

# 1 **Appendix**

2 In the Appendix, we first outline related works in Sec. A. We then demonstrate implementation details  
3 for data and models in Sec. B to supplement Sec. 2 in the main paper. Additional results are included  
4 in Sec. C to supplement Sec. 3 in the main paper. We discuss the limitations and broader impact  
5 of our work in Sec. D, and list the license of all assets in Sec. E. We also include some qualitative  
6 samples in the submitted demo\_videos folder.

## 7 **A Related Work**

### 8 **A.1 World Model**

9 World models are considered as the abstraction of the open world, and having this kind of common  
10 sense greatly helps to learn new skills effectively, thus leading to high-level intelligence [1]. Under  
11 the definition of world models following the Dreamer series [2, 3, 4], they represent the transition  
12 of environmental dynamics, taking the past states or observations and policy’s actions as input,  
13 and generating the next (latent) state together with an estimation of the reward. They also feature  
14 long-term prediction with continuous rollouts [5].

15 Abundant literature has explored world models in traditional policy learning tasks, especially utilizing  
16 the look-ahead property to learn efficient representations [6], conduct sampling-based planning [7, 8,  
17 9], and enable model-based reinforcement learning [5, 2, 3, 4].

18 Taking a step further, researchers in applications have successfully employed world models in  
19 simulated games [10, 11, 12, 2, 3, 4], navigation [13, 14], and robotics [15, 16, 17, 18]. However, to  
20 learn and apply a world model requires extensive exploration and interaction with the environment,  
21 leading to the above advancements mostly being developed in simulation or constrained environments.  
22 It is infeasible to obtain diverse hazardous driving movements in the real world [19, 20]. In this work,  
23 for the first time, we address this challenge by leveraging heterogeneous data and transferring rewards  
24 learned from simulation to diverse real-world scenarios.

### 25 **A.2 Predictive Model for Driving Scenes**

26 Driving Scenes are significantly unstructured, dynamic, and complex, compared to standard policy  
27 learning environments such as Atari [21], DM Control [22], ViZDoom [23], *etc.* In order to effectively  
28 encode observations and facilitate restoring future environments, a wide span of representations have  
29 been explored to build the world state, including the bird’s-eye-view (BEV) representation [24, 25, 26],  
30 point clouds [27, 28, 29], 3D occupancy [30, 31], and images [32, 33, 34, 35]. Meanwhile, these  
31 works mainly focus on public driving datasets, which are still limited in scales to achieve strong  
32 generalization ability. Inspired by the rapid growth of visual generative models [36, 37] and the  
33 increased data volume captured by cameras with low costs [38, 39, 40], recent world models that  
34 imagine future states in image sequence (*i.e.*, video) yield encouraging results in visual fidelity and  
35 generalization [38, 41, 35].

36 Unfortunately, prior methods still struggle to fulfill the mission of faithful simulation. Due to  
37 the insufficient learning of scenario dynamics, their imagination quality significantly degrades in  
38 challenging cases and long-horizon predictions [38, 41]. They also fall short in simulating negative  
39 consequences, such as car crashes, in response to bad ego actions, since they are mainly established  
40 on human driving logs, which are biased toward safe executions. Furthermore, the core problem for  
41 driving world models, how to deduce the reward for a given action and apply the world model for real-  
42 world driving problems, is largely understudied. In particular, with high dimensional observations and  
43 complex relationships between agents and the environment, specifying rewards for open-world driving  
44 scenarios is challenging compared to goal-conditioned reward specifications [42, 8, 13]. Among  
45 the previous works, Wang *et al.* [43] propose to construct rule-based rewards with off-the-shelf 3D  
46 perception models [44, 45], yet these models are sensitive to sensor configurations like camera poses  
47 thus hard to generalize [46]. Uncertainty-based rewards in Vista [41] struggle to consider specific  
48 types of behaviors such as off-route actions. Our work meticulously investigates these challenges to  
49 facilitate planning and simulation.

### 50 A.3 Video Generation

51 In recent years, deep generative models have made remarkable strides in both image generation [47,  
52 36] and video generation [48, 37, 49, 50]. Recent studies [51, 52] introduce the diffusion transformer  
53 architecture [53] to video generation and achieve impressive spatiotemporal consistency. However,  
54 existing video generation models trained with large-scale web data are not directly applicable as  
55 driving world models due to their imperfect prediction of driving scenarios and lack of action  
56 controllability [39]. We bridge the gap with novel designed model structures and training protocols.

## 57 B Implementation Details

### 58 B.1 Dataset

59 Our guiding observation is that each data corpus has distinct characteristics and limitations in terms  
60 of scenario diversity, planning labels feasibility, and the degree of danger, as depicted in Main  
61 Fig. 1(a). Based on that, we propose compiling our training data from diverse sources to integrate  
62 their complementary features to cover a wide scope of scenarios and ego actions. We specify each  
63 type of data source as follows.

64 **Universal Driving Videos.** Building a world model that generalizes to arbitrary scenarios requires  
65 learning from massive data with a wide coverage [15, 39, 40]. Therefore, we leverage the OpenDV  
66 dataset [39], which is the largest public driving video dataset, to pillar the scenario generalization of  
67 our world model. OpenDV dataset includes 1700 hours of uncalibrated front-view driving videos  
68 captured worldwide with a wide coverage of scenarios and camera configurations. The uncalibrated  
69 nature of this dataset allows the learned model to seamlessly adapt to new camera settings. We  
70 pseudo-labeled the dataset with high-level driving commands, including “Turning left”, “Moving  
71 forward”, and “Turning right”, by estimating the flow via the OpenCV toolkit [54]. During  
72 training, we assign a high sampling rate ( $5\times$ ) to video sequences with turning actions based on the  
73 driving command, as these cases are generally more challenging to learn than the forward movement.  
74 As a result, we collect 4M video clips from OpenCV datasets.

