

# Doe-1: Closed-Loop Autonomous Driving with Large World Model

Wenzhao Zheng<sup>\*,†</sup> Zetian Xia<sup>\*</sup> Yuanhui Huang Sicheng Zuo Jie Zhou Jiwen Lu

Department of Automation, Tsinghua University, China

wenzhao.zheng@outlook.com; xiazt21@mails.tsinghua.edu.cn

Project Page: <https://wzzheng.net/Doe>

Large Driving Models: <https://github.com/wzzheng/LDM>

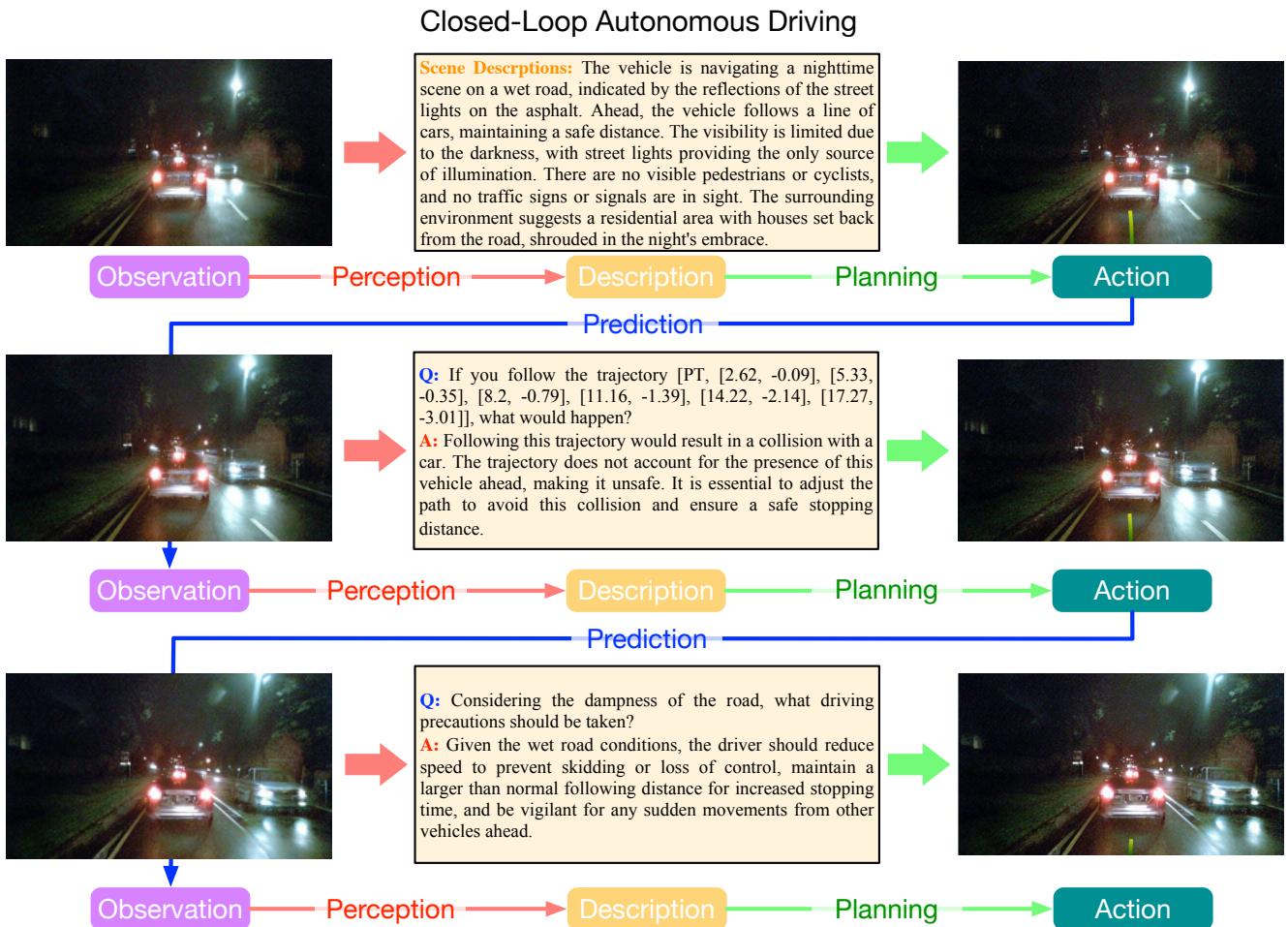


Figure 1. **Visualizations of our Doe-1 for closed-loop autonomous driving on nuScenes [11].** We propose a large driving world model (**Doe-1**) to achieve unified generative closed-loop autonomous driving. We model perception, prediction, and planning as the transitions of observation→description, description→action, and action→observation, respectively. **Doe-1** accomplishes perception, planning, and prediction in a unified autoregressive generative framework and achieves closed-loop end-to-end autonomous driving for the first time.

## Abstract

End-to-end autonomous driving has received increasing attention due to its potential to learn from large amounts of data. However, most existing methods are still open-

loop and suffer from weak scalability, lack of high-order interactions, and inefficient decision-making. In this paper, we explore a closed-loop framework for autonomous driving and propose a large **D**riving **W**orld **m**odel (**Doe-1**) for unified perception, prediction, and planning. We formulate autonomous driving as a next-token generation problem and use multi-modal tokens to accomplish differ-

\*Equal contribution. †Project leader.

ent tasks. Specifically, we use free-form texts (i.e., scene descriptions) for perception and generate future predictions directly in the RGB space with image tokens. For planning, we employ a position-aware tokenizer to effectively encode action into discrete tokens. We train a multi-modal transformer to autoregressively generate perception, prediction, and planning tokens in an end-to-end and unified manner. Experiments on the widely used nuScenes dataset demonstrate the effectiveness of **Doe-1** in various tasks including visual question-answering, action-conditioned video generation, and motion planning. Code: <https://github.com/wzzheng/Doe>.

## 1. Introduction

The emergence of GPT series [8, 9, 60] stimulates the rapid development of large models with various functions, including language modeling [2, 73, 74], visual understanding [3, 46, 47, 57], and decision-making [7, 37, 53]. The key to the success of large models is scaling up model sizes and training data [34]. When designing a model, the scalability advocates large representation capacity (e.g., transformers [17, 49, 75]) over well-designed inductive biases (e.g., convolution neural networks [23, 29]) to improve the upper bound of performance.

To build large models for autonomous driving, some methods directly apply large language models (LLMs) [13, 53, 65, 67] or vision-language models (VLMs) [62, 66, 71, 80, 81, 89, 98] for motion planning [71] or scene question-answering [55, 79]. They usually align the inputs with texts (e.g., Q-Former [39]) and output language descriptions of the planning results [53]. However, LLMs are known to share the hallucination issue [21, 42, 48], hindering the interpretability and safety of autonomous driving. To avoid this, others follow the well-tested pipeline of perception, prediction, and planning for autonomous driving and explore a scalable end-to-end model [27, 28, 33, 72, 93, 100] to jointly accomplish them. Though promising, most existing methods are still open-loop and suffer from several issues. **1) Weak scalability.** They use manually designed scene representation which cannot provide comprehensive information for downstream tasks. **2) Lack of high-order interactions.** They predict future scenes without considering the planned ego trajectory. **3) Inefficient decision-making.** They usually plan several steps ahead yet practically only use the first step to execute.

To address these, we propose a closed-loop large Driving wOrld modEl (**Doe-1**) for unified perception, prediction, and planning without intermediate latent scene representations, as shown in Figure 2. We cast autonomous driving as a scene evolution problem and represent each scene with observation, description, and action tokens. We then formulate the conventional per-

ception, planning, anaction-conditioned video generationd prediction as transitions between multi-modal tokens, i.e., observation→description, description→action, and action→observation, respectively. We then employ a generative autoregressive world model to model this evolution with next-token prediction. Concretely, we focus on vision-centric autonomous driving and adopt RGB images as observations. We tokenize images using the image vector-quantized variational autoencoder [56]. We employ free-form texts as the scene descriptions and also feed the model with question-answering pairs for on-demand perception. We represent action with displacement in the bird’s eye view (BEV) space and employ a position-aware tokenizer to encode it into discrete tokens. Our **Doe-1** then generates the observation, description, and action tokens of the next temporal frame sequentially and iteratively, which can be efficiently trained with a simple autoregressive objective. We conduct extensive experiments on the nuScenes dataset [11] to evaluate the capability of our **Doe-1**. With different prompt settings, we demonstrate that our **Doe-1** successfully achieves various tasks without finetuning, including visual question-answering, controlled image generation, and motion planning.

