Driving in the Occupancy World: Vision-Centric 4D Occupancy Forecasting and Planning via World Models for Autonomous Driving

Yu Yang [1*], Jianbiao Mei [1*], Yukai Ma [1], Siliang Du [2†], Wenqing Chen [2], Yijie Qian [1], Yuxiang Feng [1], Yong Liu [1†]

[1]Zhejiang University
[2]Huawei Technologies
{yu.yang, jianbiaomei, yukaima, yijieqian, yuxiangfeng}@zju.edu.cn
{dusiliang, chenwenqing7}@huawei.com,yongliu@iipc.zju.edu.cn

Abstract

World models envision potential future states based on various ego actions. They embed extensive knowledge about the driving environment, facilitating safe and scalable autonomous driving. Most existing methods primarily focus on either data generation or the pretraining paradigms of world models. Unlike the aforementioned prior works, we propose **Drive-OccWorld**, which adapts a vision-centric 4D forecasting world model to end-to-end planning for autonomous driving. Specifically, we first introduce a semantic and motion-conditional normalization in the memory module, which accumulates semantic and dynamic information from historical BEV embeddings. These BEV features are then conveyed to the world decoder for future occupancy and flow forecasting, considering both geometry and spatiotemporal modeling. Additionally, we propose injecting flexible action conditions, such as velocity, steering angle, trajectory, and commands, into the world model to enable controllable generation and facilitate a broader range of downstream applications. Furthermore, we explore integrating the generative capabilities of the 4D world model with end-to-end planning, enabling continuous forecasting of future states and the selection of optimal trajectories using an occupancy-based cost function. Extensive experiments on the nuScenes dataset demonstrate that our method can generate plausible and controllable 4D occupancy, opening new avenues for driving world generation and end-to-end planning.

Project Page — https://drive-occworld.github.io/

1 Introduction

Autonomous driving (AD) algorithms have advanced significantly in recent decades (Ayoub et al. 2019; Chen et al. 2023). These advancements have transitioned from modular pipelines (Guo et al. 2023; Li et al. 2023b) to end-to-end models (Hu et al. 2023b; Jiang et al. 2023), which plan trajectories directly from raw sensor data in a unified pipeline. However, due to insufficient world knowledge for forecasting dynamic environments, these methods exhibit deficiencies in generalization ability and safety robustness.

On the other hand, to embed world knowledge and simulate the real-world physics of the driving environment, recent works (Zhang et al. 2023; Min et al. 2024; Yang et al. 2024b) have introduced the world model (Ha and Schmidhuber 2018) to facilitate scalable autonomous driving. Nevertheless, most of them primarily focus on either data generation or the pretraining paradigms of world models, neglecting the enhancement of safety and robustness for end-to-end planning. For example, many studies (Ma et al. 2024a; Wang et al. 2023a; Hu et al. 2023a) aimed to generate high-fidelity driving videos through world models to provide additional data for downstream training. The very recent ViDAR (Yang et al. 2024b) pre-trained the visual encoder by forecasting point clouds from historical visual input, enhancing performance on downstream tasks such as vision-centric 3D detection and segmentation. Therefore, we believe that integrating the future forecasting capabilities of world models with end-to-end planning remains a worthwhile area for exploration.

In this work, we investigate 4D forecasting and planning using world models to implement future state prediction and end-to-end planning. With the capability to envision various futures based on different ego actions, a world model allows the agent to anticipate potential outcomes in advance. As illustrated in Figure 1, the world model predicts the future state of the environment under different action conditions, using historical observations and various ego actions. Subsequently, the planner employs a cost function that considers both safety and the 3D structure of the environment to select the most suitable trajectory, enabling the agent to navigate effectively in diverse situations. Finally, the predicted future state and selected optimal trajectory can be reintroduced into the world model for the next rollout, facilitating continuous future prediction and trajectory planning. We experimentally demonstrate that leveraging the future forecasting capability of world models enhances the planner's generalization and safety robustness while providing more explainable decision-making, as detailed in Section 4.