75 **Expert Driving Data.** Despite the large data volume and high diversity of online driving videos, these  
76 videos do not provide detailed annotations for ego actions, *e.g.*, ego trajectories, which are critical for  
77 learning world models with required action conditions [5]. The absence of such action annotations  
78 calls for the need to incorporate expert driving datasets that are rigorously curated and labeled.  
79 Therefore, we include a public driving dataset NAVSIM [55] into our compilation. We intentionally  
80 exclude commonly used nuScenes [56] and Waymo [57] datasets from training, and leverage them  
81 for held-out evaluation. Specifically, 85K data samples from navtrain split of NAVSIM [55] are  
82 included in training.

83 **Explorable simulated data.** Both online driving videos and expert driving datasets are produced  
84 by human drivers. The lack of suboptimal data would hinder the world model’s ability to emulate  
85 non-expert behaviors and corresponding outcomes, *e.g.*, collisions. We randomly sample from 220  
86 predefined routes in the Bench2Drive benchmark [58], varying the weather and time of day to enhance  
87 scenario diversity. We deploy two agents to explore the simulated environment while collecting data:  
88 One uses a well-established driving policy, PDM-Lite [59], to collect data from successful executions.  
89 Another agent for collecting non-expert data is implemented by rule-based explorations to cover  
90 a larger action space. This agent randomly samples a control configuration for steering angle and  
91 throttle and a behavior pattern from a predetermined set to execute. The total number of successful  
92 and hazardous execution cases is 88K, with each type accounting for roughly half the amount.

### 93 B.2 Model and Training

94 **ReSim World Model.** The architecture of ReSim is adapted from CogVideoX [51], consisting of a  
95 2B diffusion transformer (DiT) as denoising backbone, a T5 encoder [60] for language encoding, a 3D  
96 Causal VAE that compresses raw videos into a compact latent space. Alongside language conditions  
97 for high-level driving command, we additionally devise a lightweight trajectory encoder, composed of  
98 two attention blocks and a linear head, to integrate the action condition into the DiT input. The overall  
99 architecture is depicted in Fig. S.1 with some of our designs highlighted. Besides our key innovations

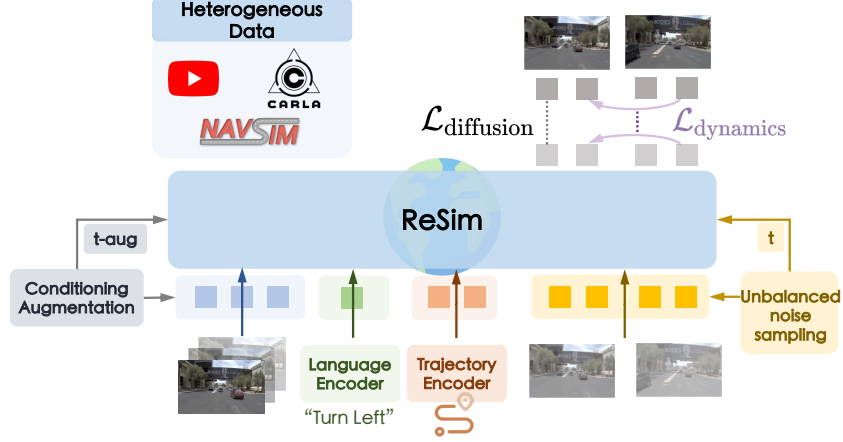


Figure S.1: **Overview of ReSim world model.** Learning from heterogeneous data compilation (Main Sec. 2.1), ReSim features designs specified in Main Sec. 2.2.

Table S.1: **Optimization configurations for different learning stages.** Traj. Enc.: Trajectory Encoder, Res.: Resolution, BS: Batch Size, LR: Learning Rate.

Stages	DiT	LoRA	Traj. Enc.	Dataset	Res.	BS	LR	Steps
1)	Trainable	-	-	OpenDV	512×896	80	$1e^{-5}$	20K
2)	Frozen	Trainable	Trainable	OpenDV, NAVSIM, CARLA	256×448	160	$5e^{-5}$	80K
3)	Trainable	Trainable	Trainable	OpenDV, NAVSIM, CARLA	512×896	80	$5e^{-5}$	50K

stated in Main Sec. 2, we also apply a conditioning augmentation strategy following [61, 12] to corrupt the video latent of historical observations to mitigate the error accumulation issue of long-term rollout. Similar to [12], the diffusion timesteps for historical context (t-aug) and future prediction (t) are separately sampled during training, while t-aug is always set to 0 during inference. This strategy improves the robustness of ReSim for multi-round prediction.

To enable classifier-free guidance for sampling [62], we randomly drop the textual command with a probability of  $p = 0.5$ . Similarly, we also drop the conditional ego trajectory at  $p = 0.5$  for NAVSIM samples. However, we retain the ego trajectory for all CARLA samples without dropout, since their abnormal and hazardous behaviors cannot be accurately inferred from historical observations only and require explicit trajectory as guidance. Moreover, exposing the model to unconditioned hazardous behaviors could interfere with the learning of expert patterns from NAVSIM. Detailed learning configurations for different stages are included in Tab. S.1. All training stages are conducted on 40 A100 GPUs, and the total training duration is around 14 days.