## 2. Related Work

**Large Models for Autonomous Driving.** The success of GPT series [8, 9, 60] confirms the power of model and data scaling and promotes the rapid development of large language models (LLMs) [2, 73, 74], which demonstrate impressive performance on a wide range of language tasks. Early methods combining LLMs with autonomous driving use ground-truth labels [13, 65, 67] or pre-trained perception models [53] to obtain detection and map results. They then translate them into text descriptions as inputs and leverage the reasoning ability of LLMs to output the planned trajectory in text. To enable LLMs to process images, large vision-language models (VLMs) [3, 46, 47, 57] usually pre-align the image representations with text space and then jointly finetune the overall model. This facilitates the emergence of end-to-end planning methods which take as inputs images and directly output the planned trajectory. Further methods explore the use of chain-of-thoughts [55, 80], instruction tuning [97], or additional visual question answering [79] to improve the interpretability. Still, existing methods struggle with the hallucination issue [21, 42, 48] commonly possessed by LLMs and VLMs and thus lack robustness and safety, which are critical for autonomous driving. To alleviate this issue, we follow the well-tested perception, prediction, and planning pipeline and propose a large driving world model (**Doe-1**) to unify all these tasks for more robust autonomous driving.

**World Models for Autonomous Driving.** The conventional world model aims to predict the next observation of

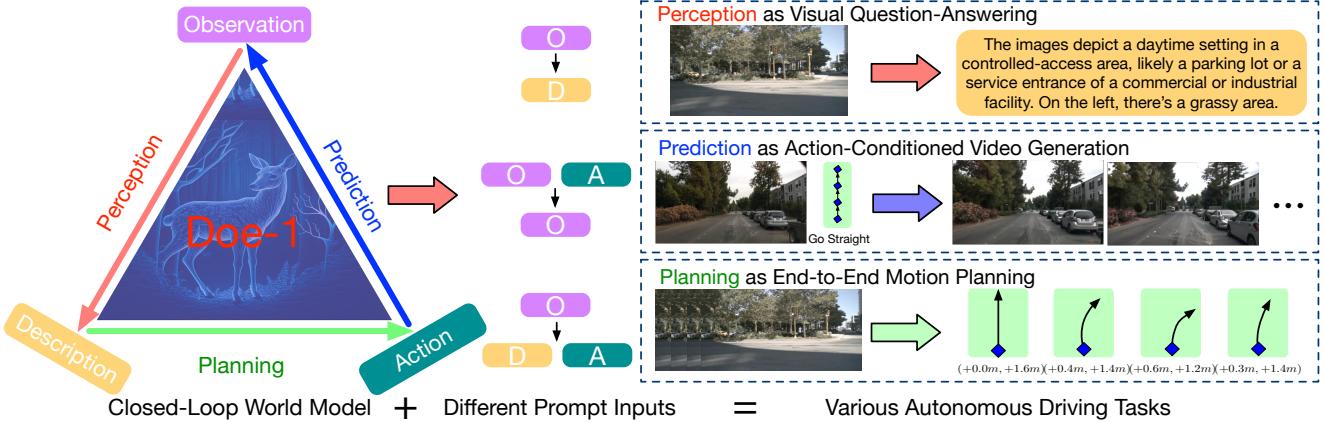


Figure 2. **Overview of the proposed Doe-1.** We formulate autonomous driving as a unified next-token generation problem and use observation, description, and action tokens to represent each scene. Without additional fine-tuning, **Doe-1** accomplishes various tasks by using different input prompts, including visual question-answering, controlled image generation, and end-to-end motion planning.

the scene given the action of the ego agent [22, 69]. Most existing methods follow this definition and focus on controlled generation in the image [20, 52, 77, 85, 90], LiDAR [35, 36, 54, 86] or 3D occupancy [78, 99] space. Vision-based methods usually utilize the capability of pre-trained image diffusion models (e.g., StableDiffusion [63]) and finetune them on the driving scenarios to predict action-conditioned futures [19, 40, 82, 91]. The other methods explore more diverse architectures (e.g., autoregressive transformer [25, 88, 99], gated recurrent unit [5], or autoencoder [1, 92]) to generate future scenes. However, most methods cannot be directly applied to trajectory planning and are usually used to compute rewards to train another planner with reinforcement learning [20]. OccWorld [99] generalizes the concept of world model and predicts the joint evolution of 3D occupancy and the ego vehicle. Still, it requires pre-trained perception models to obtain 3D occupancy and is limited by its representation ability. Differently, the proposed **Doe-1** employs one closed-loop unified framework for perception, prediction, and planning as a generalized autonomous driving model.

**End-to-End Autonomous Driving.** Perception, prediction, and planning are the well-tested pipeline for autonomous driving and are conventionally realized with separate modules [14, 30, 58, 64]. Recent methods explore the end-to-end design that jointly trains all the modules with a shared objective of trajectory planning [27, 28, 33, 72, 93, 100]. The unified framework equips them with more feasibility to scale up and more potential to achieve native large models for autonomous driving. Early methods [27, 28] adopt bird’s eye view as the scene representation to pass information throughout the model. Subsequent methods explore the use of 3D occupancy [72, 99] for more details or sparse queries [68, 96] for better efficiency. However, the manually designed scene representations might not be optimal and the induced inductive biases limit the capability of the model, which is critical for scale-up. They also suf-

fer from the lack of high-order interactions and inefficient decision-making. In this paper, we propose the first closed-loop autonomous driving model, **Doe-1**, for more scalable autonomous driving, which achieves perception with free-form texts and predicts futures directly in the image space to reduce information loss and enlarge the model capacity.

### 3. Proposed Approach

#### 3.1. Closed-Loop Autonomous Driving

Autonomous driving is a long-standing task to apply artificial intelligence to the real world, which aims to plan the future actions  $\{a\}$  for the ego vehicle from scene observations  $\{o\}$ . Several attempts [6, 12, 15, 16, 59] explore the use of deep neural networks to directly model the mapping from observations  $\{o\}$  to actions  $\{a\}$ :

$$\{o^{t-H}, \dots, o^t\} \rightarrow \{a^t, \dots, a^{t+F-1}\}, \quad (1)$$

where  $t$  is the current time stamp, and  $H$  and  $F$  is the number of concerned history and future frames, respectively. The black-box nature of deep neural networks makes the decision-making process less transparent and trustworthy.

Recent methods utilize the reasoning ability of large language models (LLMs) or vision-language models (VLMs) to improve the interpretability of the planning results [13, 55, 65, 67, 97]. They align the visual features with the text space by direct transcription [65] or learnable projection [66]. Some methods further use additional text descriptions  $d$  as an auxiliary task [79] to further alleviate the hallucination issue [48] of large models:

$$\{o^{t-H}, \dots, o^t\} \xrightarrow{\text{LLM/VLM}} d^t, \{a^t, \dots, a^{t+F-1}\}. \quad (2)$$

Still, it is doubtful whether LLMs/VLMs truly understand the 4D dynamics and interactions among traffic agents without fine-grained understanding and prediction of the scene.

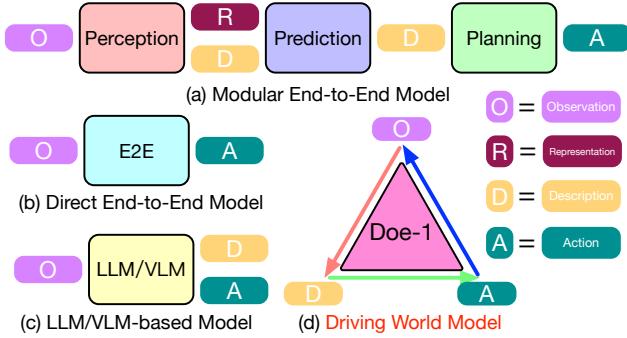


Figure 3. **Comparisons of different paradigms.** (a) The modular end-to-end model performs perception, prediction, and planning sequentially and is the most popular pipeline for autonomous driving. (b) The direct end-to-end model directly outputs the planned action given sensor inputs. (c) The LLM/VLM-based model exploits the reasoning ability of LLMs/VLMs to output actions. (d) The proposed driving world model (**Doe-1**) predicts the evolutions between observations, descriptions, and actions to achieve close-loop end-to-end autonomous driving.

