reasons for hard braking and intervention.

Table 5: The explanation of the information in dataset_info.json.

Information	Detailed explanation
meta_info	weather, location, time, frame number
description	type, tag, remark
videos	the image path for each frame
camera_info	the camera intrinsic parameters and extrinsic matrix for each frame
action_info	the coordinates of the next frame's camera position in the current camera coordinate system

The following is an example directory structure for a dataset:

```
.
dataset_info.json
action_info
    062959_s20-370_1712024694.0_1712024714.0
        0023_next_frame_position_at_current_camera.txt
        0025_next_frame_position_at_current_camera.txt
        0027_next_frame_position_at_current_camera.txt
        ...
    145325_s20-190_1683790938.0_1683790958.0
        0024_next_frame_position_at_current_camera.txt
        0026_next_frame_position_at_current_camera.txt
        0028_next_frame_position_at_current_camera.txt
        ...
    ...
camera_info
    062959_s20-370_1712024694.0_1712024714.0
        0023_camera_parameters.txt
        0025_camera_parameters.txt
        0027_camera_parameters.txt
        ...
    145325_s20-190_1683790938.0_1683790958.0
        0024_camera_parameters.txt
        0026_camera_parameters.txt
        0028_camera_parameters.txt
        ...
    ...
videos
    062959_s20-370_1712024694.0_1712024714.0
        0023_CameraFpgaP0H120.jpg
        0025_CameraFpgaP0H120.jpg
        0027_CameraFpgaP0H120.jpg
        ...
    145325_s20-190_1683790938.0_1683790958.0
        0024_CameraFpgaP0H120.jpg
        0026_CameraFpgaP0H120.jpg
        0028_CameraFpgaP0H120.jpg
        ...
    ...
```

Camera info. The 'camera info' refers to the extrinsic and intrinsic matrices of each frame of a fisheye camera. The world coordinate system is chosen as the East-North-Up (ENU) coordinate system. In the camera coordinate system, the x, y, and z axes respectively point to the right, down, and forward. We normalize the world coordinate system of the first frame to the origin, which means that the translation variables in the extrinsic matrices of each frame are subtracted by the translation variables of the first frame.

Action info. The 'action info' represents the coordinates of the next frame's camera position in the current camera coordinate system. Let the transformation matrix from the camera to the world coordinate system be $\begin{pmatrix} R & T \\ 0^3 & 1 \end{pmatrix}$. The calculation method for the action info A_n of the n-th frame is shown in formula 3. The orientation of xyz axes in matrix A_n is consistent with the camera coordinate system, where the x, y, and z axes respectively point to the right, down, and forward.

$$\begin{pmatrix} A_n \\ 1 \end{pmatrix} = \begin{pmatrix} R_n & T_n \\ 0^3 & 1 \end{pmatrix}^{-1} \begin{pmatrix} T_{n+1} \\ 1 \end{pmatrix} \tag{3}$$

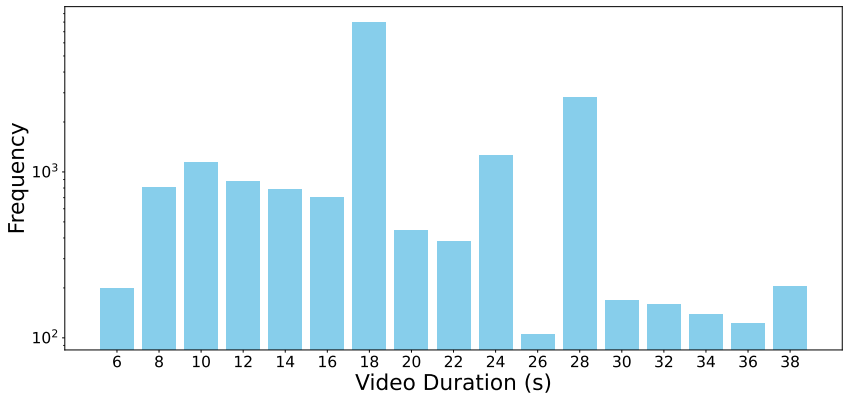


Figure 11: **Distribution of DrivingDojo video duration.**

Video info. The video is stored as a sequence of individual image frames. The distribution of video duration is shown in Figure 11, with the majority of videos lasting around 20 seconds.