**Video2Reward Model.** Video2Reward model consists of a pretrained DINOv2 [63] as backbone, and a prediction head that outputs a scalar reward. For each video sequence, all video frames are first processed separately via the image-based DINOv2 backbone. All image features are then passed to the prediction head, which aggregates all features via two consecutive spatial-temporal attention blocks and further predicts a scalar reward via an MLP.

Learning from our collected CARLA data only, Video2Reward model is supervised by the Infraction Score recorded from the CARLA simulator for each sample, which is a comprehensive evaluation of the ego driving performance [64] and penalizes behaviors such as collisions, traffic light violations, off-road deviations, and unreasonable low speed. It is trained for 20 epochs on a random subset of 35K samples from our CARLA data. We use the AdamW optimizer [65] with a learning rate of  $1 \times 10^{-3}$ . All video sequences are resized to  $224 \times 224$  as input to this model.

**Inverse Dynamics Model.** Inverse dynamics model (IDM) estimates the ego trajectory from a video clip [66, 41]. Throughout our experiments, there are two parts that require the use of IDM, *i.e.*, the *Trajectory Difference* evaluation of expert action controllability (Main Sec. 3.1) and the application of video prediction-based policy (Main Sec. 3.2). These two IDMs are trained separately on different



Figure S.2: **Example of human evaluation.** Participants are presented with synthesized videos of three anonymous candidate models. The order of different models’ generations is shuffled for each testing scenario.

128 datasets, yet share the same architecture with a visual odometry backbone from XVO [67] and a  
 129 lightweight attention head that outputs the ego trajectory with 8 waypoints in 2Hz.

130 For the *Trajectory Difference* of expert action controllability, the IDM transform model’s action-  
 131 control prediction into an estimated trajectory, and then we measure how closely the estimated  
 132 trajectory matches the ground truth according to their L2 distance. A lower distance signifies a better  
 133 action controllability of the driving world model. This IDM is trained on Waymo training set [57] for  
 134 40 epochs with a learning rate of  $1 \times 10^{-4}$ . For video prediction-based policy, the IDM transforms  
 135 ReSim’s action-free prediction (without command and ego future trajectory as condition) into an  
 136 executable trajectory for planning. The IDM is trained on navtrain split of NAVSIM for 100 epochs.  
 137 The learning rate for first 50 epochs is  $1 \times 10^{-4}$  and decreases to  $1 \times 10^{-5}$  for the last 50 epochs.

138 **Visual Odometry(VO)-based Planner.** The VO-based planner is utilized as a baseline for video  
 139 prediction-based policy as in Main Tab. 4, and an agent that drives within the simulated world of  
 140 ReSim for closed-loop visual simulation as in Main Fig. 8. It shares similar architecture and training  
 141 to the aforementioned NAVSIM IDM. The only difference is that, instead of ingesting the whole  
 142 video sequence containing both history and future frames as NAVSIM IDM, the VO-based planner  
 143 takes historical frames as input only, without any explicit clue of the future observations.

### 144 B.3 Sampling

145 With ReSim, each short-term future video is simulated by sampling with the DDIM sampler [68]  
 146 for 50 steps. The simulated outcome is a 4s video sequence in 10Hz with a resolution of  $512 \times 896$ .  
 147 The input conditions include 9 frames of historical observations in 10Hz, an optional high-level  
 148 command, and an optional ego trajectory with 8 future waypoints in 2Hz. The high-level com-  
 149 mand is in one of “Turning left”, “Moving forward”, and “Turning right”, and is clas-  
 150 sified either by estimated flow for OpenDV dataset [39] or ego trajectory for action-annotated  
 151 datasets like NAVSIM [55] following common practice in [69, 70]. We always apply a prefix  
 152 prompt, “This video depicts a realistic view from the driver’s perspective of  
 153 a car driving on the road.”, concatenated with the textual command for both training and  
 154 sampling. Empirically, this prefix helps guide the model to generate driving scenarios. Following  
 155 CogVideoX [51], we apply a decreasing classifier-free guidance strategy with guidance scale starting  
 156 from 7.5 and gradually decreasing to 1. To synthesize a longer future beyond the training horizon (4s),  
 157 we can leverage the last 9 frames from the newly generated sequence as the context for next-round  
 158 prediction iteratively. Simulating a 4-second video sequence takes two minutes on a single Nvidia  
 159 A100 GPU.

### 160 B.4 Human Evaluation

161 The human evaluation for non-expert action controllability (Main Sec. 3.1) is conducted with 15  
 162 participants and 40 questions for each participant, resulting in 600 answers in total. As showcased  
 163 in Fig. S.2, each participant is requested to choose their preferred one among the synthesized video  
 164 of three candidate models for each evaluation aspect. The candidate models are Vista [41], ReSim  
 165 w/o simulated data, and ReSim (ours), and the evaluation aspects are Visual Realism and Trajectory  
 166 Following. The association of different models and their generations is anonymous to participants.



## C Additional Results

### C.1 Action Controllability

We provide additional visualizations for zero-shot action controllability in Fig. S.3 and Fig. S.4 for nuScenes and Waymo samples, respectively. Both datasets are unseen during training. Qualitative results demonstrate that ReSim can be flexibly controlled by both ground-truth trajectory (expert action) and randomly associated trajectory (non-expert actions).