To build more trustworthy autonomous driving systems, the mainstream methods (including modern end-to-end models [27, 28, 33, 72, 93, 100]) employ a “divide and conquer” strategy and decompose autonomous driving into sequential perception  $f$ , prediction  $g$ , and planning  $h$ :

$$\begin{aligned} \{\mathbf{o}^{t-H}, \dots, \mathbf{o}^t\} &\xrightarrow{f_1} \mathbf{r}^t \xrightarrow{f_2} \mathbf{d}^t \xrightarrow{g} \\ \{\mathbf{d}^{t+1}, \dots, \mathbf{d}^{t+F}\} &\xrightarrow{h} \{\mathbf{a}^t, \dots, \mathbf{a}^{t+F-1}\}, \end{aligned} \quad (3)$$

where  $\mathbf{r}$  and  $\mathbf{d}$  denotes the scene representation and scene description, respectively. The scene representation  $\mathbf{r}$  is a set of continuous features in the 3D space (e.g., bird’s eye view [41, 43], voxels [84], tri-perspective view [30, 101], 3D Gaussians [31, 32, 87], dense points [18]), while  $\mathbf{d}$  provides high-level refined descriptions (e.g., detected bounding boxes, constructed map) of the scene. The design of  $\mathbf{d}$  introduces prior knowledge of the most important factors for decision-making and thus can improve the robustness of the model. We compare different paradigms in Figure 3.

Despite its satisfactory performance, we argue that this well-known pipeline still suffers from several issues including weak scalability, lack of high-order interactions, and inefficient decision-making.

**Weak Scalability.** One advantage of end-to-end driving is its potential to scale up model sizes and fit more training data [99]. However, most existing methods use manually designed scene representations  $\mathbf{r}$  and descriptions  $\mathbf{d}$ , which induce inductive bias and prior knowledge. Though this could be beneficial with little available training data, it limits the performance upper bound when scaling up. For example, some factors (e.g., the pose of a human, the rear light of a vehicle) can be crucial in certain corner cases, and discarding them will lead to inaccurate decisions.

**Lack of High-Order Interactions.** The current pipeline

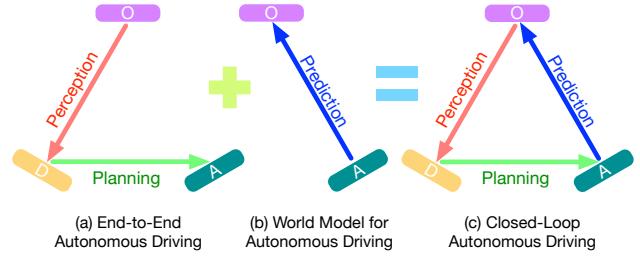


Figure 4. **Illustration of the proposed closed-loop autonomous driving paradigm.** (a) Existing end-to-end autonomous driving methods (e.g., UniAD [28], GenAD [100]) usually perform perception first and then make decisions according to the perceived descriptions. (b) Existing world models for autonomous driving (e.g., DriveDreamer [82], OccWorld [99]) predict future observations based on the current actions. (c) Close-loop autonomous driving combines the two paradigms to construct a closed loop.

(3) usually predicts the future  $F$  scenes at once before planning, assuming that the ego action **Doe-1**s not affect the environment development. However, in the interactive scenarios of autonomous driving, the future movements of other agents are highly dependent on the ego action. For example, vehicles behind usually will respond to the deceleration of the ego car. Therefore, it is not feasible to accurately predict the future without conditions of the ego action.

**Inefficient Decision-Making.** Most existing methods predict and plan several steps ahead, as stated in (3). However, in practice, driving models only take the planned next-frame action and re-plan future actions given new observations, resulting in inefficiency and redundancy of multi-step planning. Therefore, we think the model needs to be capable of predicting possible futures for possible ego actions but plan the instant (i.e., next-frame) action without explicit deduction. For example, we humans have the ability to predict how the environment respond to our actions, yet we usually make decisions instinctly and do not explicitly ponder the chain of reactions when driving.

To address these issues, this paper explores a new closed-loop autonomous driving paradigm, as shown in Figure 4 and proposes a large driving world model, **Doe-1**, to formulate autonomous driving as an autoregressive world evolution of multimodal states:

$$\mathbf{o}^t \xrightarrow{\text{Doe-1}} \mathbf{d}^t \xrightarrow{\text{Doe-1}} \mathbf{a}^t \xrightarrow{\text{Doe-1}} \mathbf{o}^{t+1} \xrightarrow{\text{Doe-1}} \mathbf{d}^{t+1} \xrightarrow{\text{Doe-1}} \dots \quad (4)$$

Using self-attention, **Doe-1** directly accesses the observations and predicts future scene evolutions from the observation space without intermediate scene representations. **Doe-1** makes predictions with ego action as the condition. Through autoregressive generation, **Doe-1** predicts multi-step futures yet only makes instant planning each time.

### 3.2. Doe-1: A Large Driving World Model

**Overview.** The overall framework of **Doe-1** is depicted in Figure 5. We use Chameleon architecture [70] as a uni-

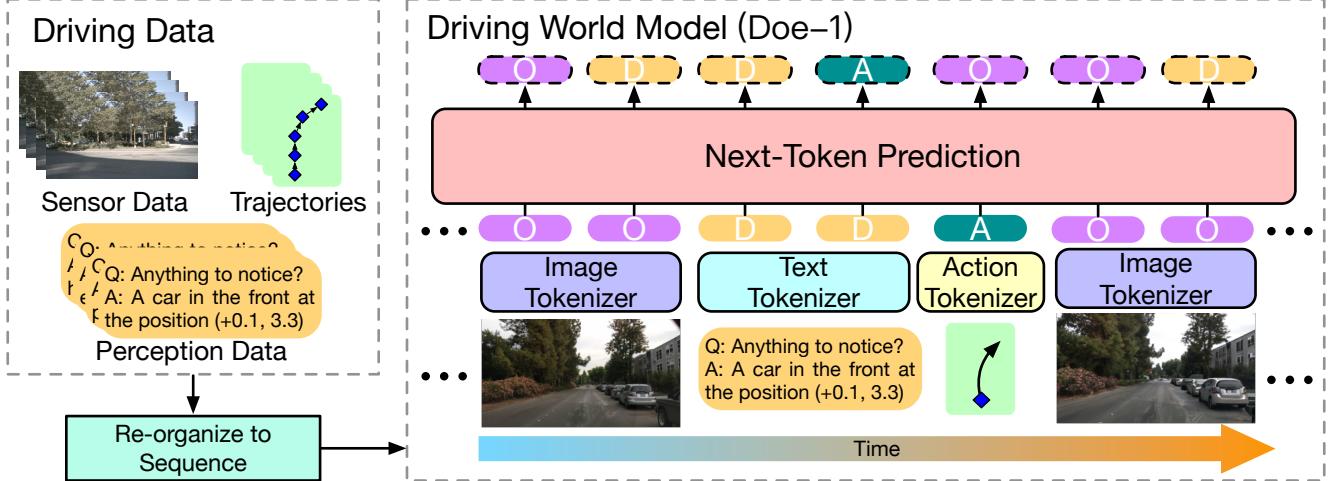


Figure 5. **Framework of the proposed Doe-1.** We first re-organize the training dataset into a temporal sequence of sensor data (image), perception data (texts), and action data (position of the next frame). We then use image, text, and action tokenizers to encode them into discrete tokens to construct a 1D token sequence. We then use a transformer-based architecture to autoregressively model this sequence and use the next-token prediction as the training objective.

fied multimodal autoregressive model to generate predictions. Following 4, the observations  $\mathbf{o} \in \mathbb{R}^{3 \times h \times w}$ , descriptions  $\mathbf{d} \in \mathbb{Z}^l$  and actions  $\mathbf{a} \in \mathbb{R}^{p \times 3}$  are tokenized into discrete space and simply concatenated into sequences, where  $(h, w)$  is the shape of observation images,  $l$  is the length of string  $\mathbf{d}$  and  $p$  is the number of motions making  $\mathbf{a}$  contains  $p$  motion parameters in 2D space. Using the past sequences as prompt, the unified model will predicted the next token iteratively. Specifically, given the tokenized observations  $\mathbf{T}_o^i \in \mathbb{Z}^{d_o}$ , the high-level descriptions  $\mathbf{T}_d^i \in \mathbb{Z}^{d_d}$  and the actions  $\mathbf{T}_a^i \in \mathbb{Z}^{d_a}$  from time  $i$  where  $0 \leq i \leq t$ , the model gives predictions tokens at time  $t + 1$  as follows:

$$\begin{aligned}\mathbf{T}_o^{t+1} &= \arg \max_{\mathbf{T}_o} p(\mathbf{T}_o | \mathbf{T}_o^0, \mathbf{T}_d^0, \mathbf{T}_a^0, \dots, \mathbf{T}_a^t), \\ \mathbf{T}_d^{t+1} &= \arg \max_{\mathbf{T}_d} p(\mathbf{T}_d | \mathbf{T}_o^0, \mathbf{T}_d^0, \mathbf{T}_a^0, \dots, \mathbf{T}_a^t, \mathbf{T}_o^{t+1}), \\ \mathbf{T}_a^{t+1} &= \arg \max_{\mathbf{T}_a} p(\mathbf{T}_a | \mathbf{T}_o^0, \mathbf{T}_d^0, \mathbf{T}_a^0, \dots, \mathbf{T}_o^{t+1}, \mathbf{T}_d^{t+1}).\end{aligned}\quad (5)$$

The  $\mathbf{o}_{t+1}$ ,  $\mathbf{d}_{t+1}$ ,  $\mathbf{a}_{t+1}$  are then decoded from the predicted tokens. To encode the sequence  $\{\mathbf{o}_i, \mathbf{d}_i, \mathbf{a}_i\}_{i=0}^t$  into discrete tokens, an observation tokenizer, a description tokenizer and an action tokenizer are used.

