Generalized Predictive Model for Autonomous Driving

Jiazhi Yang[1*] Shenyuan Gao[2,1*] Yihang Qiu[1*] Li Chen[3,1†] Tianyu Li[1] Bo Dai[1]
Kashyap Chitta[4,5] Penghao Wu[1] Jia Zeng[1] Ping Luo[3] Jun Zhang[2♮]
Andreas Geiger[4,5♮] Yu Qiao[1♮] Hongyang Li[1†]

[1] OpenDriveLab and Shanghai AI Lab [2] Hong Kong University of Science and Technology
[3] University of Hong Kong [4] University of Tübingen [5] Tübingen AI Center

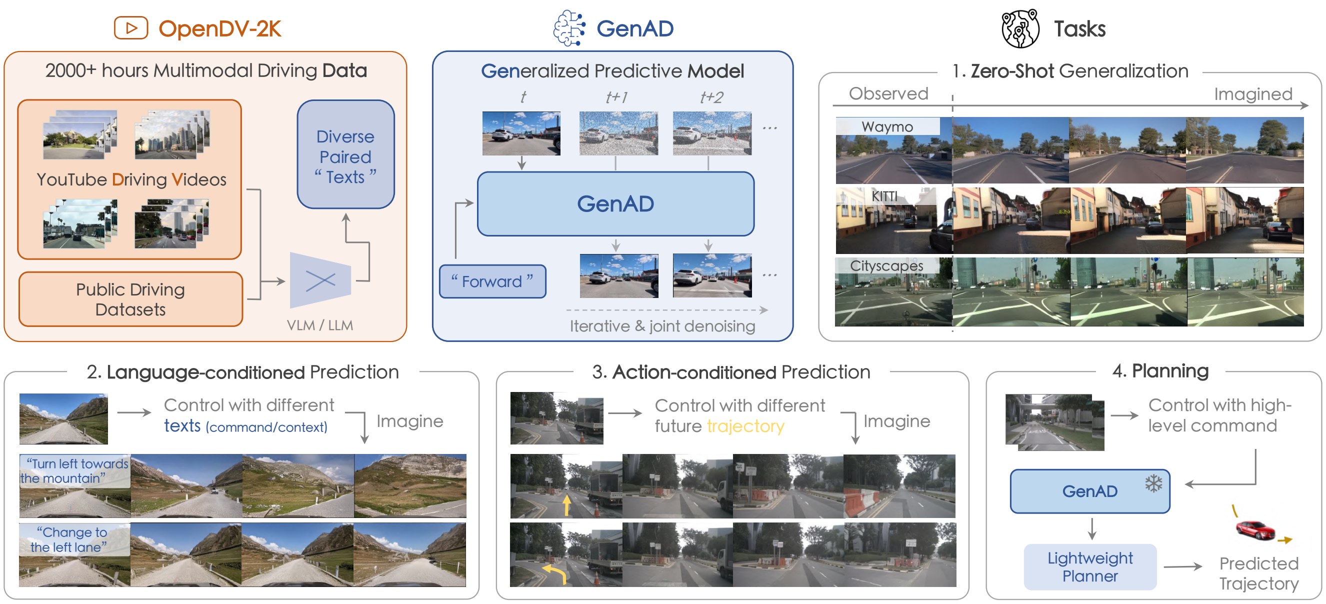

Figure 1. **Overview of the GenAD paradigm**. We aim to establish a generalized video prediction paradigm for autonomous driving by presenting the largest multimodal driving video dataset to date, **OpenDV-2K**, and a generative model that predicts the future given past visual and textual input, **GenAD**. The strong generalization and controllability of GenAD is validated spanning a diverse spectrum of tasks, including zero-shot domain transfer, language-conditioned prediction, action-conditioned prediction, and motion planning.

Abstract

In this paper, we introduce the first large-scale video prediction model in the autonomous driving discipline. To eliminate the restriction of high-cost data collection and empower the generalization ability of our model, we acquire massive data from the web and pair it with diverse and high-quality text descriptions. The resultant dataset accumulates over 2000 hours of driving videos, spanning areas all over the world with diverse weather conditions and traffic scenarios. Inheriting the merits from recent latent diffusion models, our model, dubbed GenAD, handles the challenging dynamics in driving scenes with novel temporal reasoning blocks. We showcase that it can generalize to various unseen driving datasets in a zero-shot manner, surpassing general or driving-specific video prediction counterparts. Furthermore, GenAD can be adapted into an action-conditioned prediction model or a motion planner, holding great potential for real-world driving applications.

1. Introduction

Autonomous driving agents, as a promising application of high-level artificial intelligence, perceive the surrounding

*Equal contribution, ordered by coin toss. ♮Equal co-advising.
†Project lead. Primary contact: yangjiazhi@opendrivelab.com

environment, build internal world model representations, make decisions, and take actions in response [9, 50]. However, despite dedicated efforts in academia and industry for decades, their deployment is still restricted to certain areas or scenarios, and they cannot be applied over the world seamlessly. One critical reason is the limited generalization ability of learned models in structured autonomous driving systems. Typically, perception models face challenges of generalizing to diverse environments with changes in geographical locations, sensor configurations, weather conditions, open-set objects, *etc*.; prediction and planning models fail to generalize to nondeterministic futures with rare scenarios and different driving intentions [2, 16, 54].

Motivated by how humans learn to perceive and cognize the world [27, 28, 49], we advocate employing driving videos as the universal interface that generalizes to diverse environments with dynamic futures. Based on this, a driving video predictive model is preferred to fully capture the world knowledge about driving scenarios (Fig. 1). By predicting the future, the video predictor essentially learns two vital aspects of autonomous driving: how the world operates, and how to maneuver safely in the wild.

