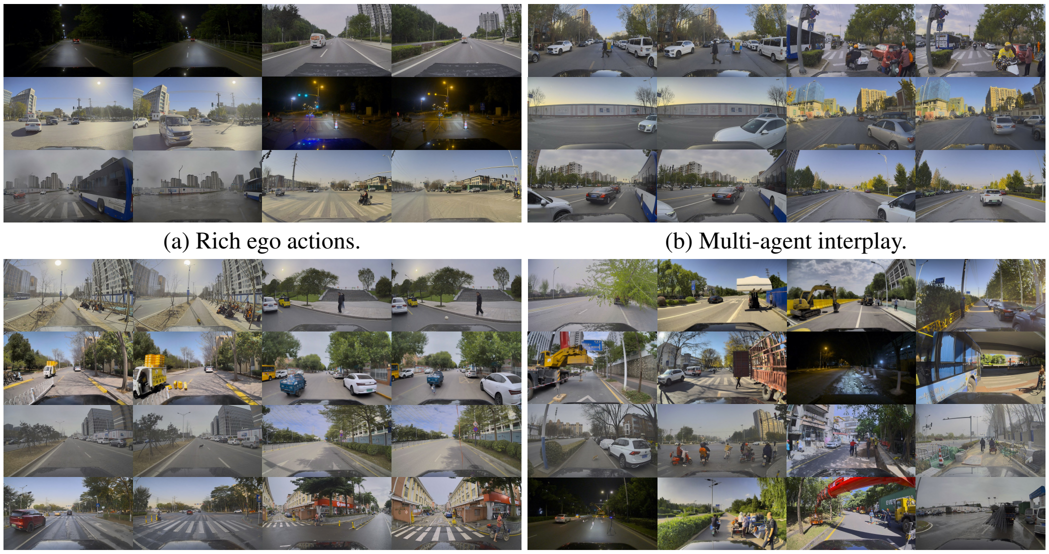
DrivingDojo Dataset: Advancing Interactive and Knowledge-Enriched Driving World Model

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

38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks.

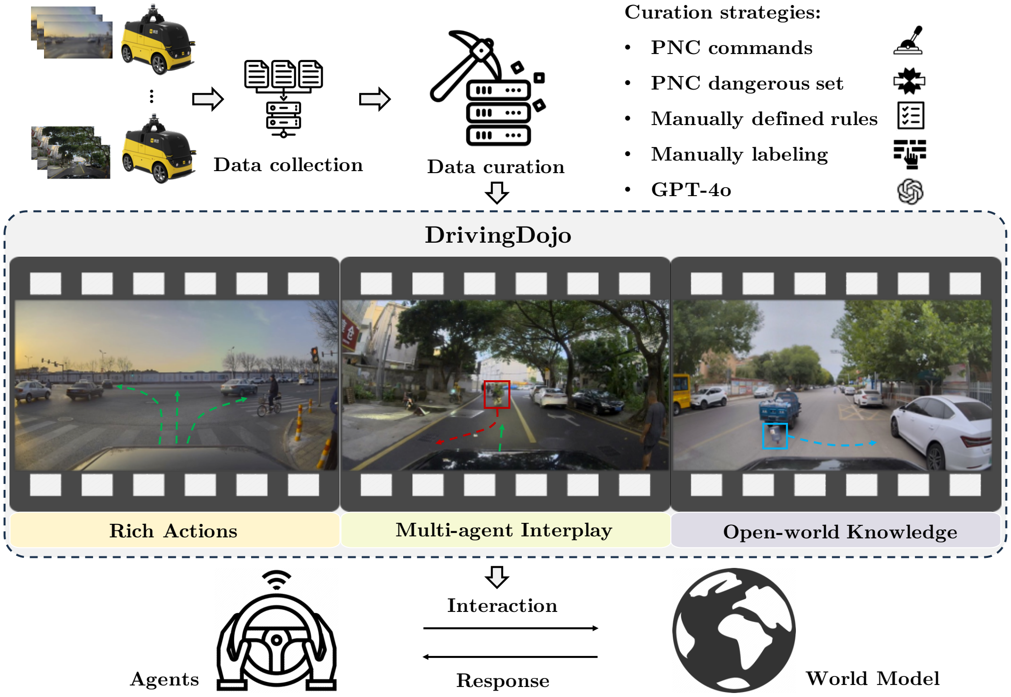


Figure 2: **Enhancing interactive and knowledge-enriched learning of world models.** Data plays a crucial role in modeling the world. DrivingDojo is a large-scale video dataset curated from millions of daily collected videos, designed to investigate real-world visual interactions. DrivingDojo features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

1 Introduction

World models [17, 20, 33, 21] have gained increasing attention due to their ability to model complex real-world physical dynamics. They also hold potential as general-purpose simulators, capable of predicting future states in response to diverse action instructions. Facilitated by advancements in video generation techniques [53, 24, 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 [47, 19, 56]. Generative world models, in particular, hold significant promise as real-world simulators and have garnered extensive research in the field of autonomous driving [28, 48, 30, 49, 54, 60, 13].