**Observation Tokenzier.** Observations are images in the RGB space  $\mathbb{R}^{3 \times h \times w}$ . To map the image  $\mathbf{o}$  to the tokens  $\mathbf{T}_o$  in a discrete space, we use the pretrained image tokenizer from Lumina-mGPT [45], which encodes an image into 1024 discrete tokens from a codebook of size 8192 and adds special tokens to indicate the shape of the image.

**Description Tokenzier.** In interactive scenarios, a description  $\mathbf{d}$  are in a char space  $\mathbb{Z}^l$ , where  $l$  denotes the length of string  $\mathbf{d}$ . We use the BPE tokenizer from Chameleon [70] to encode texts of descriptions into a discrete space.

Treating floating-point numbers as text results in the loss

of their original metric properties in the Euclidean space. For the floating-point number sequences in the texts of descriptions (such as the distance to an object from the ego), we round them to a resolution of 0.02 meters, then the discretized floating point numbers are mapped to 4000 discrete tokens in turn. In order to maintain the original metric properties of numbers after embedding, we use the Sinusoidal Positional Encoding as the embedded features of these tokens to ensure that floating-point numbers with small differences still maintain a small distance after being embedded in the high-dimensional space. For a floating-point  $x \in \mathbb{R}$ , the token  $T$  and the embedding of  $T$  is computed as:

$$\begin{aligned}T(x) &= \text{round}\left(\frac{1}{r} \times x\right) + b, \\ \text{Embedding}(x, 2i) &= \sin\left(\frac{T(x)}{\text{scale}^{2i/d}}\right), \\ \text{Embedding}(x, 2i + 1) &= \cos\left(\frac{T(x)}{\text{scale}^{(2i+1)/d}}\right),\end{aligned}\quad (6)$$

where  $r$  denotes the resolution,  $b$  denotes a bias constant determined by the codebook size and  $\text{scale}$  is a scaling constant determined by the range of  $x$ . The positional embedding is set to be non-learnable to ensure that the floating-point numbers which are not seen in the training phase are still aligned in the embedding space.

**Action Tokenzier.** The motion  $\mathbf{z}$  denotes the movement of a vehicle between two frames and is defined as a three-element array  $\mathbf{z} = (d_x, d_y, \theta)$  containing displacements along two horizontal axes and a yaw angle, and an action  $\mathbf{a} \in \mathbb{R}^{p \times 3}$  given by the model concatenates  $p$  consecutive moves to denote the future trajectory in the next  $p$  frames. To encode actions into discrete tokens, we use positional encoding similar to above and simply adjust the scaling constant. Special tokens are inserted as anchors at

the beginning and end of an action.

**Generative Architecture.** By aligning the multimodal generation, we can readily scale up the unified model. For any modality, the model follows a unified autoregressive paradigm. When predicting actions, the model generates several motion predictions for multiple future frames at once by giving  $p(\mathbf{a}_t|\mathbf{T})$  based on the previous token sequence  $\mathbf{T}$ , and  $\mathbf{a}_t$  will affect the prediction of subsequent tokens. In order to avoid the accumulation of errors between predictions at different frames, last  $p-1$  motions of the generated predictions are masked when predictions  $\mathbf{a} \in \mathbb{R}^{p \times 3}$  are generated, with the first prediction retained to provide a reference for generating subsequent observations:

$$\begin{aligned}\mathbf{a}_t &= (\mathbf{z}_t, \mathbf{z}_{t+1}, \dots, \mathbf{z}_{t+p}), \\ \text{masked}(\mathbf{a}_t) &= (\mathbf{z}_t, \mathbf{m}_1, \dots, \mathbf{m}_p),\end{aligned}\quad (7)$$

where  $\mathbf{m}_i$  denotes the  $i$ -th mask array which will be encoded into mask tokens.

### 3.3. Applications of Doe-1

As a unified multimodal world model, **Doe-1** supports inputs from multiple modalities and automatically predicts the next modality, facilitating the applications to different tasks simply by altering the prompt. Figure 6 introduces the application of **Doe-1** to different tasks of visual question-answering, motion planning, and action-conditioned video generations as examples. By designing prompt sequences, **Doe-1** can be transferred to other multimodal tasks.

**Visual Question-Answering.** Given the observation  $\mathbf{o}$ , the model is required to generate a precise description  $\mathbf{d}_o$  of the observation. Furthermore, to assess the model’s understanding of the scene and the perception ability, we require the model to complete interactive question-answering tasks based on the given observations and the generated descriptions. These tasks include identifying objects in the scene that may impact driving, describing the environment, and checking the plausibility of a given driving trajectory, among others, which requires the model to have as comprehensive a description capability of the scene as possible.

**Action-Conditioned Video Generation.** We hope that the world model, **Doe-1**, can enable closed-loop interactive simulation of driving scenarios without the need for feedback from real-world interactions. This requires **Doe-1** to be able to predict the changes in observations after taking an action. Given a sequence of observations with descriptions  $\{\mathbf{o}_j, \mathbf{d}_j\}_{j=0}^t$  and the actions history  $\{\mathbf{a}_j\}_{j=0}^t$ , the model is required to generate the observations  $\mathbf{o}_{t+1}$ .

In the action-conditioned video generation task, for the observations, we only provide the model with the image at time step 0, and the model is required to iteratively generate the next  $p$  frames of images based on the given actions.

**Motion Planning.** Given a sequence of observations  $\{\mathbf{o}_j\}_{j=0}^t$  and history motions  $\{\mathbf{z}_j\}_{j=0}^{t-1}$ , the model gener-

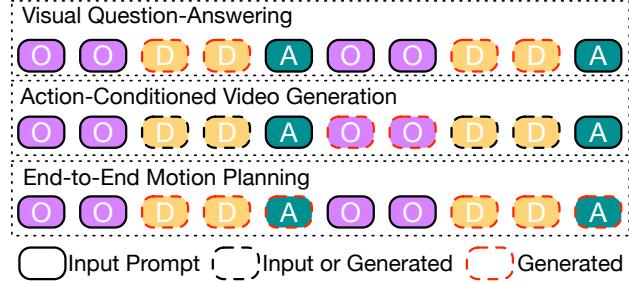


Figure 6. **Illustration of different prompt settings to apply our Doe-1 to different tasks.** For visual question-answering, we use observation tokens as prompt input and generate description tokens. For action-conditioned video generation, we use the observation and action tokens of the current frame as inputs and generate observation tokens of the next frame. For end-to-end motion planning, we use observation tokens as prompts to generate the description and action tokens.

ates descriptions  $\mathbf{d}_t$  and plan future motions  $\{\mathbf{z}_{t+i}\}_{i=0}^p$  efficiently, where  $p$  denotes the number of predicted frames.

By combining the aforementioned three tasks, the model can independently perform closed-loop driving simulation.

## 4. Experiments

In this section, we evaluate the performance of our model on a variety of driving-related tasks. Despite only using a single-view camera as input and high-level QAs as supervision, the model still exhibits promising performance.

### 4.1. Dataset

We conducted experiments on the widely used nuScenes dataset [10] for autonomous driving. It includes 1,000 driving sequences captured in diverse scenarios, including daytime/nighttime and sunny/cloudy/rainy weather. Each video clip consists of 400 frames with a frequency of 20Hz, spanning across 20 seconds as a scene. They are downsampled to 2Hz to obtain keyframes and annotated with labels including scene descriptions, 3D object bounding boxes, and semantic maps. Though it provides point clouds scanned by a 32-beam LiDAR, we mainly focus on using the RGB images captured by six surrounding cameras as inputs. We follow existing methods [100] to use the training 700 scenes for training and the validation 150 scenes for testing.