Specifically, we propose **Drive-OccWorld**, a vision-centric 4D forecasting and planning world model for autonomous driving. Our Drive-OccWorld exhibits three key features: **(1)** *Understanding how the world evolves through 4D occupancy forecasting.* Drive-OccWorld predicts plausible future states based on accumulated historical expe-

[*]These authors contributed equally.
[†]Corresponding authors.

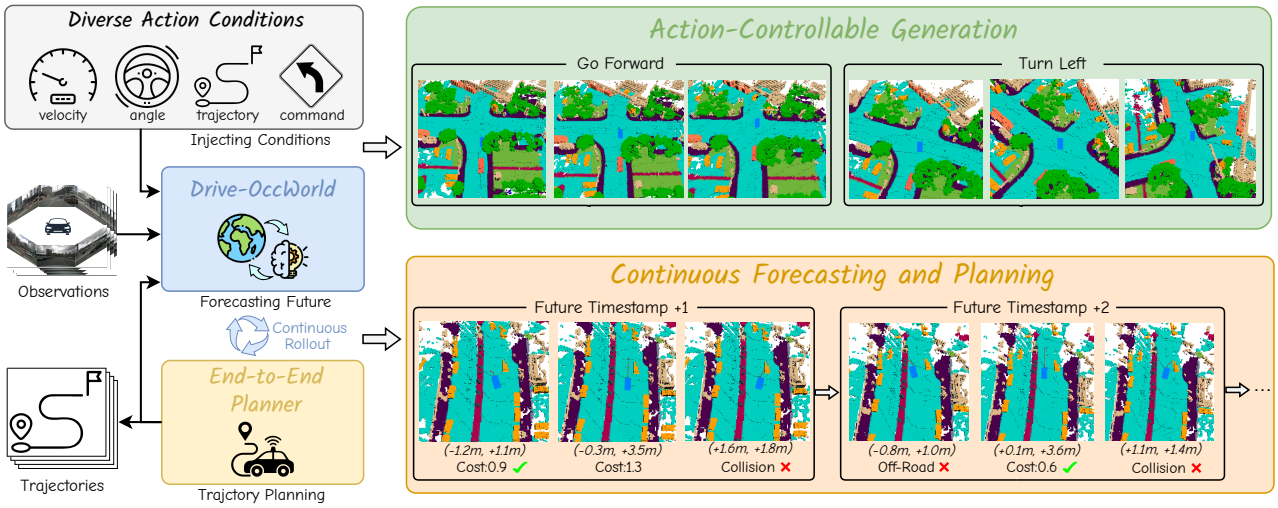

Figure 1: 4D Occupancy Forecasting and Planning via World Model. Drive-OccWorld takes observations and trajectories as input, incorporating flexible action conditions for *action-controllable generation*. By leveraging world knowledge and the generative capacity of the world model, we further integrate it with a planner for *continuous forecasting and planning*.

riences. It comprises three key components: a history encoder that encodes multi-view geometry BEV embeddings, a memory queue that accumulates historical information, and a future decoder that forecasts occupancy and flows through spatiotemporal modeling. Additionally, we introduce a semantic- and motion-conditional normalization to aggregate significant features. **(2)** *Generating various future states based on action conditions.* We incorporate a flexible set of action conditions (*e.g.*, velocity, steering angle, trajectory, and commands), which are encoded and injected into the world decoder through a unified interface, empowering the world model's capability for action-controllable generation. **(3)** *Planning trajectories with the world model.* Since the world model can forecast future occupancy and flow, providing perception and prediction results that include the fine-grained states of both agents and background elements, we further design a planner to select the optimal trajectory based on a comprehensive occupancy-based cost function.

We evaluate Drive-OccWorld on the nuScenes (Caesar et al. 2020) dataset in terms of vision-centric occupancy and flow forecasting, as well as trajectory planning. In forecasting the sequential occupancy of movable objects and their 3D backward centripetal flow, Drive-OccWorld outperforms previous methods by 2.0% in mIoU_f and 1.9% in VPQ_f. For forecasting the occupancy of both movable and static objects based on the OpenOccupancy benchmark (Wang et al. 2023b), it achieves 1.1% gains in $\widetilde{\text{mIoU}}_f$. Experiments on trajectory planning also demonstrate that Drive-OccWorld can be utilized for safe motion planning.