Recently, the community has begun to adopt video as the interface to represent observation behavior and action for various robot tasks [11]. For domains such as classical video prediction and robotics, the video backgrounds are mostly static, the movement of robots is slow, and the resolution of videos is low. In contrast, for the driving scenarios, it struggles with outdoor environments being highly dynamic, agents encompassing much larger motions, and the sensory resolution covering a large range of view. These distinctions lead to substantial challenges for autonomous driving applications. Fortunately, there are some preliminary attempts on developing a video predictive model in the driving domain [4, 15, 19, 23, 25, 33, 38, 45, 47]. Despite promising progress in terms of prediction quality, these attempts have not achieved desirable capability of generalization as in classical robot tasks (*e.g*., manipulation), being confined to either limited scenarios such as highways with low traffic density [4] and small-scale datasets [15, 23, 33, 45, 47], or restricted conditions that raises difficulties to generate diverse environments [38]. How to unveil the potential of video prediction models for driving remains seldom explored.

Motivated by the discussions above, we target at building a video predictive model for autonomous driving, capable of generalizing to new conditions and environments. To this end, we have to answer the following questions: *(1) What data can be obtained in a feasible and scalable manner? (2) How can we formulate a predictive model to capture the complex evolution of dynamic scenarios? (3) How can we apply the (foundation) model for downstream tasks?*

Scaled Data. To achieve powerful generalization ability, a substantial and diverse corpus of data is necessary. Inspired by the success of learning from Internet-scale data in foundation models [1, 26, 39], we construct our driving dataset from both the web and publicly licensed datasets. Compared to existing options, which are limited in scale and diversity due to their regulated collection processes, online data owns great diversity in several aspects: geographic locations, terrains, weather conditions, safety-critical scenarios, sensor settings, traffic elements, *etc*. To guarantee the data is of high-quality and desirable for large-scale training, we exhaustively collect driving recordings on YouTube and remove unintended corruption frames via rigorous human verification. Furthermore, videos are paired with diverse text-level conditions, including descriptions generated and refined with the aid of existing foundation models [30, 35], and high-level instructions inferred by a video classifier. Through these steps, we construct **OpenDV-2K**, the *largest* public driving dataset to date, containing more than 2000 hours of driving videos and being 374 times larger than the widely used nuScenes counterpart. Our dataset is publicly available at `https://github.com/OpenDriveLab/DriveAGI`.

Generalized Predictive Model. Learning a generalized driving video predictor bears several key challenges: generation quality, training efficiency, causal reasoning, and drastic view shift. We address these aspects by presenting a novel temporal generative model with two-stage learning. To capture the environment details, enhance generation quality, and maintain training efficiency simultaneously, we build upon the recent success of *latent diffusion models* (LDMs) [37, 41]. In the first stage, we transfer the generation distribution of LDM from its pre-trained general vision domain to the driving domain by fine-tuning it on OpenDV-2K images. In the second stage, we interleave the proposed temporal reasoning blocks into the original model and learn to predict the future given past frames and conditions. Contrary to conventional temporal modules [4, 18] that suffer from causal confusion and large motion, our solution consists of causal temporal attention and decoupled spatial attention to efficiently model the drastic spatiotemporal shift in highly dynamic driving scenes. After sufficient training, our **Gen**erative model for **A**utonomous **D**riving (**GenAD**)[1] can generalize to various scenarios in a zero-shot fashion.

Extensions for Simulation and Planning. After large-scale pre-training of video prediction, GenAD essentially understands how the world evolves and how to drive. We show how to adapt its learned knowledge for real-world driving problems, *i.e*., simulation and planning. For simulation, we fine-tune the pre-trained model with future ego trajectories as additional conditions, to associate future imaginations with different ego actions. We also empower

[1]Note that GenAD is abbreviated from both **Gen**erative models and **Gen**eralized capabilities.

	Dataset	Duration (hours)	Front-view Frames	Geographic Diversity Countries	Cities	Sensor Setup
✗	KITTI [14]	1.4	15k	1	1	fixed
✗	Cityscapes [10]	0.5	25k	3	50	fixed
✗	Waymo Open* [43]	11	390k	1	3	fixed
✗	Argoverse 2* [48]	4.2	300k	1	6	fixed
✓	nuScenes [6]	5.5	241k	2	2	fixed
✓	nuPlan* [7]	120	4.0M	2	4	fixed
✓	Talk2Car [12]	4.7	-	2	2	fixed
✓	ONCE [34]	144	7M	1	-	fixed
✓	Honda-HAD [24]	32	1.2M	1	-	fixed
✓	Honda-HDD-Action [40]	104	1.1M	1	-	fixed
✓	Honda-HDD-Cause [40]	32	-	1	-	fixed
✓	OpenDV-YouTube (Ours)	1747	60.2M	$\geq 40^\dagger$	$\geq 244^\dagger$	uncalibrated
-	**OpenDV-2K** (Ours)	**2059**	**65.1M**	$\geq \mathbf{40}^\dagger$	$\geq \mathbf{244}^\dagger$	**uncalibrated**

Table 1. **OpenDV-2K comparison at a glance to existing counterparts in terms of scale and diversity**. Note that datasets with ✓ are included in OpenDV-2K (last row). *Perception subset in Waymo Open, Argoverse 2, and nuPlan. †Estimated by GPT [36] from video titles.

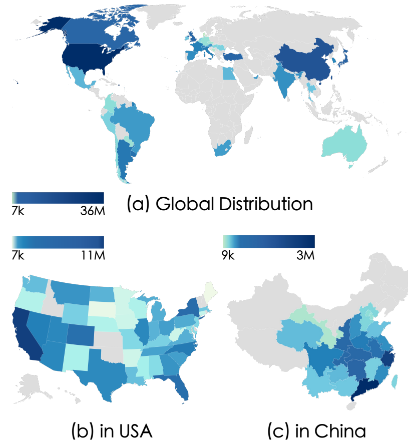

Figure 2. **Geographic distribution of OpenDV-2K**. Our dataset covers ample driving scenarios around the world.