B Implementation Details

B.1 Experiment Setup

During the experiment, we employed two settings for training the model. In the first setting, the focus is on visual prediction: the model predicts subsequent video content based solely on the initial frame image. In the second setting, we employ action-controlled video generation. Here, alongside the initial frame image, action information for the subsequent frames is provided to the model, enabling it to predict the ensuing video content.

Visual prediction. In this setup, we trained a high-resolution version of the model, 1024×576 resolution for 11 frames, aimed at better capturing the generation of long-tail objects. Additionally, we developed a low-resolution version of the model, 576×320 resolution for 30 frames, to simulate various vehicle behaviors and interaction events. We fine-tune all parameters of the U-Net model.

Action instruction following. In this setup, we trained model using 576×320 resolution for 30 frames. We fine-tune all U-Net parameters together with a new action encoder.

B.2 Training

We initialize the model using the SVD-XT checkpoint. Following SVD, our model is trained with the EDM framework [31]. During training, we set the fps to 5 and the motion_bucket_id to 127. We utilize the AdamW optimizer [34] with a learning rate of 1×10^{-5}. The training process is conducted on 16 NVIDIA A100 (80G) GPUs with 32 batch size for 50K iterations. To allow classifier-free guidance [24], we drop out action feature with a ratio of 20%.

B.3 Evaluation

During inference, we generate videos using the DDIM sampler for 25 steps.

Visual Quality. To evaluate the quality of the generated video, we utilize FID (Frechet Inception Distance) [22] and FVD (Frechet Video Distance) [45] as the main metrics. For FID calculation on videos, we randomly select 5,000 frames for evaluation. Additionally, for FVD calculation, we generate 256 videos for evaluation. The results are the average of 10 calculations. We use the official UCF FVD evaluation code[2].

Action instruction following (AIF). For each generated video with action instructions, we estimate the camera poses for each frame in the video, align the scale of the estimated trajectory with the instruction trajectory, and compare the vehicle motion in each frame with the respective action instructions. We estimate the ego trajectories in generated videos with the offline visual structure-from-motion (SfM) implementation COLMAP [40, 41]. For videos generated based on initial images from DrivingDojo, we fix the camera intrinsic parameters as the ground truth values for videos from DrivingDojo. For videos generated from initial images with unknown camera intrinsics (e.g. images from OpenDV-2K), we estimate the camera intrinsics together with the camera extrinsics of images. We perform feature point extraction, feature point matching, and sparse scene reconstruction with the official implementation of COLMAP[3] to estimate the poses of cameras. In our experiments, we generate videos in 30 frames and align the scale of estimated trajectories with the instruction trajectories based on the motions in the first $N = 10$ frames. We report the mean value of the absolute error between estimated motions and instruction motions in all video frames.

C Additional Experiments

Table 6: **Action instruction following under zero-shot evaluation.** * denotes the model is applied zero-shot to this dataset without fine-tuning.

Training set	Test set	FID	FVD	$E_x^{\text{AIF}}(\downarrow)$	$E_y^{\text{AIF}}(\downarrow)$
DrivingDojo	OpenDV-2K*	**24.27**	**442.67**	**0.175m**	**0.121m**
ONCE	OpenDV-2K*	28.37	473.59	0.295m	0.308m
nuScenes	OpenDV-2K*	37.90	794.36	0.488m	0.465m

C.1 Action Instruction Following

As shown in Table 6, we compared the performance of models trained on different datasets and their zero-shot generalization performance on new datasets. The results indicate that models trained on our dataset exhibit higher generation quality and significantly improved action-following ability. Especially, we noticed that richer driving actions in the autonomous driving datasets lead to significantly better AIF performance of models trained on them. According to Figure 3a, videos in DrivingDojo averagely contain far richer driving actions compared to ONCE or nuScenes. This leads to the far

[2] https://github.com/SongweiGe/TATS/
[3] https://github.com/colmap/colmap

better AIF performance of model trained on DrivingDojo compared to those trained on ONCE or nuScenes. we observed that the model trained on the ONCE dataset will always generate videos in which the vehicle moves in a straight line, even with action instructions to turn left/right or change lanes. This leads to its especially poor AIF performance in the lateral direction (E_y^{AIF}). We speculate that this is because the driving action of making turns or changing lanes is very rare in the ONCE dataset, as shown in Figure 3a, which results in the lack of ability of the model trained on the ONCE dataset to follow the lateral motion instructions. Moreover, the even more lacking driving actions in the nuScenes dataset lead to a worse AIF performance of the world model.