### C.2 Ablation Study for Simulated Data

As shown in Fig. S.5, jointly training with simulated data improves the controllability of ReSim in open-world scenarios. Samples are from OpenDV validation set [39] with randomly associated trajectories from other labeled datasets.

### C.3 Action-free Prediction

We show the action-free prediction ability of ReSim in Fig. S.6. When conditioned on historical frames only without action inputs, ReSim synthesizes a possible future outcome, that might differ from the ground-truth due to the multi-modality of driving scenarios [71].

### C.4 Long-horizon Prediction

We compare ReSim with Vista [41] on long-horizon prediction in Fig. S.7. Starting from the same scenario, ReSim can emulate a more visually rich future in a longer horizon. This generation process does not use any action conditions, and both models perform multi-round rollouts that iteratively condition on the previously generated sequence to extend the prediction horizon.

### C.5 Failure Mode

Although ReSim exhibits improved fidelity and controllability over previous methods, it still faces challenges as in Fig. S.8. We discuss the limitation in Sec. D (Societal Impact).

## D Limitations and Broader Impact

**Inference Efficiency.** Despite the improved fidelity and controllability of our proposed ReSim, its real-world application is still potentially bottlenecked by the inference efficiency since diffusion models typically require multiple rounds of denoising process to ensure the generation quality [43, 37, 41]. To improve the inference latency, one potential solution is to reduce the number of denoising steps during the sampling phase. Recent advances in robotics [72] have proven that even with a single forward pass of the generative denoising network, the produced representation would greatly benefit downstream planning performance. Another approach is to distill a large yet slow diffusion model into a smaller one, which can be real-time deployed [73, 74].

**World Model for Policy Training.** Besides the onboard deployment of the heavy world model, another promising direction is to apply the world model as a dynamic environment to train policies [5, 2, 75]. This is beneficial as we can then deploy the policy to the autonomy directly, instead of the world model, upon the training convergence of the policy model. Inspired by the tremendous success of large-scale policy learning within the abstract simulator without visual signals [76], the proposed ReSim offers a great opportunity to reproduce and go beyond the human-level robustness in the regime of vision-based driving [69, 77] by scaling up ReSim’s visual simulation. We will follow this research direction in future work.

**Closed-loop Benchmark.** As illustrated in the results in Main Sec. 3.2, ReSim can reactively expose the policy to new states beyond the human driving logs when serving as a closed-loop visual simulator, in contrast to current predominant evaluation benchmarks for end-to-end autonomous driving [56, 57, 55]. However, since ReSim is trained on front-view observations only, common planning methods with multi-view camera inputs, such as UniAD [69] and VAD [77], cannot be

211 readily applied in such simulation. Moreover, how to fairly benchmark different policies quantitatively  
212 using ReSim is still worth exploration.

213 **Societal Impact.** Though meticulously developed with state-of-the-art performance shown in the  
214 results, ReSim might still exhibit uncontrollable visual artifacts in generation due to the stochastic  
215 nature of the diffusion framework. It might also hallucinate in complex scenarios with multiple agents  
216 involved, and further pose risks for downstream applications. Despite the training on large-scale  
217 datasets, the uncurated data distribution, such as geographical regions, might lead to biased behavior  
218 of the learned model. We hope our work could shed light on the construction of open-world neural  
219 simulation for physical intelligence spanning both driving and robotics, by leveraging the visual  
220 richness of the real world and the action flexibility of the simulated world collectively.

## 221 E License of Assets

222 Our training and evaluation are conducted on publicly licensed datasets and benchmarks [56, 78, 57,  
223 55, 39]. To improve action diversity, we collected some data from the CARLA simulator [79] under  
224 the CC-BY License. The scenario configurations for the CARLA data follow Bench2Drive [58] under  
225 CC BY-NC-SA 4.0. ReSim is developed upon CogVideoX [51], with both code and model under  
226 the Apache License 2.0. We adopt public visual encoders, including DINOv2 [63] (under Apache  
227 License 2.0) and XVO [67] (under CC BY-NC-SA 4.0) for the construction of our Video2Reward  
228 and inverse dynamics model, respectively. Vista [41] is leveraged as a comparative baseline, which is  
229 under Apache License 2.0. We will release our code and models under the Apache License 2.0.

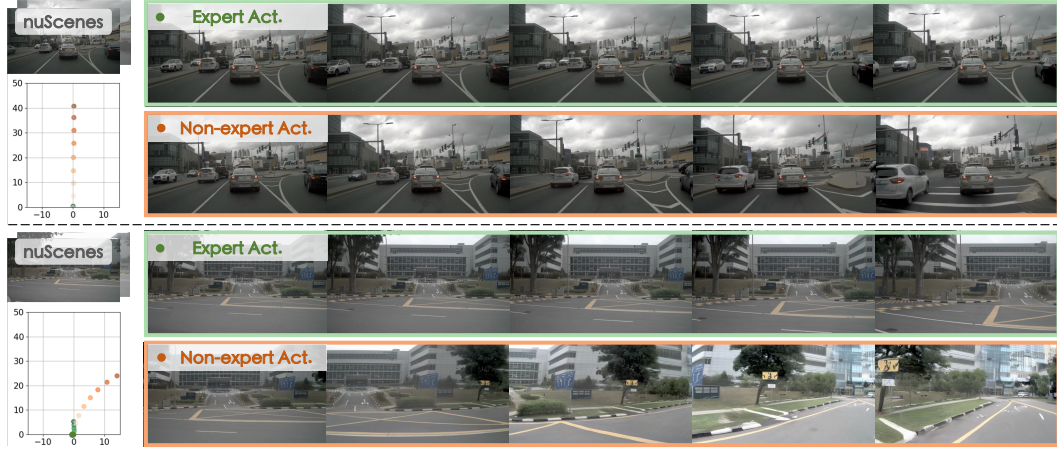


Figure S.3: **Visualizations for zero-shot action controllability on nuScenes.** The **expert** actions are recorded ground-truth from the driving log, while **non-expert** actions are randomly sampled from other scenarios. Best viewed zoomed in.