GenAD to perform planning on challenging benchmarks by using a lightweight planner to translate latent features into the future trajectory of the ego vehicle. On account of its pre-trained ability to predict accurate future frames, our algorithm exhibits promising results in both simulation consistency and planning reliability.

2. OpenDV-2K Dataset

We introduce OpenDV-2K, a large-scale multimodal dataset for autonomous driving, to support the training of a generalized video prediction model. The main component is a vast corpus of high-quality YouTube driving videos, which are collected from all over the world, and are gathered into our dataset after a careful curation process. We automatically create language annotations for these videos using vision-language models. To further improve its diversity in sensor configurations and language expressions, we merge 7 publicly licensed datasets into our OpenDV-2K, as shown in Tab. 1. As a result, OpenDV-2K occupies a total of 2059 hours of videos paired with texts, including 1747 hours from YouTube and 312 hours from public datasets. We use OpenDV-YouTube and OpenDV-2K to specify the YouTube split and the overall dataset, respectively.

2.1. Diversity over Prior Datasets

A brief comparison with other public datasets is provided in Tab. 1. Beyond its significant scale, the proposed OpenDV-2K represents *diversity* across various aspects as follows.

Globe-wise Geographic Distribution. Due to the global nature of online videos, OpenDV-2K covers more than 40 countries and 244 cities worldwide. This is a tremendous improvement over previous public datasets, which are typically gathered in a small number of restricted areas. We plot the specific distribution of OpenDV-YouTube in Fig. 2.

Open-world Driving Scenarios. Our dataset provides a huge amount of realistic driving experience in the open world, covering rare environments like forests, extreme weather conditions like heavy snow, and appropriate driving behaviors in response to interactive traffic situations. These data are crucial for diversity and generalization yet are seldom collected in existing public datasets.

Unrestricted Sensor Configurations. Current driving datasets are confined to specific sensor configurations, including intrinsic and extrinsic camera parameters, image, sensor type, optics, *etc.*, which poses great challenges for deploying the learned models with different sensors [32]. In contrast, YouTube driving videos are recorded in various types of vehicles with flexible camera setups, which aids in the robustness of the trained model when deployed using a novel camera setting.

2.2. Towards High-quality Multimodal Dataset

Driving Video Collection and Curation. Finding clean driving videos from the vast pool of the web is a tedious and costly task. To simplify the process, we start by selecting certain video uploaders, *i.e.*, YouTubers. Judging from the average length and overall quality, we collect 43 YouTubers with 2139 high-quality front-view driving videos. To make sure there is no overlap between training and validation sets, we take all videos from 3 YouTubers for validation, with the remaining videos as the training set. To rule out non-driving frames like video introductions and subscription reminders, we discard a certain length of segments at the beginning and end of each video. Each frame is then described with language contexts using a VLM model, BLIP-2 [30]. We further remove the black frames and transition frames, which are not ideal for training, by manually checking if there are certain keywords in these contexts. We give an illustration of the dataset construction pipeline in Appendix C.1.1, and we introduce how to generate the contexts below.

Language Annotation for YouTube Videos. To create a predictive model that can be controlled by natural language to simulate different futures accordingly, To make the predictive model controllable and improve the sample quality [3], it is crucial to pair the driving videos with meaningful and varied language annotations. We construct two types of texts for OpenDV-YouTube, *i.e*., driving commands for ego-vehicle and frame descriptions, namely "command" and "context", to help the model comprehend ego actions and open-world concepts, respectively. For commands, we train a video classifier on Honda-HDD-Action [40] for 14 types of actions to label ego behaviors in a 4s sequence. These categorical commands will be further mapped to multiple free-form expressions from a predefined dictionary. For contexts, we leverage an established vision-language model, BLIP-2 [30], to describe the main objects and scenarios for each frame. For more details on annotations, please refer to Appendix C.1.2.

Enlarging Language Spectrum with Public Datasets. Considering that BLIP-2 annotations are generated for static frames without comprehension of dynamic driving scenarios such as the traffic light transitions, we exploit several public datasets that provide linguistic descriptions for driving scenarios [6, 7, 12, 24, 34, 40]. However, their metadata is relatively sparse with only a few words such as "sunny road". We further enhance their text quality using GPT [36] to form a descriptive "context" and generate a "command" by categorizing the logged trajectory for each video clip. Ultimately, we integrate these datasets with OpenDV-YouTube to establish OpenDV-2K dataset, as shown in the last row of Tab. 1.

3. GenAD Framework

In this section, we introduce the training and design of the GenAD model. As shown in Fig. 3, GenAD is trained in two stages, *i.e*., image domain transferring and video prediction pre-training. The first stage adapts the general text-to-image model to the driving domain (Sec. 3.1). The second stage lifts the text-to-image model to a video prediction model with our proposed temporal reasoning block and modified training schemes (Sec. 3.2). In Sec. 3.3, we explore how the predictive model can be extended to action-conditioned prediction and planning.

3.1. Image Domain Transfer

On-board cameras capture a large field of views with abundant visual contents, including the road, background buildings, surrounding vehicles, *etc*., which require strong and robust generation capability to produce continuous and realistic driving scenarios. To facilitate the learning process, we start with independent image generation in the first stage. Concretely, we initialize our model with SDXL [37], which is a large-scale latent diffusion model (LDM) for text-to-image generation, to leverage its ability to synthesize high-quality images with plenty of visual details. It is implemented as a denoising UNet $\mathbf{f}_\theta$ with several stacked convolution and attention blocks, which learns to synthesize images by denoising the noisy latents [41]. Specifically, given a noisy input latent $\mathbf{x}_t$ corrupted by the forward diffusion process, it is trained to predict the added noise $\boldsymbol{\epsilon}$ of $\mathbf{x}_t$ via the following objective:

$$\mathcal{L}_{\text{img}} := \mathbb{E}_{\mathbf{x},\boldsymbol{\epsilon}\sim\mathcal{N}(0,1),\mathbf{c},t}\Big[\|\boldsymbol{\epsilon} - \mathbf{f}_\theta(\mathbf{x}_t;\mathbf{c},t)\|_2^2\Big], \quad (1)$$

where $\mathbf{x}$ and $\mathbf{x}_t$ are the clean and noisy latent, respectively, t denotes the timestep for different noise scales, and $\mathbf{c}$ is the text condition that guides the denoising process, which is a concatenation of context and command. For training efficiency, the learning process takes place in a compressed latent space [13, 37, 41] instead of pixel space. During sampling, the model generates images from standard Gaussian noise by denoising the last-step predictions iteratively.