D Visualizations

In this section, we show the model generation demos trained on the DrivingDojo dataset. As shown in Figure 12, our model can generate high-resolution, complex driving scenarios.

Figure 12: **Examples of high-resolution and complex scenarios generation.** For illustration purposes, we represent each video example with a single frame.

D.1 Diverse Actions

As shown in Figure 13, we demonstrate how actions control the generation of different futures, such as moving forward, backward, and stopping.

D.2 Dynamic Interaction

As shown in Figure 14, we observe that choosing different actions can influence the behavior of other vehicles, resulting in different responses from the world model. For instance, in the first example, if we choose to proceed slowly, the vehicle on the left decides to stop and yield. Conversely, if our vehicle stops, the left vehicle perceives an obstruction and slightly reverses to make way. In the second example, when we choose to brake, the right vehicle quickly cuts in front of us, while if we choose to proceed straight, the right vehicle waits in place.

D.3 Open-world Knowledge

As illustrated in Figure 15, we demonstrate the model's ability to simulate various open-world objects, such as encountering construction zones, rare objects like ladders or balloons on the road, and simulating a puddle of water on the ground.

D.4 AIF Visualization

We showcase examples of estimated trajectories from generated videos in Figure 16. In each frame, the red dot represents the current estimated camera pose and the black dots represent the camera poses in past frames.

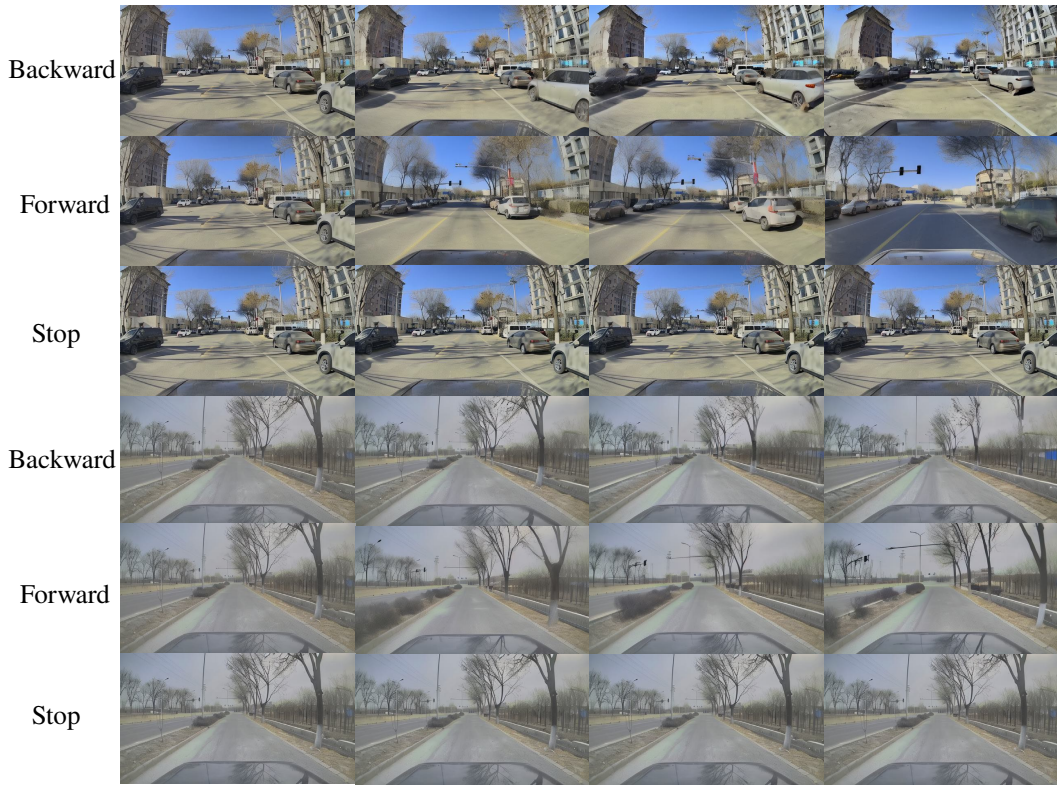


Figure 13: **Examples of diverse action-based video generation.**

Figure 14: **Simulation of interaction with other agents.**

Figure 15: **Simulation of various open-world objects on the road.**

Figure 16: **Examples of ego trajectories estimated based on the generated videos.**

E License of Assets

We report licenses of all artifacts used in this work in this section.