### 4.2. Task Descriptions

We evaluate our **Doe-1** on various tasks including visual question-answering, motion planning, and action-conditioned video generation.

**Visual Question-Answering.** We evaluate the perception ability of our method on the OmniDrive-nuScenes [79] benchmark, which supplements the original nuScenes data with high-quality visual question-answering (QA) text pairs generated by GPT4. The QA pairs include perception, reasoning, and planning in the 3D space. The goal of visual

Table 1. Visual question-answering results on OmniDrive-nuScenes [79]. \* denotes only using the front camera as input.

Method	Counterfactual		METEOR↑	Caption CIDEr ↑	ROUGE↑
	AP (%) ↑	AR (%) ↑			
OmniDrive-3D [79]	52.3	<b>59.6</b>	38.0	68.6	32.6
OmniDrive-2D [79]	51.5	55.0	<b>38.3</b>	67.1	32.5
OmniDrive-BEV [79]	45.6	49.5	35.6	59.5	27.8
<b>Doe-1*</b>	<b>54.5</b>	54.0	37.6	<b>72.8</b>	<b>35.6</b>



**Scene Descriptions:** The scene depicts a quiet residential street lined with tall trees and apartment buildings. Ahead, the road stretches straight, with a few parked vehicles visible in the distance. To the left, there is a grassy area with more trees. To the right, a green fence separates the road from a sidewalk. The overall atmosphere appears peaceful and suburban.

Figure 7. Visualizations of the visual-question answering results of our Doe-1. Doe-1 produces language descriptions and answers questions about the scene.

QA is to generate correct answers to questions about the visual observations. We use the widely used language-based metrics (i.e., METEOR [4], ROUGE [44], and CIDEr [76]) to compute the sentence similarities between the generated and ground-truth answers. For counterfactual reasoning, we ask the model to predict what will happen if executing a given trajectory. Following OmniDrive [79], we extract keywords from the predictions, including “safety”, “collision”, “running a red light”, and “out of the drivable area”. We compare them with the ground truth and compute the average precision (AP) and recall (AR) for all categories.

**Action-Conditioned Video Generation.** The objective of action-conditioned video generation is to generate high-quality images consistent with the controlling conditions. World models in autonomous driving are usually formulated as generating future observations given the actions (e.g., high-level command, trajectory). Following existing methods [20, 82], we report the Fréchet Inception Distance (FID) [24] score to measure the image generation quality.

**End-to-End Motion Planning.** End-to-end motion planning aims to produce safe and consistent future ego trajectories from observations. To assess performance, we use L2 error and collision rate as the key evaluation metrics. We evaluate the performance of our GaussianAD with the L2 error and collision rate following existing end-to-

<b>Q:</b> What kind of vehicles can you see around you?
<b>A:</b> There is a car parked on the right side of the street, a white van ahead in the distance, and another car parked to the right-front.
<b>Q:</b> Are there any pedestrians or cyclists visible in the vicinity?
<b>A:</b> No, there are no pedestrians or cyclists visible in the immediate vicinity.
<b>Q:</b> What is the appropriate driving behavior in this scenario?
<b>A:</b> The appropriate driving behavior would be to maintain a moderate speed, continue straight, and stay in the current lane while being mindful of the parked vehicles.
<b>Q:</b> If you follow the trajectory [PT, [3.32, -0.04], [6.83, -0.12], [10.28, -0.18], [13.64, -0.19], [17.09, -0.14], [20.18, -0.04]], what would happen?
<b>A:</b> Following this trajectory would result in running the red light, which is a traffic violation. This action is unsafe and illegal. Additionally, it could potentially lead to a collision with the moving car at (22.2, -0.9).

end methods [27, 28, 99, 100]. The L2 error computes the L2 distance between the planned and ground-truth paths. The collision rate reflects how often the autonomous vehicle collides with other objects while executing the planned trajectory. For fair comparisons, we use a 5-frame (i.e., 2 seconds) history and compute the metrics for the 1s, 2s, and 3s future.

### 4.3. Implementation Details

We use the pre-trained Lumina-mGPT 7B [45] to initialize our model and fine-tune the model for 5 epochs on the BDD100k [94] dataset to improve the conditioned image generation capability for driving scene images. We take as input images with a resolution of  $672 \times 384$  and only adopt the images from the camera. The resolution of the action tokenizer is set to 0.02m, where the scaling factor is set to 10000 for displacements and 1000 for yaw angles.

For training, we use the AdamW [50] optimizer with a cosine learning rate scheduler. The initial learning rate is set to  $1 \times 10^{-5}$  with a weight decay of 0.1. To emphasize the accuracy of action prediction, we increase the loss weight for action tokens in the input sequence by a factor of 5. Z-loss is applied with a weight of  $1 \times 10^{-5}$  to stabilize the training. We train our model for 16 epochs on 8 A100 GPUs with a total batch size of 24.

Table 2. Action-conditioned video generation results on the nuScenes [10] dataset.

Method	DriveGAN [38]	DriveDreamer [82]	WoVoGen [51]	Drive-WM [83]	GenAD [90]	Vista [20]	<b>Doe-1</b>
Type	GAN	Diffusion	Diffusion	Diffusion	Diffusion	Diffusion	Auto-Regressive
Resolution	$256 \times 256$	$128 \times 192$	$256 \times 448$	$192 \times 384$	$256 \times 448$	$576 \times 1024$	$384 \times 672$
FID ↓	73.4	52.6	27.6	15.8	15.4	<b>6.9</b>	15.9

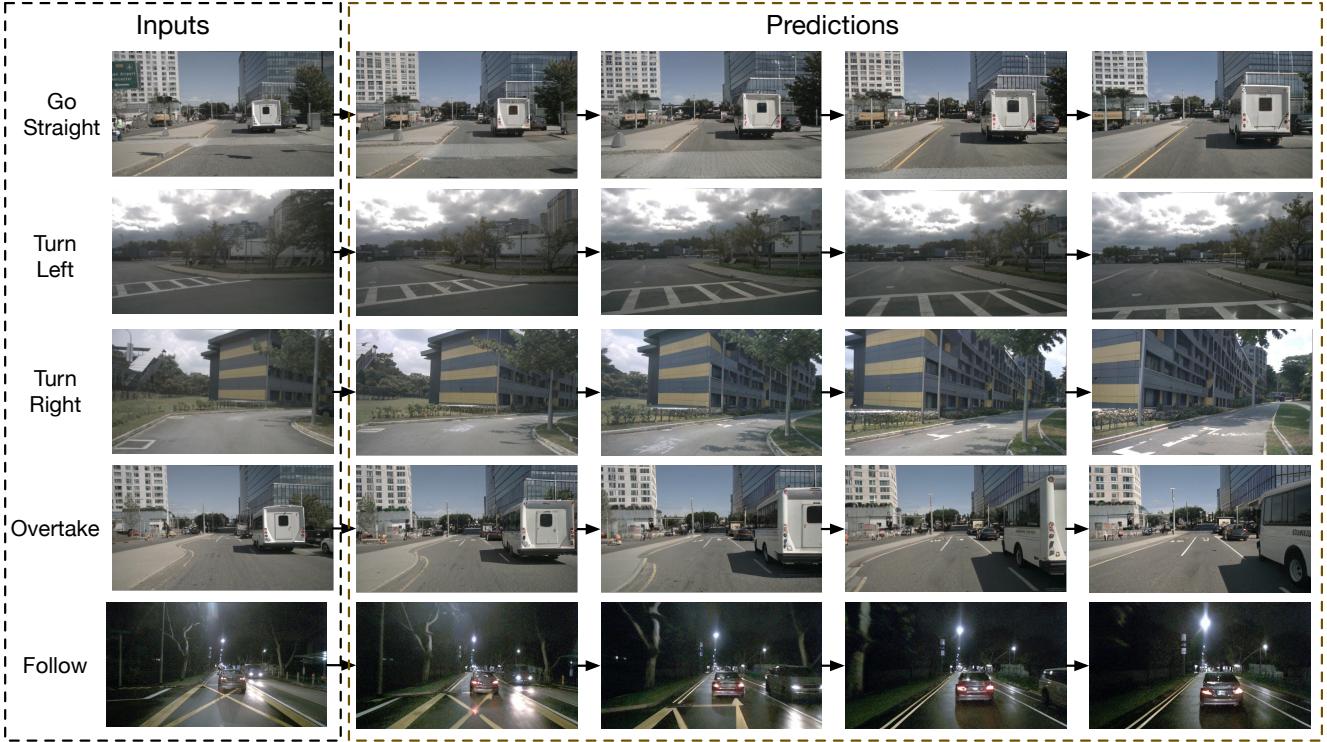


Figure 8. Visualizations of action-conditioned video generation results of our **Doe-1**. **Doe-1** generates high-quality videos in consistent with the 3D structures and action conditions.