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 [45], and ONCE [37], 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 potential of world models in action instruction following by including rich longitudinal maneuvers like acceleration, emergency braking and stop-and-go as well as lateral ones like U-turn, overtaking, and lane change. Besides, we explicitly curate the dataset to include a large volume of trajectories

Table 1: **A comparison of driving datasets for world model**. This comparison emphasizes the diversity of the video content, placing less focus on annotations or sensor data. * 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 [45]	1k	11	✓			
OpenDV-2k [54]	2k	2059		✓		
nuPlan [7]	-	1500	✓	✓	✓	
DrivingDojo (Ours)	18k	150*	✓	✓	✓	✓

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

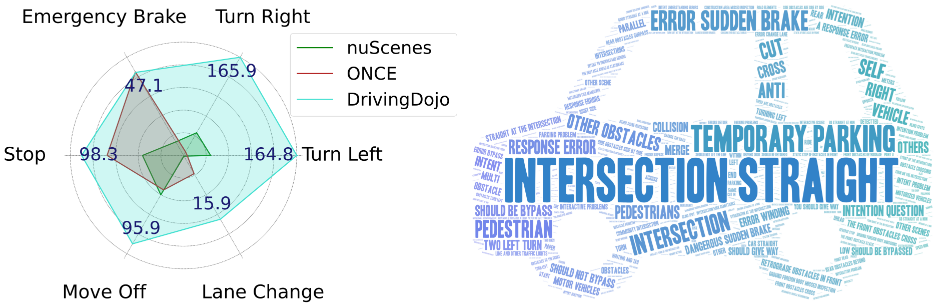
Datasets for perception. The driving dataset has played a crucial role in advancing computer vision in recent years, aiming to achieve comprehensive perception and understanding surrounding the ego vehicle. Initially, perception in autonomous driving relied on 2D image-based perception. Datasets like Cityscapes [10], Mapillary Vistas [39], and BDD100k [58] provided instance-level masks for learning tasks. With the integration of LiDAR sensors and advancements in 3D perception, datasets like KITTI [14], nuScenes [6], and Waymo [45] have emerged as standard benchmarks for various 3D perception tasks. Additionally, datasets like ONCE [37], Argoverse [8, 50], and others [29, 15, 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 [26], and nuScenes Prediction [6] challenge. Additionally, the Interaction dataset [59] 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 [17, 33] 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 [18, 20, 9, 52, 44, 43, 21]. 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 [51] 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 [55] 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 [16, 61] and can be categorized into two main branches. The first branch explores agent policies in virtual simulators. MILE [27] employed imitation learning to jointly learn the dynamics model and driving behavior in CARLA [11]. Think2Drive [34] 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 [28] introduced a generative world model for autonomous driving, capable of simulating realistic driving videos from inputs like images, texts, and actions. DriveDreamer [48] emphasized scenario generation, leveraging HD maps and 3D boxes to enhance video quality. Drive-WM [49] 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 [30] constructed a general world model based on MLLM and diffusion models, using vision-action pairs to auto-regressively predict current frame control signals. DriveDreamer2 [60] 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 [54] 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 dataset constitution.** The dataset is organized into three subsets: DrivingDojo-Action, DrivingDojo-Interplay, and DrivingDojo-Open, to support research on specific tasks.

Dataset	Videos	Type	Camera	Ego Trajectory	Text Description
DrivingDojo	17.8k	total	✓	✓	✓
DrivingDojo-Action	7.9k	rich ego-actions	✓	✓	
DrivingDojo-Interplay	6.2k	multi-agent interplay	✓	✓	
DrivingDojo-Open	3.7k	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 [37] 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 [45], or are constructed from large-scale internet videos that lack proper curation and balancing, like OpenDV-2K [54].

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 [37] 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 3.7k 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.

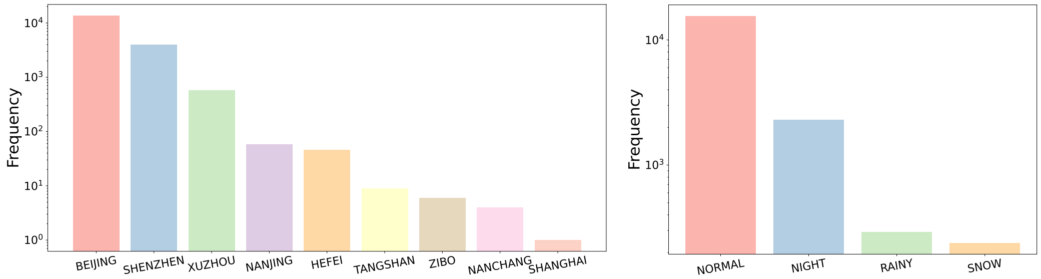


Figure 4: **Descriptive statistics of the DrivingDojo dataset.** The dataset was collected from various regions across China, including nighttime and rainy/snowy conditions.