However, the original SDXL is trained on data in the general domain, such as portraits and artistic paintings, which are not concerned with autonomy systems. To adapt the model to synthesize images for driving, we fine-tune it on text-to-image generation using image-text pairs in OpenDV-2K with the same objective as Eq. (1). Following the original training of SDXL, all parameters θ of the UNet are fine-tuned at this stage, whereas the CLIP text encoders [39] and the autoencoder [13] remain frozen.

3.2. Video Prediction Pre-training

In the second stage, with a few frames of a consecutive video as past observations, GenAD is trained to reason about all visual observations and predict several future frames in plausible ways. Similar to the first stage stage, the prediction process can also be guided by text conditions. However, predicting the highly dynamic driving world temporally is challenging due to two fundamental barriers.

1. *Causal Reasoning*: To predict plausible futures following the temporal causality of the driving world, the model needs to comprehend the intentions of all other agents together with the ego vehicle, and understand underlying traffic rules, *e.g*., how the traffic will change with the transition of traffic lights.
2. *Drastic View Shift*: Contrary to typical video generation benchmarks which mainly have a static background with slow motion of centered objects, the view of driving changes drastically over time. Each pixel in every frame may move to a distant location in the next frame.

We propose temporal reasoning blocks to address these problems. As illustrated in Fig. 3(c), each block is composed of three successive attention layers, *i.e*., the causal

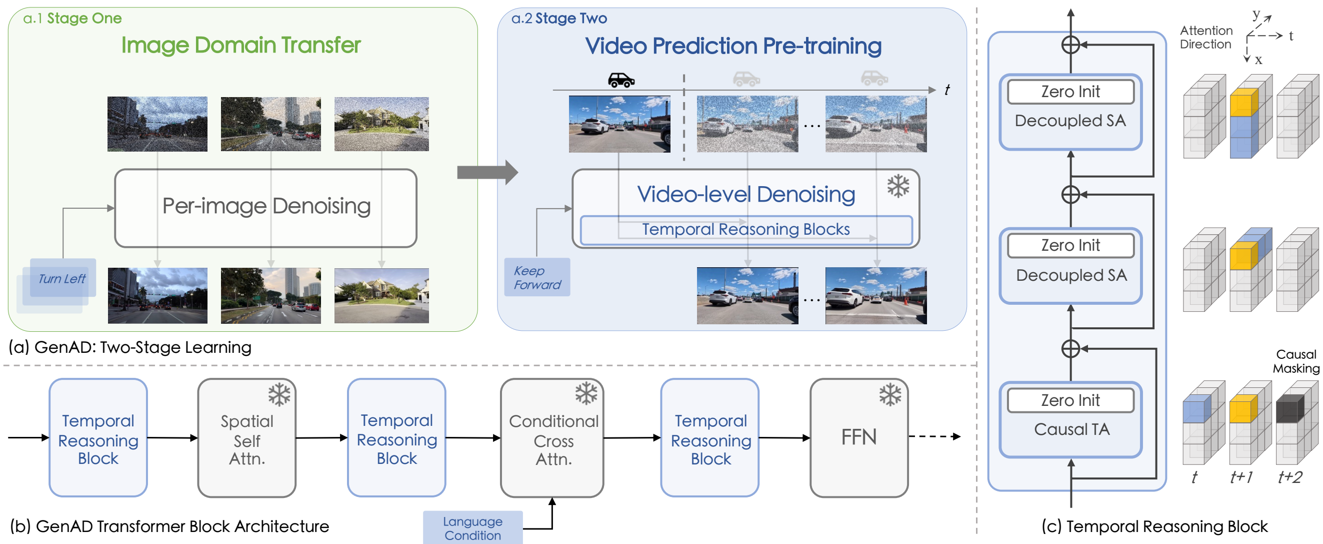

Figure 3. **Framework of GenAD.** (**a**) The two-stage learning for GenAD is composed of transferring the image domain of an image diffusion model to the driving field (a.1 Stage one), and video prediction pre-training for modeling the temporal dependency of videos (a.2 Stage two). (**b**) One transformer block in GenAD for the second stage training has interleaved temporal reasoning blocks before each frozen layer to align spatiotemporal features. (**c**) The proposed Temporal Reasoning Block includes one causal temporal attention (TA) and two decoupled spatial attention (SA) layers to extract features in different axes. A query grid attends to itself as well as blue grids while the dark gray grid is masked out in causal attention. '`Zero init`' is appended at the end of each attention block to stabilize training.

temporal attention layer and two decoupled spatial attention layers, which are tailored for the causal reasoning and modeling large shifts in the driving scenes, respectively.

Causal Temporal Attention. Since the model after the stage-one training can only process each frame independently, we leverage temporal attention to exchange information among different video frames. The attention takes place in the time axis and models the temporal dependency of each grid-wise feature. However, directly adapting bidirectional temporal attention here as [4, 18, 46, 51] can hardly acquire the ability of causal reasoning, since the predictions will be inevitably dependent on the subsequent frames instead of past conditions. Therefore, we restrict the attention direction by adding a causal attention mask, as shown in the last row of Fig. 3(c), to encourage the model to fully exploit knowledge from past observations and faithfully reason about the future as if in real-world driving. We empirically found that the causality constraint greatly regularizes the predicted frames to be coherent with past frames. Following common practice, we also add temporal bias implemented as relative position embeddings on the time axis [42] to distinguish different frames of a sequence for temporal attention.