Our main contributions can be summarized as follows:

- We propose Drive-OccWorld, a vision-centric world model designed for forecasting 4D occupancy and dynamic flow, achieving new state-of-the-art performance.
- We develop a simple yet efficient semantic- and motion-conditional normalization module for semantic enhancement and motion compensation, which improves forecasting and planning performance.
- We incorporate flexible action conditions into Drive-OccWorld to enable action-controllable generation and explore integrating the world model with an occupancy-based planner for continuous forecasting and planning.

2 Related Works

2D Image-based World Models aim to predict future driving videos using reference images and various conditions (*e.g.*, actions, HDMaps, 3D boxes, and text prompts). GAIA-1 (Hu et al. 2023a) employs an autoregressive transformer as a world model to predict future driving videos. Other methods, such as DriveDreamer (Wang et al. 2023a), ADriver-I (Jia et al. 2023), DrivingDiffusion (Li, Zhang, and Ye 2023), GenAD (Yang et al. 2024a), Vista (Gao et al. 2024), Delphi (Ma et al. 2024a), and Drive-WM (Wang et al. 2024b), utilize *latent diffusion models* (Rombach et al. 2022; Blattmann et al. 2023) for driving video generation. These methods focus on designing modules to incorporate actions, BEV layouts, and other priors into the denoising process, resulting in more coherent and plausible video generations.

3D Volume-based World Models forecast future states in the form of point clouds or occupancy. Copilot4D (Zhang et al. 2023) tokenizes LiDAR observations with VQVAE (Van Den Oord, Vinyals et al. 2017) and predicts future point clouds via discrete diffusion. ViDAR (Yang et al. 2024b) implements a visual point cloud forecasting task to pre-train visual encoders. UnO (Agro et al. 2024) forecasts a continuous occupancy field with self-supervision from LiDAR data. OccWorld (Zheng et al. 2023) and OccSora (Wang et al. 2024a) compact the occupancy input with a scene tokenizer and use a generative transformer to predict future occupancy. UniWorld (Min et al. 2023) and DriveWorld (Min et al. 2024)