3.4 Data Curation and Statistics

Dataset statistics. The DrivingDojo dataset contains around 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 [31] 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 [40].

4.3 Evaluation Metrics

Visual quality. To evaluate the quality of the generated video, we utilize FID (Frechet Inception Distance) [23] and FVD (Frechet Video Distance) [46] 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 [41, 42]: $\{\widetilde{A}_t, ..., \widetilde{A}_{t+k}\} = \text{SfM}(\{I_t, ..., I_{t+k}\})$, 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	OpenDV-2K	OpenDV-2K	18.27	321.05
SVD	-	OpenDV-2K	24.17	580.94
SVD	nuScenes†	OpenDV-2K	21.05	395.04
SVD	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 [54] 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.

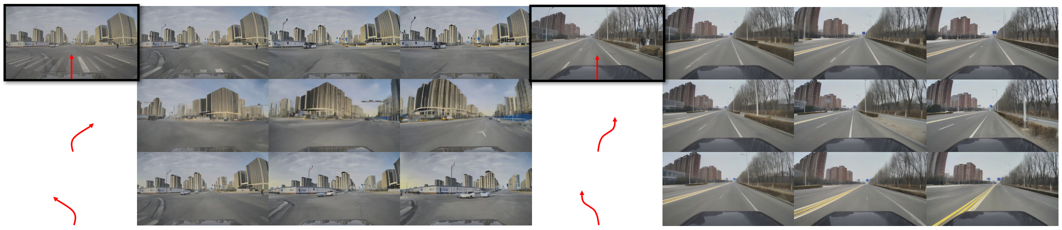


Figure 5: **Predicting multiple futures based on different actions.** Left: going straight, turning left, and turning right at a crossing; Right: changing to the left lane, staying in the current lane, and changing to the right lane.

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 10 cm in the lateral or longitudinal directions. In row 3, 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 [54] 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**. GT refers to using real images to test the accuracy of the reconstructed trajectory. * 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(GT)	-	-	0.036m	0.019m
In-Domain	DrivingDojo	37.07	658.72	0.100m	0.062m
Out-of-Domain	DrivingDojo	38.30	716.44	0.173m	0.110m
Out-of-Domain	OpenDV-2K*	24.27	442.67	0.238m	0.136m

Table 5: **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**	**0.238m**	**0.136m**
ONCE	OpenDV-2K*	28.37	473.59	0.255m	0.23d9m
nuScenes	OpenDV-2K*	37.90	794.36	0.387m	0.254m

Zero-shot evaluation. As shown in Table 5, 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 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.

AIF visualization. We showcase examples of estimated trajectories from generated videos in Figure 6. In each frame, the red dot represents the current estimated camera pose and the black dots represent the camera poses in past frames.

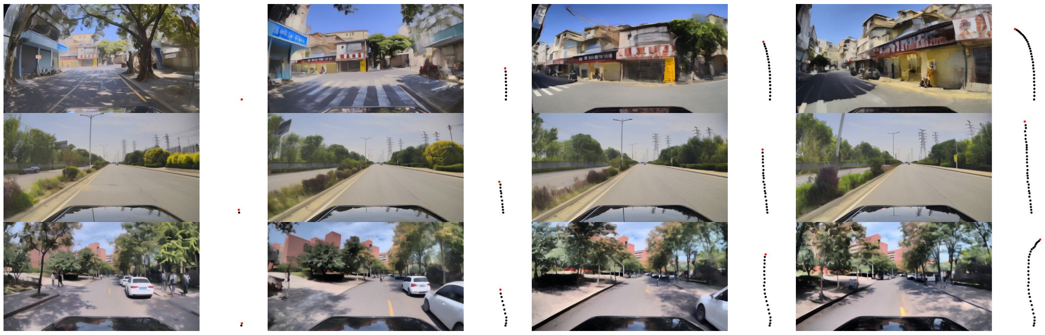
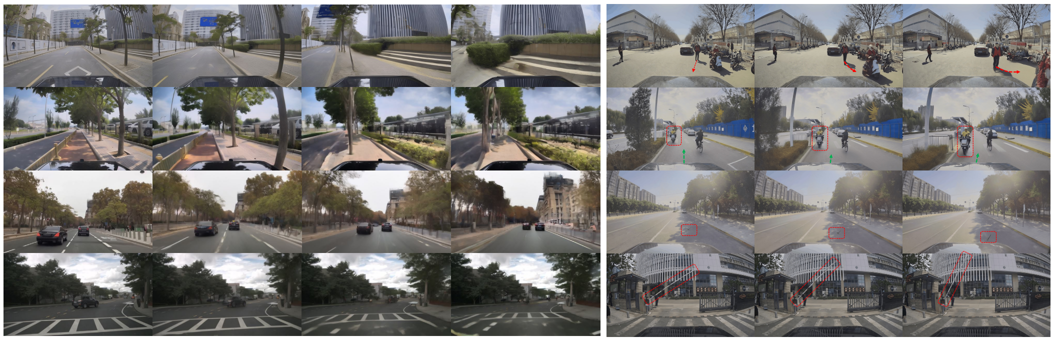