Decoupled Spatial Attention. As driving videos feature fast perspective changes, features in a specific grid could vary greatly in different timesteps and are hard to correlate and learn by temporal attention, which suffers from a limited receptive field. In light of this, we introduce spatial attention to propagate each grid feature in spatial axes to aid in gathering information for temporal attention. We implement a decoupled variant of self-attention for its efficiency with linear computational complexity, compared to quadratic full self-attention. As shown in Fig. 3(c), the two decoupled attention layers propagate features in horizontal and vertical axes, respectively.

Deep Interaction. Intuitively, the spatial blocks fine-tuned in stage one refine features of each frame independently towards photorealism, whereas the temporal blocks introduced in stage two align features of all video frames towards coherency and consistency. To further boost the spatiotemporal feature interaction, we interleave the proposed temporal reasoning blocks with the original Transformer blocks in SDXL, *i.e.*, spatial attention, cross attention, and feed-forward network, as shown in Fig. 3(b).

Zero Initialization. Similar to the previous practices [1, 52], for each block that is newly introduced in stage two, we initialized all parameters of its final layer as zero. This avoids disrupting the prior knowledge of the well-trained image generation model in the beginning and stabilizes the training process.

Training. GenAD is trained to predict the future by jointly denoising from the noisy latents with the guidance of past frames and text conditions. We first project T consecutive frames of a video clip into a batch of latents $\mathbf{v} = \{\mathbf{v}^m, \mathbf{v}^n\}$, where the leading m frame latents $\mathbf{v}^m$ are clean, representing historical observations, and other $n = T - m$ frame latents

$\mathbf{v}^n$ indicate the future to be predicted. $\mathbf{v}^n$ are then corrupted to $\mathbf{v}_t^n$ by the forward diffusion process, where t indexes a randomly sampled noise scale. The model is trained to predict the noise of $\mathbf{v}_t^n$ conditioned on observations $\mathbf{v}^m$ and text $\mathbf{c}$. The learning objective of the video prediction model is formulated as follows:

$$\mathcal{L}_{\text{vid}} := \mathbb{E}_{\mathbf{v},\boldsymbol{\epsilon}\sim\mathcal{N}(0,1),\mathbf{c},t}\Big[\|\boldsymbol{\epsilon} - \mathbf{f}_{\theta,\phi}(\mathbf{v}_t^n;\mathbf{v}^m,\mathbf{c},t)\|_2^2\Big], \quad (2)$$

where θ denotes the inherited stage-one model and ϕ represents the newly inserted temporal reasoning blocks. Following [4], we freeze θ and only train the temporal reasoning blocks to avoid perturbing the generation ability of the image generation model and focus on learning temporal dependencies in videos. Notably, only the outputs from the corrupted frames $\mathbf{v}_t^n$ contribute to the training loss while those from condition frames $\mathbf{v}^m$ are ignored.

3.3. Extensions

Relying on the well-trained video prediction capability in driving scenarios, we further exploit the potential of the pre-trained model in action-controlled prediction and planning, which are important for real-world driving systems. Here, we explore the downstream tasks on nuScenes [6] which provides recorded poses.

Action-conditioned Prediction. To make our predictive model controllable with exact ego actions and act as a simulator [25], we fine-tune the model with the paired future trajectory as an additional condition. Specifically, we map the raw trajectory to a high-dimensional feature with Fourier embeddings [44]. After further projection by a linear layer, it is added to the original conditions. Thus, the ego actions are injected into the network through the conditional cross-attention layer in Fig. 3(b).

Planning. By learning to predict the future, GenAD acquires strong representations of complex driving scenes, which can be further exploited for planning. Specifically, we extract spatiotemporal features of two historical frames through the UNet encoder of the *frozen* GenAD, which is nearly half the size of the entire model, and feed them to a multi-layer perceptron (MLP) to predict future waypoints. With the frozen GenAD encoder and a learnable MLP layer, the training process of our planner can be sped up by 3400 times compared to an end-to-end planning model UniAD [22], validating the effectiveness of the learned spatiotemporal feature of GenAD.

4. Experiments

4.1. Setup and Protocols

GenAD is learned in two stages on OpenDV-2K but with different learning objectives (in Sec. 3) and input formats. In stage one, the model takes input (image, text) pairs and is trained on text-to-image generation. We broadcast the command annotation, which is labeled for each 4s video sequence, to all frames included. The model is trained for 300K iterations on 32 NVIDIA Tesla A100 GPUs with a total batch size of 256. In the second stage, GenAD is trained to jointly denoise future latents conditioned on past latents and texts. Its inputs are (video clip, text) pairs where each video clip is 4s at 2Hz. The current version of GenAD is trained on 64 GPUs for 112.5K iterations with a total batch size of 64. The input frames are resized to 256×448 for training in both stages, and the text condition $\mathbf{c}$ is dropped at a probability of $p = 0.1$ to enable classifier-free guidance [17] in sampling, which is commonly used in diffusion models to improve sample quality. More training and sampling details are in Appendix D.

Method	Training Dataset	Pred.	nuScenes FID (↓)	nuScenes FVD (↓)
DriveGAN [25]	nuScenes	✓	73.4	502
DriveDreamer* [45]	nuScenes	✓	52.6	452
DrivingDiffuion* [31]	nuScenes	✗	15.8	332
GenAD-nus (Ours)	nuScenes	✓	**15.4**	244
GenAD (Ours)	OpenDV-2K	✓	**15.4**	**184**

Table 2. **Video generation quality compared to state-of-the-arts trained on nuScenes**. "Pred.": evaluation by future prediction. *: requiring 3D layout inputs.