Model We use the pre-trained stable video diffusion [2] checkpoints from the huggingface platform. These checkpoints are released under the stable video diffusion non-commercial community license agreement[4] for research purpose.

Our Dataset Our dataset is collected and curated by the autonomous driving team of Meituan Inc. The road test and data collection procedures conform to privacy and security requirements of local authorities. The authors have obtained the permission for publicly releasing this dataset from both the management team and the company legal team. All personal identifiable information has been removed by both algorithm and subsequent manual inspection. We release the dataset under the CC BY-NC 4.0 license.

Other Datasets We use other public datasets in this work including nuScenes [6], ONCE [36] and OpenDV-2k [53]. The nuScenes [6] dataset is released under the CC BY-NC-SA 4.0 license with Dataset Terms [5]. The ONCE dataset is also released under the CC BY-NC-SA 4.0 license with Dataset Terms [6]. The OpenDV-2K dataset is constructed from publicly licensed datasets and youtube videos that the authors claimed to support academic usage licenses.

[4] https://huggingface.co/stabilityai/stable-video-diffusion-img2vid-xt/blob/main/LICENSE

[5] https://www.nuscenes.org/terms-of-use

[6] https://once-for-auto-driving.github.io/terms_of_use.html

F Datasheet

F.1 Motivation

- **For what purpose was the dataset created?** Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

  We introduce DrivingDojo, the first dataset tailor-made for training interactive world models with complex driving dynamics. Our dataset features video clips with a complete set of driving maneuvers, diverse multi-agent interplay, and rich open-world driving knowledge, laying a stepping stone for future world model development.

- **Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?**

  Institute of Automation, Chinese Academy of Sciences, University of Chinese Academy of Sciences, Meituan Inc., and Centre for Artificial Intelligence and Robotics, HKISI_CAS.

- **Who funded the creation of the dataset?**If there is an associated grant, please provide the name of the grantor and the grant name and number.

  This work was supported in part by the National Key R&D Program of China (No. 2022ZD0116500), the National Natural Science Foundation of China (No. U21B2042, No. 62320106010), and in part by the 2035 Innovation Program of CAS, and the InnoHK program.

- **Any other comments?**

  No.

F.2 Composition

- **What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?** Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

  The instances of our DrivingDojo dataset are videos with ego actions and DrivingDojo-Open subset is also with text descriptions for each scene.

- **How many instances are there in total (of each type, if appropriate)?**

  There are 18.2k videos for the whole DrivingDojo dataset, in which the DrivingDojo-Action subset has 7.9k videos, DrivingDojo-Interplay subset has 6.4k videos, and DrivingDojo-Open has 3.9k videos.

- **Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?** If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

  The DrivingDojo dataset is sampled from a data pool of around 7500 hours. About representativeness, please refer to the Data Curation section (Sec. 3.4 and Sec. A.2).

- **What data does each instance consist of?** "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

  DrivingDojo-Action and DrivingDojo-Interplay subsets consist of videos and ego actions, and DrivingDojo-Open subset consists of videos, ego actions, and text descriptions.

- **Is there a label or target associated with each instance?** If so, please provide a description.

  Yes. There is a text description label for each instance in DrivingDojo-Open subset, which describes the open-world knowledge in the scene.

- **Is any information missing from individual instances?** If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

  No.

- **Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?** If so, please describe how these relationships are made explicit.

  No.

- **Are there recommended data splits (e.g., training, development/validation, testing)?** If so, please provide a description of these splits, explaining the rationale behind them.

  No. There is no need for the validation/testing split. We care about zero-shot generation.

- **Are there any errors, sources of noise, or redundancies in the dataset?** If so, please provide a description.