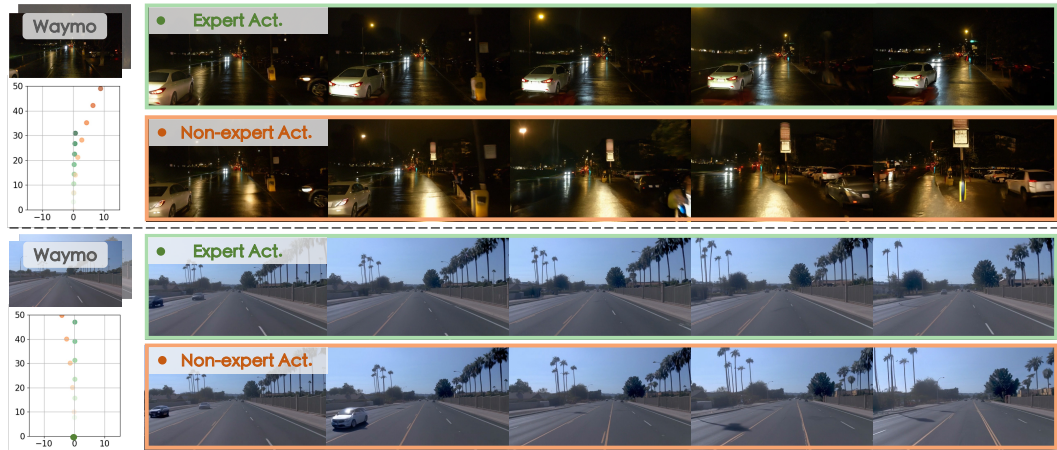


Figure S.4: **Visualizations for zero-shot action controllability on Waymo.** The **expert** actions are recorded ground-truth from the driving log, while **non-expert** actions are randomly sampled from other scenarios. Best viewed zoomed in.

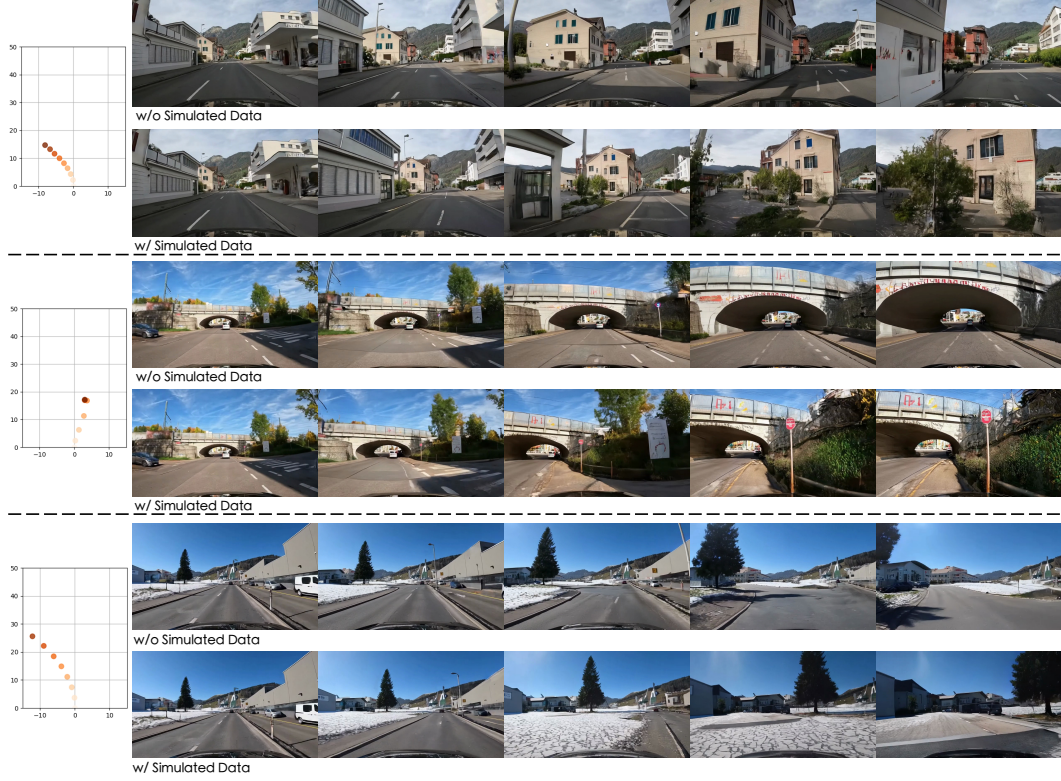


Figure S.5: **Additional ablations for incorporating simulated data in training.** Simulated data improves controllability of ReSim for **non-expert** actions. Historical frames are not shown for brevity.