4.2. Results of Video Prediction Pre-training

Comparison to Recent Video Generation Approaches. We compare GenAD to recent advances on an unseen set with geofencing from OpenDV-YouTube, Waymo [43], KITTI [14], and Cityscapes [10] in a *zero-shot* generation manner. Fig. 4 depicts the qualitative results. Image-to-video models I2VGen-XL [53] and VideoCrafter1 [8] can not strictly follow the given frames to make predictions, yielding poor consistency between the predicted frames and past frames. The video prediction model DMVFN [21] that is trained on Cityscapes suffers from the unfavorable shape distortions in its predictions, especially on the three unseen datasets. In contrast, GenAD exhibits remarkable zero-shot generalization ability and visual quality although *none* of these sets are included in the training.

Comparison to nuScenes Experts. We also compare GenAD with the most recent available driving video generation models which are exclusively trained for nuScenes. Tab. 2 shows that GenAD surpasses all previous methods in both image fidelity (FID) and video coherence (FVD). Specifically, GenAD significantly reduces FVD by **44.5%** compared to DrivingDiffusion [31], without taking 3D future layouts as additional inputs. For fair comparisons, we train a model variant (GenAD-nus) on nuScenes dataset only. We find that although GenAD-nus performs on par

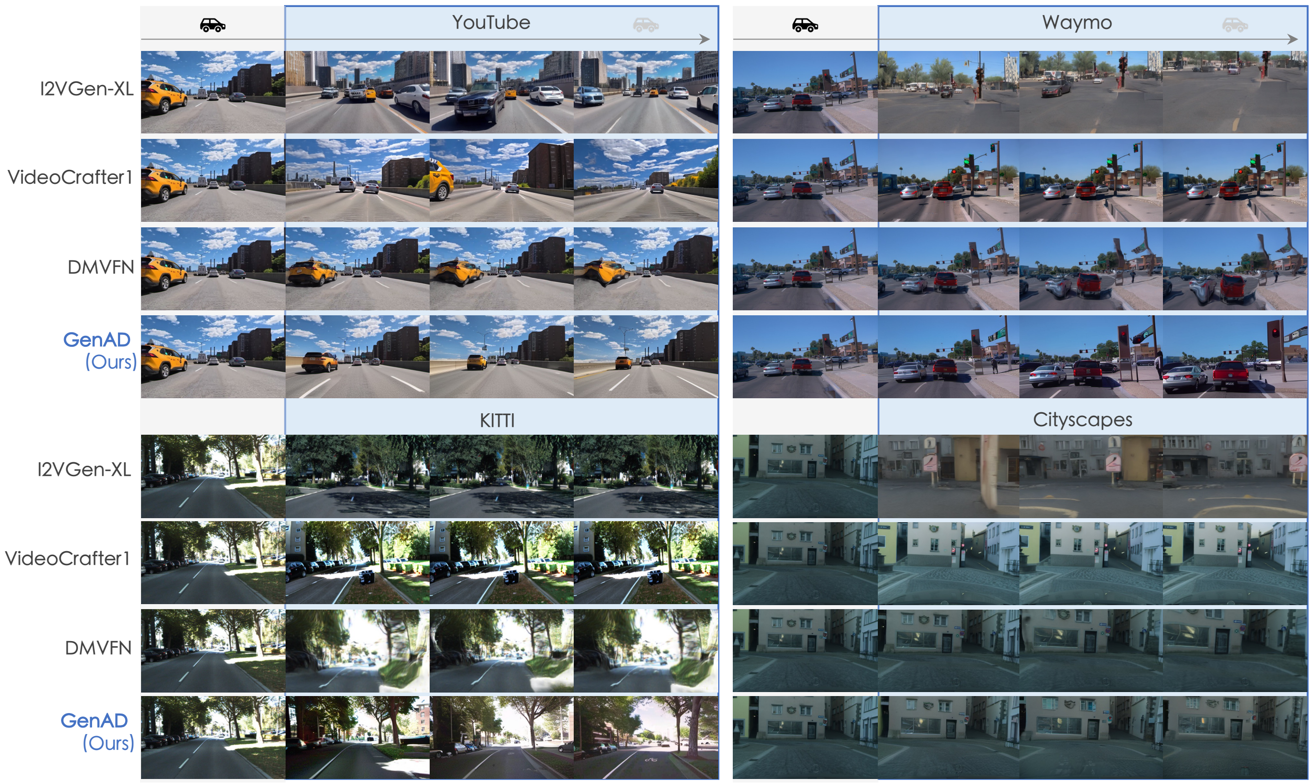

Figure 4. **Task on zero-shot video prediction for unseen scenarios**. We show the generation results (in blue boxes) of different models given the same starting frames. GenAD makes more robust, realistic, and reasonable future predictions on unseen datasets (scenarios). More comparisons and visualizations are shown in Appendix.

Figure 5. **Task on langauge-conditioned prediction**. Given two frames of a rainy scenario in the intersection and three high-level text conditions, GenAD simulates reasonable futures accordingly.

with GenAD on nuScenes, it struggles to generalize to unseen datasets like Waymo, where the generation degrades to the nuScenes visual pattern. In contrast, GenAD trained on OpenDV-2K exhibits strong generalization ability across datasets as shown in Fig. 4.

We provide language-conditioned prediction samples on nuScenes in Fig. 5, where GenAD simulates various futures from the same start following different textual instructions. The impressive generation quality is exhibited in the intricate details of the environment, and the natural transition of ego motion.