#### 4.4. Close-Loop Autonomous Driving

Figure 1 shows the visualizations of **Doe-1** for closed-loop autonomous driving. We model perception, planning, and prediction as the transitions of observation→description, description→action, and action→observation, respectively. We observe that the proposed method can correctly generate scene descriptions, answer questions about the scene, plan ego trajectory, and predict future observations correctly with one trained model without finetuning.

#### 4.5. Visual Question-Answering

We use visual question-answering to evaluate the perception ability of our **Doe-1**. We compare our method with OmniDrive [79] with 3D Q-Former (OmniDrive-3D), 2D Q-Former (OmniDrive-2D), and dense BEV (OmniDrive-BEV) on the OmniDrive-nuScenes [79] benchmark, as shown in Table 1. We use bold numbers to denote the best results. Note that OmniDrive uses surrounding cameras as inputs, while our **Doe-1** only uses the front camera. Still, we see that our model achieves competitive results on both visual caption and counterfactual reasoning tasks.

**Visualizations.** We provide a qualitative analysis of the visual question-answering results in Figure 7. We see that our **Doe-1** correctly describes the scene and answers the questions about the input image.

#### 4.6. Action-Conditioned Video Generation

We evaluate the prediction ability of **Doe-1** on the action-conditioned video generation, where we adopt accurate actions (displacements in the BEV space) as the condition. We compare our model with existing real-world world models in Table 2. We see that **Doe-1** achieves comparable performance with Drive-WM [83] and GenAD [90], yet underperforms Vista [20]. Still, our model is the first to use the autoregressive architecture instead of diffusion, facilitating the joint perception and planning of **Doe-1**.

**Visualizations.** Figure 8 shows the generated sequences of images given an image and trajectory as conditions. We see that **Doe-1** generates high-quality images following the prompt actions. They show consistency in the 3D structure and demonstrate the ability of **Doe-1** to understand the evolutions of the 3D world.

Table 3. **End-to-end motion planning results on the nuScenes [10] dataset.** \* represents only using the front camera as input, and † denotes using the metric adopted in VAD [33].

Method	Input	Auxiliary Supervision	L2 (m) ↓				Collision Rate (%) ↓			
			1s	2s	3s	Avg.	1s	2s	3s	Avg.
IL [61]	LiDAR	None	0.44	1.15	2.47	1.35	0.08	0.27	1.95	0.77
NMP [95]	LiDAR	Box & Motion	0.53	1.25	2.67	1.48	0.04	0.12	0.87	0.34
FF [26]	LiDAR	Freespace	0.55	1.20	2.54	1.43	0.06	0.17	1.07	0.43
EO [35]	LiDAR	Freespace	0.67	1.36	2.78	1.60	0.04	0.09	0.88	0.33
ST-P3 [27]	Camera	Map & Box & Depth	1.33	2.11	2.90	2.11	0.23	0.62	1.27	0.71
UniAD [28]	Camera	Map & Box & Motion & Tracklets & Occ	0.48	0.96	1.65	1.03	0.05	<b>0.17</b>	<b>0.71</b>	<b>0.31</b>
OccNet [72]	Camera		1.29	2.13	2.99	2.14	0.21	0.59	1.37	0.72
OccWorld [99]	Camera	3D-Occ & Map & Box	0.52	1.27	2.41	1.40	0.12	0.40	2.08	0.87
VAD-Tiny [33]	Camera	Map & Box & Motion	0.60	1.23	2.06	1.30	0.31	0.53	1.33	0.72
VAD-Base [33]	Camera	Map & Box & Motion	0.54	1.15	1.98	1.22	<b>0.04</b>	0.39	1.17	0.53
GenAD [100]	Camera	Map & Box & Motion	<b>0.36</b>	<b>0.83</b>	<b>1.55</b>	<b>0.91</b>	0.06	0.23	1.00	0.43
<b>Doe-1</b>	Camera*	QA	0.50	1.18	2.11	1.26	<b>0.04</b>	0.37	1.19	0.53
VAD-Tiny† [33]	Camera	Map & Box & Motion	0.46	0.76	1.12	0.78	0.21	0.35	0.58	0.38
VAD-Base† [33]	Camera	Map & Box & Motion	0.41	0.70	1.05	0.72	0.07	0.17	0.41	0.22
OccWorld-D† [99]	Camera	3D-Occ	0.39	0.73	1.18	0.77	0.11	0.19	0.67	0.32
GenAD† [100]	Camera	Map & Box & Motion	0.28	0.49	0.78	0.52	0.08	0.14	<b>0.34</b>	<b>0.19</b>
OmniDrive† [79]	Camera	Map & Box & Motion & QA	<b>0.14</b>	<b>0.29</b>	<b>0.55</b>	<b>0.33</b>	<b>0.00</b>	<b>0.13</b>	0.78	0.30
<b>Doe-1†</b>	Camera*	QA	0.37	0.67	1.07	0.70	0.02	0.14	0.47	0.21

Table 4. **Ablation study on planning strategies.** “Description” means no other prompt between observations and actions. “Mask” means the plan of preceding frames will not be masked during training and evaluation.

Ablation	L2 (m) ↓				Collision Rate (%) ↓			
	1s	2s	3s	Avg.	1s	2s	3s	Avg.
Description	0.70	1.68	2.99	1.79	0.07	0.44	2.69	1.07
Mask	0.95	2.94	4.59	2.83	1.11	2.82	4.23	2.72
Full Model	<b>0.50</b>	<b>1.18</b>	<b>2.11</b>	<b>1.26</b>	<b>0.04</b>	<b>0.37</b>	<b>1.19</b>	<b>0.53</b>

#### 4.7. End-to-End Motion Planning

We evaluate the action planning performance of our **Doe-1** following existing end-to-end autonomous driving methods [28, 33, 100], as shown in Table 3. Additionally, we also compute the average performance across all previous frames for each time step at the bottom of the table following VAD [33]. Though our **Doe-1** does not achieve the best results, it demonstrates competitive performance with existing methods using only question-answering pairs as the auxiliary supervision. Note that using more supervision signals generally results in better performance, with the cost of expensive annotations. Also, our model only takes the front camera as input, while the other vision-based methods use surrounding cameras. Still, our model plans the future trajectory with a satisfactory collision rate. In particular, **Doe-1** delivers a small collision rate within 1 second, which is the most important factor in the practical close-loop scenario.

#### 4.8. Analysis

**Effect of Different Planning Strategies.** **Doe-1** leverages the perceived descriptions before generating the current action and masks the previous frames of the generated actions to avoid accumulation of error. Table 4 demonstrates the effectiveness of our design, which shows that the planning performance is influenced by the constraints of the textual modality. The mask mechanism also effectively prevents significant error accumulation.

#### 5. Conclusion

In this paper, we have presented a large driving world model (**Doe-1**) for closed-loop autonomous driving. While existing end-to-end autonomous driving methods demonstrate strong planning performance, they are still open-loop and suffer from information loss with hand-crafted scene representations. We address this with a next-token prediction formulation and model perception, prediction, and planning with the transitions between multi-modal tokens. We have conducted extensive experiments on the widely used nuScenes dataset and demonstrated the effectiveness of **Doe-1** on visual question-answering, action-conditioned video generation, and end-to-end motion planning.

**Limitations.** **Doe-1** only takes the front-view images as inputs due to the inefficiency of using multi-view inputs. However, surround-view information is critical for safe autonomous driving and is an interesting future direction.