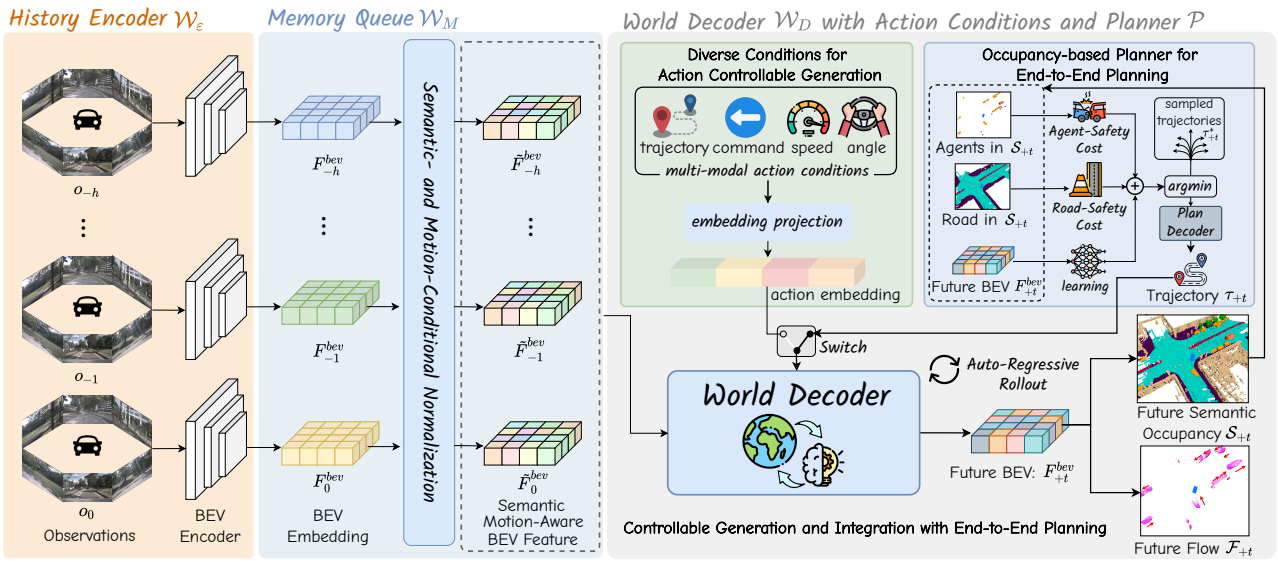

Figure 2: Overview of Drive-OccWorld. (a) The *history encoder* extracts multi-view image features and transforms them into BEV embeddings. (b) The *memory queue* employs semantic- and motion-conditional normalization to aggregate historical information. (c) The *world decoder* incorporates action conditions to generate various future occupancies and flows. Integrating the world decoder with an occupancy-based planner enables continuous forecasting and planning.

propose 4D pre-training via 4D occupancy reconstruction.

In this work, we investigate potential applications of the world model by injecting action conditions to enable action-controllable generation and integrating this generative capability with end-to-end planners for safe driving.

3 Method

3.1 Preliminary

An end-to-end autonomous driving model aims to control a vehicle (*i.e.*, plan trajectories) directly based on sensor inputs and ego actions (Hu et al. 2023b). Formally, given historical sensor observations $\{o_{-h}, \ldots, o_{-1}, o_0\}$ and ego trajectories $\{\tau_{-h}, \ldots, \tau_{-1}, \tau_0\}$ over h timestamps, an end-to-end model $\mathcal{A}$ predicts desirable ego trajectories $\{\tau_1, \ldots, \tau_f\}$ for the future f timestamps:

$$\mathcal{A}(\{o_{-h}, \ldots, o_{-1}, o_0\}, \{\tau_{-h}, \ldots, \tau_{-1}, \tau_0\}) = \{\tau_1, \ldots, \tau_f\} \tag{1}$$

A driving world model $\mathcal{W}$ can be viewed as a generative model that takes prior observations and ego actions $\{a_{-h}, \ldots, a_{-1}, a_0\}$ as input, generating plausible future states $\{s_1, \ldots, s_f\}$ of the environment:

$$\mathcal{W}(\{o_{-h}, \ldots, o_{-1}, o_0\}, \{a_{-h}, \ldots, a_0\}) = \{s_1, \ldots, s_f\} \tag{2}$$

where ego actions a can be injected into the controllable generation process in various forms, *i.e.*, velocity, steering angle, ego trajectory, and high-level commands.

Given the world model's ability to foresee future states, we propose integrating it with a planner to fully exploit the capabilities of the world model in end-to-end planning. Specifically, we introduce an auto-regressive framework termed Drive-OccWorld, which consists of a generative world model $\mathcal{W}$ to forecast future occupancy and flow states, and an occupancy-based planner $\mathcal{P}$ that employs a cost function to select the optimal trajectory based on evaluating future predictions. Formally, we formulate Drive-OccWorld as follows, which auto-regressively predicts the future state and trajectory at the next timestamp:

$$\mathcal{W}(\{o_{-h}, \ldots, o_{-1}, o_0\}, \{s_1, \ldots, s_{t-1}, s_t\}, \{a_{-h}, \ldots, a_{-1}, a_0, \ldots, a_{t-1}, a_t\}) = s_{t+1} \tag{3}$$

$$\mathcal{P}(f_o(s_{t+1}, \tau^*_{t+1})) = \tau_{t+1} \tag{4}$$

where f_o is the occupancy-based cost function, and τ^*_{t+1} denotes sampled trajectory proposals at the $t+1$ timestamp.

Notably, for action-controllable generation, a can be injected into $\mathcal{W}$ as conditions in the form of velocity, *etc.*, and $\mathcal{P}$ is discarded to prevent potential ego-status leakage. In end-to-end planning, the predicted trajectory τ_{t+1} serves as the action condition a_{t+1} to forecast the next state s_{t+2}, leading to a continuous rollout of forecasting and planning.