Ablation Study. We perform ablations by training each variant on a subset of OpenDV-2K for 75K steps. Starting from the baseline with plain temporal attentions [4, 18], we gradually introduce our proposed components. Notably, by interleaving the temporal blocks with the spatial blocks, the FVD significantly improves (-17%) due to more sufficient spatiotemporal interactions. Both temporal causality and decoupled spatial attention contribute to better CLIP-SIM, improving the temporal consistency between future predictions and the condition frames. To be clear, the slight increase in FID and FVD, shown in fourth and third rows of Tab. 3 respectively, does not faithfully reflect a decline in generation quality as discussed in [4, 5, 37]. The effectiveness of each design is shown in Fig. 6.

4.3. Results of Extensions

Action-conditioned Prediction. We further showcase the performance of the action-conditioned model fine-tuned on nuScenes, GenAD-act, in Fig. 7 and Tab. 4. Given two starting frames and a trajectory $\mathbf{w}$ composed of 6 future way-

Figure 6. **Case study for model designs**. All components help alleviate artifacts and improve the consistency of future predictions.

Method	YouTube FID (↓)	YouTube FVD (↓)	YouTube CLIPSIM (↑)
Baseline	18.32	244.44	0.8405
+ Deep Interaction	17.96	201.69	0.8409
+ Temporal Causality	**16.54**	207.45	0.8550
+ Decoupled Spatial Attn.	17.67	**189.54**	**0.8652**

Table 3. **Ablation on model designs in GenAD**. All proposed designs contribute to the final performance.

Method	Condition	nuScenes Action Prediction Error (↓)
Ground truth	-	0.90
GenAD	text	2.54
GenAD-act	text + traj.	**2.02**

Table 4. **Task on action-conditioned prediction**. Compared to GenAD with text conditions only, GenAD-act enables more precise future predictions that follow the action condition.

points, GenAD-act imagines 6 future frames following the trajectory sequence. To evaluate the consistency between the input trajectory $\mathbf{w}$ and predicted frames, we establish an inverse dynamics model (IDM) on nuScenes as the evaluator, which projects a video sequence into a corresponding ego trajectory. We leverage the IDM to translate predicted frames into the trajectory $\hat{\mathbf{w}}$, and calculate the L2 distance between $\mathbf{w}$ and $\hat{\mathbf{w}}$ as the Action Prediction Error. Specifically, GenAD-act substantially reduces the Action Prediction Error by 20.4% compared to GenAD with text condition, allowing for more accurate future simulations.

Planning Results. Tab. 5 depicts the planning results on nuScenes where ground truth poses for the ego vehicle are available. By freezing GenAD encoder and only optimizing

Figure 7. **Task on action-conditioned prediction (simulation)**. Given the same starting frames and different future trajectories (shown in yellow dots in the first column), GenAD-act can simulate diverse futures following different ego intentions. More visualizations are in Appendix.

Method	# Trainable Params.	nuScenes ADE (↓)	nuScenes FDE (↓)
ST-P3* [20]	10.9M	2.65	3.73
UniAD* [22]	58.8M	1.03	1.65
GenAD (Ours)	0.8M	1.23	2.31

Table 5. **Task on open-loop planning**. A lightweight MLP with *frozen* GenAD gets competitive planning results with 73× fewer trainable parameters and front-view image alone. *: multi-view inputs. Evaluation protocols are aligned with UniAD [22].

an additional MLP on top of it, the model can effectively learn to plan. Notably, by pre-extracting image features through the UNet encoder of GenAD, the entire learning process for planning adaptation takes only 10 minutes on a single NVIDIA Tesla V100 device, which is 3400 times more efficient than the training of the UniAD planner [22].

5. Limitations and Discussion

We study the system-level development of GenAD, a large-scale generalized video predictive model for autonomous driving. We also validate the adaptation of the learned representation of GenAD to driving tasks, *i.e.*, learning a "world model" and motion planning. Although we obtain improved generalization to open domains, the increased model capacity poses challenges in both training efficiency and real-time deployment. We envision the unified video prediction task will serve as a scalable objective for future research on representation learning and policy learning. Another interesting direction involves distilling the encoded knowledge for a wider range of downstream tasks [29].

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, 2022. 2, 5

[2] Mohammadhossein Bahari, Saeed Saadatnejad, Ahmad Rahimi, Mohammad Shaverdikondori, Amir Hossein Shahidzadeh, Seyed-Mohsen Moosavi-Dezfooli, and Alexandre Alahi. Vehicle trajectory prediction works, but not everywhere. In *CVPR*, 2022. 2

[3] Fan Bao, Chongxuan Li, Jiacheng Sun, and Jun Zhu. Why are conditional generative models better than unconditional ones? *arXiv preprint arXiv:2212.00362*, 2022. 4

[4] 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, 5, 6, 7

[5] 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. In *NeurIPS*, 2022. 7

[6] 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. 3, 4, 6

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

[8] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, Chao Weng, and Ying Shan. VideoCrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023. 6

[9] Li Chen, Penghao Wu, Kashyap Chitta, Bernhard Jaeger, Andreas Geiger, and Hongyang Li. End-to-end autonomous driving: Challenges and frontiers. *arXiv preprint arXiv:2306.16927*, 2023. 2

[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, 6

[11] Yilun Dai, Mengjiao Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Josh Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning universal policies via text-guided video generation. In *NeurIPS*, 2023. 2

[12] Thierry Deruyttere, Simon Vandenhende, Dusan Grujicic, Luc Van Gool, and Marie-Francine Moens. Talk2Car: Taking control of your self-driving car. In *EMNLP*, 2019. 3, 4

[13] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, 2021. 4

[14] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The KITTI dataset. *IJRR*, 2013. 3, 6

[15] Agrim Gupta, Stephen Tian, Yunzhi Zhang, Jiajun Wu, Roberto Martín-Martín, and Li Fei-Fei. MaskViT: Masked visual pre-training for video prediction. In *ICLR*, 2023. 2