  Yes. The sources of noise may be inaccurate poses, camera noises, and human-sourced text noises.

- **Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?** If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

  Yes. the DrivingDojo dataset is self-contained.

- **Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)?** If so, please provide a description.

  No.

- **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?** If so, please describe why.

  No.

F.3 Collection Process

- **How was the data associated with each instance acquired?** Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

  DrivingDojo dataset is collected using the platform of Meituan's autonomous delivery vehicles.

- **What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)?** How were these mechanisms or procedures validated?

  DrivingDojo dataset is collected using the platform of Meituan's autonomous delivery vehicles with fish-eye RGB cameras. The cameras are calibrated. The text labels are manually validated.

- **If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?**

  Please refer to the Data Curation section (Sec. 3.4 and Sec. A.2).

- **Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?**

  The data collectors are employed by Meituan Inc. and are paid by Meituan Inc.

- **Over what timeframe was the data collected?** Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

  The data are collected from May 2022 to May 2024. This timeframe matches the creation timeframe of the data associated with the instances.

- **Were any ethical review processes conducted (e.g., by an institutional review board)?** If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

  Yes. The ethical review is conducted before the release by Meituan Inc.

F.4 Preprocessing/cleaning/labeling

- **Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?** If so, please provide a description. If not, you may skip the remaining questions in this section.

  No.

- **Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?** If so, please provide a link or other access point to the "raw" data.

  N/A.

- **Is the software that was used to preprocess/clean/label the data available?** If so, please provide a link or other access point.

  N/A.

- **Any other comments?**

  No.

F.5 Uses

- **Has the dataset been used for any tasks already?** If so, please provide a description.

  The DrivingDojo dataset has been used for driving world models. The experiments are in Sec. 5 in the main paper and Sec. C in the appendix.

- **Is there a repository that links to any or all papers or systems that use the dataset?** If so, please provide a link or other access point.

  Yes. Please refer to the webset:https://drivingdojo.github.io.

- **What (other) tasks could the dataset be used for?**

  The DrivingDojo dataset could be used for training end-to-end autonomous driving models.

- **Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?** For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other risks or harms (e.g., legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?

  No.

- **Are there tasks for which the dataset should not be used?** If so, please provide a description.

Due to the known biases of the dataset, under no circumstance should any models be put into production using the dataset as is. It is neither safe nor responsible. As it stands, the dataset should be solely used for research purposes in its uncurated state.

- **Any other comments?**

  No.

F.6 Distribution

- **Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?** If so, please provide a description.

  Yes, the dataset will be open-source.

- **How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?** Does the dataset have a digital object identifier (DOI)?

  On our website: https://drivingdojo.github.io.

- **When will the dataset be distributed?**

  We have released some demos on the project page. The whole DrivingDojo dataset will be public in the camera-ready version.

- **Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?** If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

  DrivingDojo dataset will be distributed under the CC BY-NC 4.0 license.

- **Have any third parties imposed IP-based or other restrictions on the data associated with the instances?** If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

  No.

- **Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?** If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

  No.

- **Any other comments?**

  No.

F.7 Maintenance

- **Who will be supporting/hosting/maintaining the dataset?**

  Institute of Automation, Chinese Academy of Sciences and Meituan Inc. will maintain DrivingDojo dataset.

- **How can the owner/curator/manager of the dataset be contacted (e.g., email address)?**

  The main maintainer Yuqi Wang's e-mail: wangyuqi2020@ia.ac.cn.

- **Is there an erratum?** If so, please provide a link or other access point.

  There is no erratum for our initial release. Errata will be documented as future releases on the dataset website.

- **Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?** If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (e.g., mailing list, GitHub)?

Yes. We will update the DrivingDojo dataset. Especially, we will adapt to end-to-end autonomous driving tasks in the future. The update will be released on the website and GitHub.

- **If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were the individuals in question told that their data would be retained for a fixed period of time and then deleted)?** If so, please describe these limits and explain how they will be enforced.

  N/A.

- **Will older versions of the dataset continue to be supported/hosted/maintained?** If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.

  Yes. We will maintain the older versions of the dataset on the website and GitHub.

- **If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?** If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.

  Yes. The dataset is open source under the CC BY-NC 4.0 license. So it is open to other contributors.

- **Any other comments?**

  No.