Figure S.6: **Visualizations for action-free future prediction.** ReSim can predict the future without action conditions by inferring from historical frames only.



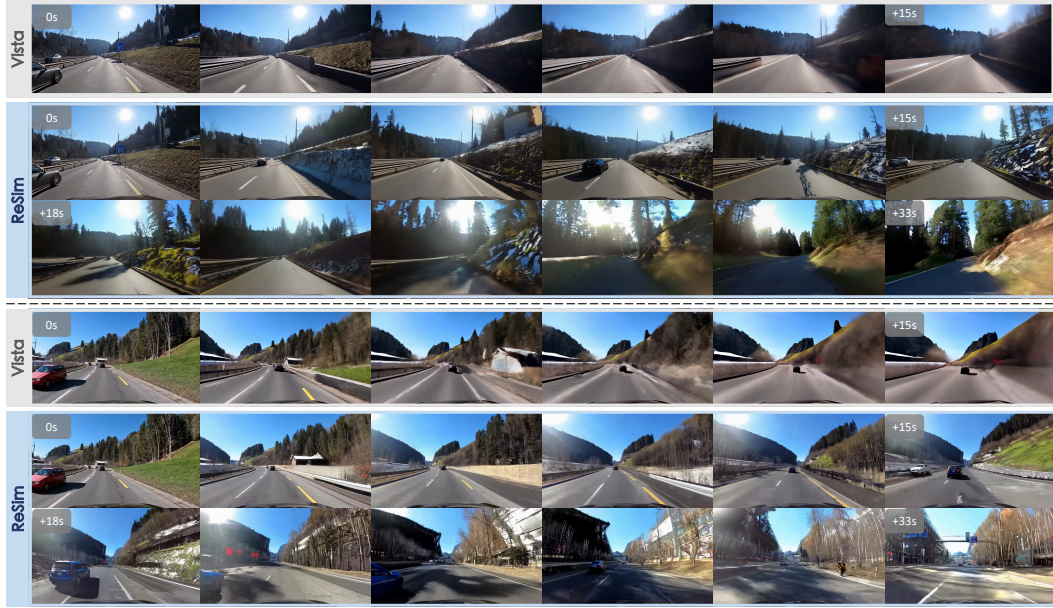


Figure S.7: **Long-term future prediction.** Compared to Vista whose prediction fidelity severely degrades in 15s, ReSim can predict consistent future states with rich details in more than 30s.



Figure S.8: **Failure modes.** ReSim still struggles in certain scenarios, such as falsely crossing the parapet, poor consistency for occluded objects, and producing visual artifacts for extreme cases. Best viewed zoomed in.



## References

- [1] Yann LeCun. A path towards autonomous machine intelligence. *Open Review*, 62, 2022. 1
- [2] Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to Control: Learning behaviors by latent imagination. In *ICLR*, 2020. 1, 5
- [3] Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with discrete world models. In *ICLR*, 2021. 1
- [4] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023. 1
- [5] David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *NeurIPS*, 2018. 1, 2, 5
- [6] Jialong Wu, Haoyu Ma, Chaoyi Deng, and Mingsheng Long. Pre-training contextualized world models with in-the-wild videos for reinforcement learning. In *NeurIPS*, 2023. 1
- [7] Frederik Ebert, Chelsea Finn, Sudeep Dasari, Annie Xie, Alex Lee, and Sergey Levine. Visual Foresight: Model-based deep reinforcement learning for vision-based robotic control. *arXiv preprint arXiv:1812.00568*, 2018. 1
- [8] Chelsea Finn and Sergey Levine. Deep visual foresight for planning robot motion. In *ICRA*, 2017. 1
- [9] 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. 1
- [10] Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *ICML*, 2024. 1
- [11] Eloi Alonso, Adam Jelley, Vincent Micheli, Anssi Kanervisto, Amos Storkey, Tim Pearce, and François Fleuret. Diffusion for world modeling: Visual details matter in atari. In *NeurIPS*, 2024. 1
- [12] Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. Diffusion models are real-time game engines. In *ICLR*, 2024. 1, 3
- [13] Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. Navigation world models. In *CVPR*, 2025. 1
- [14] Jing Yu Koh, Honglak Lee, Yinfei Yang, Jason Baldridge, and Peter Anderson. Pathdreamer: A world model for indoor navigation. In *ICCV*, 2021. 1
- [15] Mengjiao Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. In *ICLR*, 2024. 1, 2
- [16] Russell Mendonca, Shikhar Bahl, and Deepak Pathak. Structured world models from human videos. In *RSS*, 2023. 1
- [17] Yilun Du, Sherry Yang, Pete Florence, Fei Xia, Ayzaan Wahid, brian ichter, Pierre Sermanet, Tianhe Yu, Pieter Abbeel, Joshua B. Tenenbaum, Leslie Pack Kaelbling, Andy Zeng, and Jonathan Tompson. Video language planning. In *ICLR*, 2024. 1
- [18] Sherry Yang, Jacob Walker, Jack Parker-Holder, Yilun Du, Jake Bruce, Andre Barreto, Pieter Abbeel, and Dale Schuurmans. Video as the new language for real-world decision making. In *ICML*, 2024. 1
- [19] Dian Chen and Philipp Krähenbühl. Learning from all vehicles. In *CVPR*, 2022. 1
- [20] Henry X Liu and Shuo Feng. Curse of rarity for autonomous vehicles. *Nature Communications*, 2024. 1
- [21] Marc G Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The arcade learning environment: An evaluation platform for general agents. *JAIR*, 2013. 1
- [22] Saran Tunyasuvunakool, Alistair Muldal, Yotam Doron, Siqi Liu, Steven Bohez, Josh Merel, Tom Erez, Timothy Lillicrap, Nicolas Heess, and Yuval Tassa. Dm\_control: Software and tasks for continuous control. *Software Impacts*, 2020. 1
- [23] Michał Kempka, Marek Wydmuch, Grzegorz Runc, Jakub Toczek, and Wojciech Jaśkowski. ViZDoom: A doom-based ai research platform for visual reinforcement learning. In *CIG*, 2016. 1