Figure 6: **Examples of ego trajectories estimated based on the generated videos.**

5.3 Real-world Simulation

Action generalization. Our model demonstrates robust generalization capabilities in two key aspects. As illustrated in Figure 7a, 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 [54] 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.



(a) Action generalization

(b) Interaction simulation

Figure 7: **Qualitative examples of our model's capability.**

Dynamic agents. We showcase our model's ability to simulate interactions with dynamic agents in Figure 7b. The results indicate that the model can provide reasonable responses based on our actions. 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 7b, 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 dataset currently comprises only single-camera videos. Our primary focus is to maximize video diversity, which has led us to reduce the number of sensors used, enabling us to capture a wider range of scenes. Additionally, this paper primarily explores the value of the dataset, treating the model aspect as a baseline without any specialized 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 8, 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.

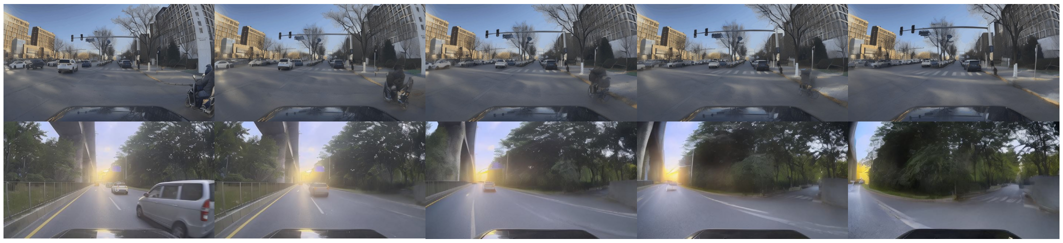


Figure 8: **Examples of hallucination**. Top: object suddenly disappears. bottom: a non-existed road.

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, 57, 22] and faster generation [38, 36] are left for future research.

Driving policy. The long-tail cases in our dataset are valuable for driving policy research. While this work focuses on visual prediction in world models, future studies can investigate how this data improves driving policy.

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.

Acknowledgments and Disclosure of Funding

We would like to thank Meituan's autonomous vehicle team for providing the data and computing resources. 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, and in part by the Meituan Collaborative Research Project.

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. 10

[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, 9, 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, 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] Shenyuan Gao, Jiazhi Yang, Li Chen, Kashyap Chitta, Yihang Qiu, Andreas Geiger, Jun Zhang, and Hongyang Li. Vista: A generalizable driving world model with high fidelity and versatile controllability. *arXiv preprint arXiv:2405.17398*, 2024. 2

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

[15] 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

[16] 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

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

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

[19] 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

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

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

[22] 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. 10

[23] 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, 23

[24] 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

[25] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 23

[26] 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

[27] 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

[28] 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

[29] 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

[30] 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

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

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

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

[34] 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

[35] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2018. 23

[36] 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. 10

[37] 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

[38] Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *CVPR*, 2023. 10

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

[40] 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

[41] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 7, 23

[42] 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, 23

[43] 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

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

[45] 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

[46] 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, 23

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

[48] 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

[49] 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

[50] 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

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

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

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

[54] 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, 9, 25

[55] 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

[56] Sherry Yang, Jacob C Walker, Jack Parker-Holder, Yilun Du, Jake Bruce, Andre Barreto, Pieter Abbeel, and Dale Schuurmans. Position: Video as the new language for real-world decision making. In *ICML*, 2024. 2

[57] 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. 10

[58] 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

[59] 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

[60] 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

[61] Zheng Zhu, Xiaofeng Wang, Wangbo Zhao, Chen Min, Nianchen Deng, Min Dou, Yuqi Wang, Botian Shi, Kai Wang, Chi Zhang, et al. Is sora a world simulator? a comprehensive survey on general world models and beyond. *arXiv preprint arXiv:2405.03520*, 2024. 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]