## References

- [1] Ben Agro, Quinlan Sykora, Sergio Casas, Thomas Gilles, and Raquel Urtasun. Uno: Unsupervised occupancy fields for perception and forecasting. In *CVPR*, pages 14487–14496, 2024. 3
- [2] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 2
- [3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 2
- [4] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for mt evaluation with improved correlation with human judgments. In *ACL Workshop*, 2005. 7
- [5] Daniel Bogdolla, Yitian Yang, and J Marius Zöllner. Muvo: A multimodal generative world model for autonomous driving with geometric representations. *arXiv preprint arXiv:2311.11762*, 2023. 3
- [6] Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, et al. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*, 2016. 3
- [7] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023. 2
- [8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *NeurIPS*, 33:1877–1901, 2020. 2
- [9] Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. 2
- [10] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liang, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, pages 11621–11631, 2020. 6, 8, 9
- [11] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liang, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020. 1, 2
- [12] Dian Chen, Brady Zhou, Vladlen Koltun, and Philipp Krähenbühl. Learning by cheating. In *CoRL*, pages 66–75, 2020. 3
- [13] Long Chen, Oleg Sinavski, Jan Hünermann, Alice Karnsund, Andrew James Willmott, Danny Birch, Daniel Maund, and Jamie Shotton. Driving with llms: Fusing object-level vector modality for explainable autonomous driving. *arXiv preprint arXiv:2310.01957*, 2023. 2, 3
- [14] Jie Cheng, Ren Xin, Sheng Wang, and Ming Liu. Mpnp: Multi-policy neural planner for urban driving. In *IROS*, pages 10549–10554, 2022. 3
- [15] Kashyap Chitta, Aditya Prakash, Bernhard Jaeger, Zehao Yu, Katrin Renz, and Andreas Geiger. Transfuser: Imitation with transformer-based sensor fusion for autonomous driving. *TPAMI*, 2022. 3
- [16] Felipe Codevilla, Matthias Müller, Antonio López, Vladlen Koltun, and Alexey Dosovitskiy. End-to-end driving via conditional imitation learning. In *ICRA*, pages 4693–4700, 2018. 3
- [17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2020. 2
- [18] Xin Fei, Wenzhao Zheng, Yueqi Duan, Wei Zhan, Masayoshi Tomizuka, Kurt Keutzer, and Jiwen Lu. Driv3r: Learning dense 4d reconstruction for autonomous driving. *arXiv preprint arXiv:2412.06777*, 2024. 4
- [19] Ruiyuan Gao, Kai Chen, Enze Xie, Lanqing Hong, Zhen-guo Li, Dit-Yan Yeung, and Qiang Xu. Magicdrive: Street view generation with diverse 3d geometry control. *arXiv preprint arXiv:2310.02601*, 2023. 3
- [20] Shenyuan Gao, Jiazhai 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. 3, 7, 8
- [21] Anisha Gunjal, Jihan Yin, and Erhan Bas. Detecting and preventing hallucinations in large vision language models. In *AAAI*, pages 18135–18143, 2024. 2
- [22] David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018. 3
- [23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2
- [24] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 33:6840–6851, 2020. 7
- [25] 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. 3
- [26] Peiyun Hu, Aaron Huang, John Dolan, David Held, and Deva Ramanan. Safe local motion planning with self-supervised freespace forecasting. In *CVPR*, 2021. 9
- [27] 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. 2, 3, 4, 7, 9
- [28] Yihan Hu, Jiazhai Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhui Wang, et al. Planning-oriented autonomous driving. In *CVPR*, pages 17853–17862, 2023. 2, 3, 4, 7, 9

- [29] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *CVPR*, 2017. 2
- [30] Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Tri-perspective view for vision-based 3d semantic occupancy prediction. In *CVPR*, pages 9223–9232, 2023. 3, 4
- [31] Yuanhui Huang, Amonnut Thammatadatrakoon, Wenzhao Zheng, Yunpeng Zhang, Dalong Du, and Jiwen Lu. Gaussianformer-2: Probabilistic gaussian superposition for efficient 3d occupancy prediction. *arXiv preprint arXiv:2412.04384*, 2024. 4
- [32] Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Gaussianformer: Scene as gaussians for vision-based 3d semantic occupancy prediction. *arXiv preprint arXiv:2405.17429*, 2024. 4
- [33] Bo Jiang, Shaoyu Chen, Qing Xu, Bencheng Liao, Jiajie Chen, Helong Zhou, Qian Zhang, Wenyu Liu, Chang Huang, and Xinggang Wang. Vad: Vectorized scene representation for efficient autonomous driving. *arXiv preprint arXiv:2303.12077*, 2023. 2, 3, 4, 9
- [34] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020. 2
- [35] Tarasha Khurana, Peiyun Hu, Achal Dave, Jason Ziglar, David Held, and Deva Ramanan. Differentiable raycasting for self-supervised occupancy forecasting. In *ECCV*, 2022. 3, 9
- [36] Tarasha Khurana, Peiyun Hu, David Held, and Deva Ramanan. Point cloud forecasting as a proxy for 4d occupancy forecasting. In *CVPR*, pages 1116–1124, 2023. 3
- [37] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024. 2
- [38] Seung Wook Kim, Jonah Philion, Antonio Torralba, and Sanja Fidler. DriveGAN: Towards a Controllable High-Quality Neural Simulation. In *CVPR*, 2021. 8
- [39] 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*, pages 19730–19742, 2023. 2
- [40] 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. 3
- [41] Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. *arXiv preprint arXiv:2206.10092*, 2022. 4
- [42] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023. 2
- [43] Zhiqi Li, Wenhui Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Qiao Yu, and Jifeng Dai. Bevformer: Learning bird’s-eye-view representation from multi-camera images via spatiotemporal transformers. In *ECCV*, 2022. 4
- [44] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, 2004. 7
- [45] Dongyang Liu, Shitian Zhao, Le Zhuo, Weifeng Lin, Yu Qiao, Hongsheng Li, and Peng Gao. Lumina-mgpt: Illuminate flexible photorealistic text-to-image generation with multimodal generative pretraining. *arXiv preprint arXiv:2408.02657*, 2024. 5, 7
- [46] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, pages 26296–26306, 2024. 2
- [47] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 36, 2024. 2
- [48] Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*, 2024. 2, 3
- [49] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 10012–10022, 2021. 2
- [50] Ilya Loshchilov, Frank Hutter, et al. Fixing weight decay regularization in adam. *arXiv preprint arXiv:1711.05101*, 5, 2017. 7
- [51] 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. 8
- [52] Enhui Ma, Lijun Zhou, Tao Tang, Zhan Zhang, Dong Han, Junpeng Jiang, Kun Zhan, Peng Jia, Xianpeng Lang, Haiyang Sun, et al. Unleashing generalization of end-to-end autonomous driving with controllable long video generation. *arXiv preprint arXiv:2406.01349*, 2024. 3
- [53] Jiageng Mao, Yuxi Qian, Hang Zhao, and Yue Wang. Gpt-driver: Learning to drive with gpt. *arXiv preprint arXiv:2310.01415*, 2023. 2
- [54] Benedikt Mersch, Xieyanli Chen, Jens Behley, and Cyrill Stachniss. Self-supervised point cloud prediction using 3d spatio-temporal convolutional networks. In *CoRL*, pages 1444–1454. PMLR, 2022. 3
- [55] Ming Nie, Renyuan Peng, Chunwei Wang, Xinyue Cai, Jianhua Han, Hang Xu, and Li Zhang. Reason2drive: Towards interpretable and chain-based reasoning for autonomous driving. In *ECCV*, pages 292–308, 2025. 2, 3
- [56] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. *arXiv preprint arXiv:1711.00937*, 2017. 2
- [57] OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023. 2
- [58] Tung Phan-Minh, Elena Corina Grigore, Freddy A Boulton, Oscar Beijbom, and Eric M Wolff. Covernet: Multimodal behavior prediction using trajectory sets. In *CVPR*, 2020. 3