- [24] 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. In *NeurIPS*, 2022. 1
- [25] 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). In *ECCV*, 2024. 1
- [26] Yumeng Zhang, Shi Gong, Kaixin Xiong, Xiaoqing Ye, Xiao Tan, Fan Wang, Jizhou Huang, Hua Wu, and Haifeng Wang. BEVWorld: A multimodal world model for autonomous driving via unified bev latent space. *arXiv preprint arXiv:2407.05679*, 2024. 1
- [27] Daniel Bogdoll, Yitian Yang, and J Marius Zöllner. MUVO: A multimodal generative world model for autonomous driving with geometric representations. *arXiv preprint arXiv:2311.11762*, 2023. 1
- [28] Zetong Yang, Li Chen, Yanan Sun, and Hongyang Li. Visual point cloud forecasting enables scalable autonomous driving. In *CVPR*, 2024. 1
- [29] Lunjun Zhang, Yuwen Xiong, Ze Yang, Sergio Casas, Rui Hu, and Raquel Urtasun. Learning unsupervised world models for autonomous driving via discrete diffusion. In *ICLR*, 2024. 1
- [30] Wenzhao Zheng, Weiliang Chen, Yuanhui Huang, Borui Zhang, Yueqi Duan, and Jiwen Lu. OccWorld: Learning a 3D occupancy world model for autonomous driving. In *ECCV*, 2024. 1
- [31] Sicheng Zuo, Wenzhao Zheng, Yuanhui Huang, Jie Zhou, and Jiwen Lu. GaussianWorld: Gaussian world model for streaming 3D occupancy prediction. *arXiv preprint arXiv:2412.10373*, 2024. 1
- [32] Seung Wook Kim, Jonah Philion, Antonio Torralba, and Sanja Fidler. DriveGAN: Towards a controllable high-quality neural simulation. In *CVPR*, 2021. 1
- [33] Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jiagang Zhu, and Jiwen Lu. DriveDreamer: Towards real-world-driven world models for autonomous driving. In *ECCV*, 2024. 1
- [34] 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. 1
- [35] Mariam Hassan, Sebastian Stapf, Ahmad Rahimi, Pedro Rezende, Yasaman Haghighi, David Brüggemann, Isinsu Katircioglu, Lin Zhang, Xiaoran Chen, Suman Saha, et al. GEM: A generalizable ego-vision multimodal world model for fine-grained ego-motion, object dynamics, and scene composition control. In *CVPR*, 2025. 1
- [36] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1, 2
- [37] 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. 1, 2, 5
- [38] 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. 1
- [39] 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. In *CVPR*, 2024. 1, 2, 4, 5, 6
- [40] Yuqi Wang, Ke Cheng, Jiawei He, Qitai Wang, Hengchen Dai, Yuntao Chen, Fei Xia, and Zhaoxiang Zhang. DrivingDojo Dataset: Advancing interactive and knowledge-enriched driving world model. In *NeurIPS Datasets and Benchmarks*, 2024. 1, 2
- [41] 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. In *NeurIPS*, 2024. 1, 3, 4, 5, 6
- [42] Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. DINO-WM: World models on pre-trained visual features enable zero-shot planning. *arXiv preprint arXiv:2411.04983*, 2024. 1
- [43] 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. In *CVPR*, 2024. 1, 5