[16] Niklas Hanselmann, Katrin Renz, Kashyap Chitta, Apratim Bhattacharyya, and Andreas Geiger. KING: Generating safety-critical driving scenarios for robust imitation via kinematics gradients. In *ECCV*, 2022. 2

[17] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS Workshops*, 2021. 6

[18] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. In *NeurIPS*, 2022. 2, 5, 7

[19] 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

[20] Shengchao Hu, Li Chen, Penghao Wu, Hongyang Li, Junchi Yan, and Dacheng Tao. ST-P3: End-to-end vision-based autonomous driving via spatial-temporal feature learning. In *ECCV*, 2022. 8

[21] Xiaotao Hu, Zhewei Huang, Ailin Huang, Jun Xu, and Shuchang Zhou. A dynamic multi-scale voxel flow network for video prediction. In *CVPR*, 2023. 6

[22] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, Lewei Lu, Xiaosong Jia, Qiang Liu, Jifeng Dai, Yu Qiao, and Hongyang Li. Planning-oriented autonomous driving. In *CVPR*, 2023. 6, 8

[23] 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

[24] Jinkyu Kim, Teruhisa Misu, Yi-Ting Chen, Ashish Tawari, and John Canny. Grounding human-to-vehicle advice for self-driving vehicles. In *CVPR*, 2019. 3, 4

[25] Seung Wook Kim, Jonah Philion, Antonio Torralba, and Sanja Fidler. DriveGAN: Towards a controllable high-quality neural simulation. In *CVPR*, 2021. 2, 6

[26] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *ICCV*, 2023. 2

[27] Yann LeCun. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62, 2022. 2

[28] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 2015. 2

[29] Daiqing Li, Huan Ling, Amlan Kar, David Acuna, Seung Wook Kim, Karsten Kreis, Antonio Torralba, and Sanja Fidler. DreamTeacher: Pretraining image backbones with deep generative models. In *ICCV*, 2023. 8

[30] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023. 2, 3, 4

[31] Xiaofan Li, Yifu Zhang, and Xiaoqing Ye. DrivingDiffusion: Layout-guided multi-view driving scene video generation with latent diffusion model. *arXiv preprint arXiv:2310.07771*, 2023. 6

[32] Zhenyu Li, Zehui Chen, Ang Li, Liangji Fang, Qinhong Jiang, Xianming Liu, and Junjun Jiang. Unsupervised domain adaptation for monocular 3d object detection via self-training. In *ECCV*, 2022. 3

[33] Jiachen Lu, Ze Huang, Jiahui Zhang, Zeyu Yang, and Li Zhang. Wovogen: World volume-aware diffusion for controllable multi-camera driving scene generation. *arXiv preprint arXiv:2312.02934*, 2023. 2

[34] Jiageng Mao, Minzhe Niu, Chenhan Jiang, Hanxue Liang, Xiaodan Liang, Yamin Li, Chao Ye, Wei Zhang, Zhenguo Li, Jie Yu, Hang Xu, and Chunjing Xu. One million scenes for autonomous driving: ONCE dataset. In *NeurIPS Datasets and Benchmarks*, 2021. 3, 4

[35] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2

[36] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *NeurIPS*, 2022. 3, 4

[37] 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. *arXiv preprint arXiv:2307.01952*, 2023. 2, 4, 7

[38] Xiaojuan Qi, Zhengzhe Liu, Qifeng Chen, and Jiaya Jia. 3D motion decomposition for rgbd future dynamic scene synthesis. In *CVPR*, 2019. 2

[39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 2, 4

[40] Vasili Ramanishka, Yi-Ting Chen, Teruhisa Misu, and Kate Saenko. Toward driving scene understanding: A dataset for learning driver behavior and causal reasoning. In *CVPR*, 2018. 3, 4

[41] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2, 4

[42] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. *arXiv preprint arXiv:1803.02155*, 2018. 5

[43] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott M. Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. In *CVPR*, 2020. 3, 6

[44] Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. In *NeurIPS*, 2020. 6

[45] 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, 6

[46] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, Yuwei Guo, Tianxing Wu, Chenyang Si, Yuming Jiang, Cunjian Chen, Chen Change Loy, Bo Dai, Dahua Lin, Yu Qiao, and Ziwei Liu. LAVIE: High-quality video generation with cascaded latent diffusion models. *arXiv preprint arXiv:2309.15103*, 2023. 5

[47] 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. *arXiv preprint arXiv:2311.17918*, 2023. 2

[48] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, Deva Ramanan, Peter Carr, and James Hays. Argoverse 2: Next generation datasets for self-driving perception and forecasting. In *NeurIPS Datasets and Benchmarks*, 2021. 3

[49] Daniel M Wolpert, Zoubin Ghahramani, and Michael I Jordan. An internal model for sensorimotor integration. *Science*, 1995. 2

[50] Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, Rui Zheng, Xiaoran Fan, Xiao Wang, Limao Xiong, Yuhao Zhou, Weiran Wang, Changhao Jiang, Yicheng Zou, Xiangyang Liu, Zhangyue Yin, Shihan Dou, Rongxiang Weng, Wensen Cheng, Qi Zhang, Wenjuan Qin, Yongyan Zheng, Xipeng Qiu, Xuanjing Huang, and Tao Gui. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864*, 2023. 2

[51] David Junhao Zhang, Jay Zhangjie Wu, Jia-Wei Liu, Rui Zhao, Lingmin Ran, Yuchao Gu, Difei Gao, and Mike Zheng Shou. Show-1: Marrying pixel and latent diffusion models for text-to-video generation. *arXiv preprint arXiv:2309.15818*, 2023. 5

[52] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 5

[53] Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and

Jingren Zhou. I2VGen-XL: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv preprint arXiv:2311.04145*, 2023. 6

[54] Ruizhao Zhu, Peng Huang, Eshed Ohn-Bar, and Venkatesh Saligrama. Learning to drive anywhere. In *CoRL*, 2023. 2