- [59] Aditya Prakash, Aseem Behl, Eshed Ohn-Bar, Kashyap Chitta, and Andreas Geiger. Exploring data aggregation in policy learning for vision-based urban autonomous driving. In *CVPR*, pages 11763–11773, 2020. 3
- [60] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 2
- [61] Nathan D Ratliff, J Andrew Bagnell, and Martin A Zinkevich. Maximum margin planning. In *ICML*, pages 729–736, 2006. 9
- [62] Katrin Renz, Long Chen, Ana-Maria Marcu, Jan Hünermann, Benoit Hanotte, Alice Karnsund, Jamie Shotton, Elahe Arani, and Oleg Sinavski. Carllava: Vision language models for camera-only closed-loop driving. *arXiv preprint arXiv:2406.10165*, 2024. 2
- [63] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 3
- [64] Oliver Scheel, Luca Bergamini, Maciej Wolczyk, Błażej Osiński, and Peter Ondruska. Urban driver: Learning to drive from real-world demonstrations using policy gradients. In *CoRL*, pages 718–728, 2022. 3
- [65] Hao Sha, Yao Mu, Yuxuan Jiang, Li Chen, Chenfeng Xu, Ping Luo, Shengbo Eben Li, Masayoshi Tomizuka, Wei Zhan, and Mingyu Ding. Languagempc: Large language models as decision makers for autonomous driving. *arXiv preprint arXiv:2310.03026*, 2023. 2, 3
- [66] Hao Shao, Yuxuan Hu, Letian Wang, Steven L Waslander, Yu Liu, and Hongsheng Li. Lmdrive: Closed-loop end-to-end driving with large language models. *arXiv preprint arXiv:2312.07488*, 2023. 2, 3
- [67] SP Sharan, Francesco Pittaluga, Manmohan Chandraker, et al. Llm-assist: Enhancing closed-loop planning with language-based reasoning. *arXiv preprint arXiv:2401.00125*, 2023. 2, 3
- [68] Wenchao Sun, Xuewu Lin, Yining Shi, Chuang Zhang, Haoran Wu, and Sifa Zheng. Sparsedrive: End-to-end autonomous driving via sparse scene representation. *arXiv preprint arXiv:2405.19620*, 2024. 3
- [69] Richard S Sutton. Dyna, an integrated architecture for learning, planning, and reacting. *ACM Sigart Bulletin*, 2(4):160–163, 1991. 3
- [70] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024. 4, 5
- [71] Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Yang Wang, Zhiyong Zhao, Kun Zhan, Peng Jia, Xianpeng Lang, and Hang Zhao. Drivevml: The convergence of autonomous driving and large vision-language models. *arXiv preprint arXiv:2402.12289*, 2024. 2
- [72] Wenwen Tong, Chonghao Sima, Tai Wang, Li Chen, Silei Wu, Hanming Deng, Yi Gu, Lewei Lu, Ping Luo, Dahua Lin, et al. Scene as occupancy. In *ICCV*, pages 8406–8415, 2023. 2, 3, 4, 9
- [73] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *corr*, abs/2302.13971, 2023. doi: 10.48550. *arXiv preprint arXiv:2302.13971*. 2
- [74] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 2
- [75] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017. 2
- [76] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based image description evaluation. In *CVPR*, 2015. 7
- [77] Lening Wang, Wenzhao Zheng, Dalong Du, Yunpeng Zhang, Yilong Ren, Han Jiang, Zhiyong Cui, Haiyang Yu, Jie Zhou, Jiwen Lu, and Shanghang Zhang. Stag-1: Towards realistic 4d driving simulation with video generation model. *arXiv preprint arXiv:2405.00001*, 2024. 3
- [78] Lening Wang, Wenzhao Zheng, Yilong Ren, Han Jiang, Zhiyong Cui, Haiyang Yu, and Jiwen Lu. Occsora: 4d occupancy generation models as world simulators for autonomous driving. *arXiv preprint arXiv:2405.20337*, 2024. 3
- [79] Shihao Wang, Zhiding Yu, Xiaohui Jiang, Shiyi Lan, Min Shi, Nadine Chang, Jan Kautz, Ying Li, and Jose M Alvarez. Omnidrive: A holistic llm-agent framework for autonomous driving with 3d perception, reasoning and planning. *arXiv preprint arXiv:2405.01533*, 2024. 2, 3, 6, 7, 8, 9
- [80] Tianqi Wang, Enze Xie, Ruihang Chu, Zhenguo Li, and Ping Luo. Drivecot: Integrating chain-of-thought reasoning with end-to-end driving. *arXiv preprint arXiv:2403.16996*, 2024. 2
- [81] Wenhui Wang, Jiangwei Xie, ChuanYang Hu, Haoming Zou, Jianan Fan, Wenwen Tong, Yang Wen, Silei Wu, Hanming Deng, Zhiqi Li, et al. Drivemlm: Aligning multi-modal large language models with behavioral planning states for autonomous driving. *arXiv preprint arXiv:2312.09245*, 2023. 2
- [82] 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. 3, 4, 7, 8
- [83] 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. 8
- [84] Yi Wei, Linqing Zhao, Wenzhao Zheng, Zheng Zhu, Jie Zhou, and Jiwen Lu. Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving. In *ICCV*, pages 21729–21740, 2023. 4
- [85] Yuqing Wen, Yucheng Zhao, Yingfei Liu, Fan Jia, Yanhui Wang, Chong Luo, Chi Zhang, Tiancai Wang, Xiaoyan Sun,

- and Xiangyu Zhang. Panacea: Panoramic and controllable video generation for autonomous driving. In *CVPR*, pages 6902–6912, 2024. 3
- [86] Xinshuo Weng, Jianren Wang, Sergey Levine, Kris Kitani, and Nicholas Rhinehart. Inverting the pose forecasting pipeline with spf2: Sequential pointcloud forecasting for sequential pose forecasting. In *CoRL*, pages 11–20, 2021. 3
- [87] Yuqi Wu, Wenzhao Zheng, Sicheng Zuo, Yuanhui Huang, Jie Zhou, and Jiwen Lu. Embodiedoccc: Embodied 3d occupancy prediction for vision-based online scene understanding. *arXiv preprint arXiv:2412.04380*, 2024. 4
- [88] Zixun Xie, Sicheng Zuo, Wenzhao Zheng, Yunpeng Zhang, Dalong Du, Jie Zhou, Jiwen Lu, and Shanghang Zhang. Gpd-1: Generative pre-training for driving. *arXiv preprint arXiv:2412.08643*, 2024. 3
- [89] Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kwan-Yee K Wong, Zhenguo Li, and Hengshuang Zhao. Drivegpt4: Interpretable end-to-end autonomous driving via large language model. *RAL*, 2024. 2
- [90] Jiazhai 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*, pages 14662–14672, 2024. 3, 8
- [91] Kairui Yang, Enhui Ma, Jibin Peng, Qing Guo, Di Lin, and Kaicheng Yu. Bevcontrol: Accurately controlling street-view elements with multi-perspective consistency via bev sketch layout. *arXiv preprint arXiv:2308.01661*, 2023. 3
- [92] Yu Yang, Jianbiao Mei, Yukai Ma, Siliang Du, Wenqing Chen, Yijie Qian, Yuxiang Feng, and Yong Liu. Driving in the occupancy world: Vision-centric 4d occupancy forecasting and planning via world models for autonomous driving. *arXiv preprint arXiv:2408.14197*, 2024. 3
- [93] Tengju Ye, Wei Jing, Chunyong Hu, Shikun Huang, Lingsheng Gao, Fangzhen Li, Jingke Wang, Ke Guo, Wencong Xiao, Weibo Mao, et al. Fusionad: Multi-modality fusion for prediction and planning tasks of autonomous driving. *arXiv preprint arXiv:2308.01006*, 2023. 2, 3, 4
- [94] Fisher Yu, Wenqi Xian, Yingying Chen, Fangchen Liu, Mike Liao, Vashisht Madhavan, Trevor Darrell, et al. Bdd100k: A diverse driving video database with scalable annotation tooling. *arXiv preprint arXiv:1805.04687*, 2(5):6, 2018. 7
- [95] Wenyuan Zeng, Wenjie Luo, Simon Suo, Abbas Sadat, Bin Yang, Sergio Casas, and Raquel Urtasun. End-to-end interpretable neural motion planner. In *CVPR*, 2019. 9
- [96] Diankun Zhang, Guoan Wang, Runwen Zhu, Jianbo Zhao, Xiwu Chen, Siyu Zhang, Jiahao Gong, Qibin Zhou, Wenyuan Zhang, Ningzi Wang, et al. Sparsead: Sparse query-centric paradigm for efficient end-to-end autonomous driving. *arXiv preprint arXiv:2404.06892*, 2024. 3
- [97] Ruijun Zhang, Xianda Guo, Wenzhao Zheng, Chenming Zhang, Kurt Keutzer, and Long Chen. Instruct large language models to drive like humans. *arXiv preprint arXiv:2406.07296*, 2024. 2, 3
- [98] Zaibin Zhang, Shiyu Tang, Yuanhang Zhang, Talas Fu, Yifan Wang, Yang Liu, Dong Wang, Jing Shao, Lijun Wang, and Huchuan Lu. Ad-h: Autonomous driving with hierarchical agents. *arXiv preprint arXiv:2406.03474*, 2024. 2
- [99] 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. 3, 4, 7, 9
- [100] Wenzhao Zheng, Ruiqi Song, Xianda Guo, and Long Chen. Genad: Generative end-to-end autonomous driving. In *ECCV*, 2024. 2, 3, 4, 6, 7, 9
- [101] Sicheng Zuo, Wenzhao Zheng, Yuanhui Huang, Jie Zhou, and Jiwen Lu. Pointocc: Cylindrical tri-perspective view for point-based 3d semantic occupancy prediction. *arXiv preprint arXiv:2308.16896*, 2023. 4