- [44] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. BEVFormer: Learning bird’s-eye-view representation from multi-camera images via spatiotemporal transformers. In *ECCV*, 2022. 1
- [45] Bencheng Liao, Shaoyu Chen, Xinggang Wang, Tianheng Cheng, Qian Zhang, Wenyu Liu, and Chang Huang. MapTR: Structured modeling and learning for online vectorized hd map construction. In *ICLR*, 2023. 1
- [46] Xinzhu Ma, Wanli Ouyang, Andrea Simonelli, and Elisa Ricci. 3D object detection from images for autonomous driving: a survey. *IEEE TPAMI*, 2023. 1
- [47] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *ICLR*, 2024. 2
- [48] 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
- [49] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. AnimateDiff: Animate your personalized text-to-image diffusion models without specific tuning. In *ICLR*, 2024. 2
- [50] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent Video Diffusion Models for High-Fidelity Long Video Generation. *arXiv preprint arXiv:2211.13221*, 2022. 2
- [51] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. CogVideoX: Text-to-video diffusion models with an expert transformer. In *ICLR*, 2025. 2, 4, 6
- [52] Shoufa Chen, Mengmeng Xu, Jiawei Ren, Yuren Cong, Sen He, Yanping Xie, Animesh Sinha, Ping Luo, Tao Xiang, and Juan-Manuel Perez-Rua. GenTron: Diffusion transformers for image and video generation. In *CVPR*, 2024. 2
- [53] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023. 2
- [54] Itseez. Open source computer vision library. <https://github.com/itseez/opencv>, 2015. 2
- [55] Daniel Dauner, Marcel Hallgarten, Tianyu Li, Xinshuo Weng, Zhiyu Huang, Zetong Yang, Hongyang Li, Igor Gilitschenski, Boris Ivanovic, Marco Pavone, et al. NAVSIM: Data-driven non-reactive autonomous vehicle simulation and benchmarking. In *NeurIPS Datasets and Benchmarks*, 2024. 2, 4, 5, 6
- [56] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yuxin Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020. 2, 5, 6
- [57] 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, 4, 5, 6
- [58] Xiaosong Jia, Zhenjie Yang, Qifeng Li, Zhiyuan Zhang, and Junchi Yan. Bench2Drive: Towards multi-ability benchmarking of closed-loop end-to-end autonomous driving. In *NeurIPS Datasets and Benchmarks*, 2024. 2, 6
- [59] Jens Beißwenger. PDM-Lite: A rule-based planner for carla leaderboard 2.0. [https://github.com/OpenDriveLab/DriveLM/blob/DriveLM-CARLA/pdm\\_lite/docs/report.pdf](https://github.com/OpenDriveLab/DriveLM/blob/DriveLM-CARLA/pdm_lite/docs/report.pdf), 2024. 2
- [60] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 2020. 2
- [61] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *JMLR*, 2022. 3
- [62] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS Workshops*, 2021. 3
- [63] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *TMLR*, 2024. 3, 6

- 376 [64] CARLA autonomous driving leaderboard. <https://leaderboard.carla.org/>, 2022. 3
- 377 [65] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. *arXiv preprint*  
378 *arXiv:1711.05101*, 2017. 3
- 379 [66] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. VideoGPT: Video generation using  
380 vq-vae and transformers. *arXiv preprint arXiv:2104.10157*, 2021. 3
- 381 [67] Lei Lai, Zhongkai Shanguan, Jimuyang Zhang, and Eshed Ohn-Bar. XVO: Generalized visual odometry  
382 via cross-modal self-training. In *ICCV*, 2023. 4, 6
- 383 [68] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint*  
384 *arXiv:2010.02502*, 2020. 4
- 385 [69] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei  
386 Lin, et al. Planning-oriented autonomous driving. In *CVPR*, 2023. 4, 5
- 387 [70] Shengchao Hu, Li Chen, Penghao Wu, Hongyang Li, Junchi Yan, and Dacheng Tao. ST-P3: End-to-end  
388 vision-based autonomous driving via spatial-temporal feature learning. In *ECCV*, 2022. 4
- 389 [71] Li Chen, Penghao Wu, Kashyap Chitta, Bernhard Jaeger, Andreas Geiger, and Hongyang Li. End-to-end  
390 autonomous driving: Challenges and frontiers. *IEEE TPAMI*, 2024. 5
- 391 [72] Yucheng Hu, Yanjiang Guo, Pengchao Wang, Xiaoyu Chen, Yen-Jen Wang, Jianke Zhang, Koushil Sreenath,  
392 Chaochao Lu, and Jianyu Chen. Video prediction policy: A generalist robot policy with predictive visual  
393 representations. *arXiv preprint arXiv:2412.14803*, 2024. 5
- 394 [73] Yuanzhi Zhu, Hanshu Yan, Huan Yang, Kai Zhang, and Junnan Li. Accelerating video diffusion models  
395 via distribution matching. *arXiv preprint arXiv:2412.05899*, 2024. 5
- 396 [74] Tianwei Yin, Qiang Zhang, Richard Zhang, William T Freeman, Fredo Durand, Eli Shechtman, and Xun  
397 Huang. From slow bidirectional to fast causal video generators. *arXiv preprint arXiv:2412.07772*, 2024. 5
- 398 [75] Mitchell Goff, Greg Hogan, George Hotz, Armand du Parc Locmaria, Kacper Raczy, Harald Schäfer,  
399 Adeeb Shihadeh, Weixing Zhang, and Yassine Yousfi. Learning to drive from a world model. *arXiv*  
400 *preprint arXiv:2504.19077*, 2025. 5
- 401 [76] Marco Cusumano-Towner, David Hafner, Alex Hertzberg, Brody Huval, Aleksei Petrenko, Eugene Vinitsky,  
402 Erik Wijmans, Taylor Killian, Stuart Bowers, Ozan Sener, et al. Robust autonomy emerges from self-play.  
403 *arXiv preprint arXiv:2502.03349*, 2025. 5
- 404 [77] Bo Jiang, Shaoyu Chen, Qing Xu, Bencheng Liao, Jiajie Chen, Helong Zhou, Qian Zhang, Wenyu Liu,  
405 Chang Huang, and Xinggang Wang. VAD: Vectorized scene representation for efficient autonomous  
406 driving. In *ICCV*, 2023. 5
- 407 [78] Holger Caesar, Juraj Kabzan, Kok Seang Tan, Whye Kit Fong, Eric M. Wolff, Alex H. Lang, Luke Fletcher,  
408 Oscar Beijbom, and Sammy Omari. nuPlan: A closed-loop ml-based planning benchmark for autonomous  
409 vehicles. In *CVPR Workshops*, 2021. 6
- 410 [79] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An  
411 open urban driving simulator. In *CoRL*, 2017. 6