In the following sections, we will detail the world model's structure, equipping $\mathcal{W}$ with action-controllable generation and integrating it with $\mathcal{P}$ for end-to-end planning.

3.2 4D Forecasting with World Model

As depicted in Figure 2, Drive-OccWorld comprises three components: **(1)** a *History Encoder* $\mathcal{W}_\mathcal{E}$, which takes historical camera images as input, extracts multi-view geometry features, and transforms them into BEV embeddings. Following previous works (Yang et al. 2024b; Min et al. 2024), we utilize the visual BEV encoder (Li et al. 2022) as our history encoder. **(2)** a *Memory Queue* $\mathcal{W}_\mathcal{M}$ *with Semantic- and Motion-Conditional Normalization*, which employs a simple yet efficient normalization operation in latent space

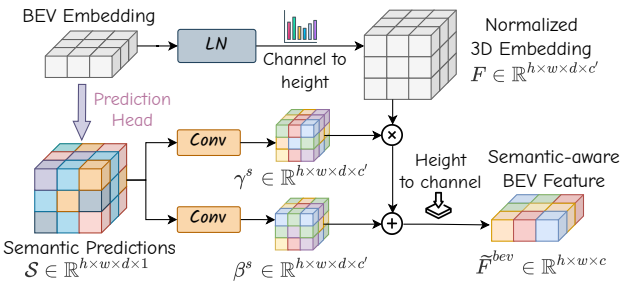

Figure 3: Overview of semantic-conditional normalization.

to aggregate semantic information and compensate for dynamic motions, thereby accumulating more representative BEV features. **(3)** a *World Decoder* $\mathcal{W}_\mathcal{D}$, which extracts world knowledge through temporal modeling with historical features to forecast future semantic occupancies and flows. Flexible action conditions can be injected into $\mathcal{W}_\mathcal{D}$ for controllable generation. An occupancy-based planner $\mathcal{P}$ is integrated for continuous forecasting and planning.

Semantic- and Motion-Conditional Normalization is designed to enhance historical BEV embeddings by incorporating semantic and dynamic information. For example, consider the BEV embedding $\boldsymbol{F}^{bev} \in \mathbb{R}^{h\times w\times c}$, where h and w are the spatial resolutions of the BEV, and c denotes the channel dimension. We first apply layer normalization without affine mapping, then modulate it into $\tilde{\boldsymbol{F}}^{bev}$ using an adaptive affine transformation, with the scale and shift parameters (γ^*, β^*) derived from semantic or motion labels:

$$\tilde{\boldsymbol{F}}^{bev} = \gamma^* \cdot \mathit{LayerNorm}(\boldsymbol{F}^{bev}) + \beta^* \tag{5}$$

Specifically, for semantic-conditional normalization, (γ^s, β^s) are inferred from voxel-wise semantic predictions. As illustrated in Figure 3, we implement a lightweight head along with the argmax function to predict voxel-wise semantic labels $\mathcal{S} \in \mathbb{R}^{h\times w\times d\times 1}$, where d denotes the height of the voxelized 3D space. The semantic labels are encoded as one-hot embeddings and convolved to produce modulation parameters for the affine transformation as Eq. 5. This method efficiently enhances the semantic discrimination of BEV embeddings, as demonstrated in the experiments.

In motion-conditional normalization, we account for the movements of both the ego vehicle and other agents across various timestamps. Specifically, the ego-pose transformation matrix $E_{-t}^{+t} = [R_{-t}^{+t}, T_{-t}^{+t}]$, which represents the rotation and translation of the ego vehicle from timestamp $-t$ to $+t$, is flattened and encoded into an embedding processed by MLPs to generate affine transformation parameters (γ^e, β^e). To address the movements of other agents, we predict voxel-wise 3D backward centripetal flow $\mathcal{F} \in \mathbb{R}^{h\times w\times d\times 3}$ that points from the voxel at time t to its corresponding 3D instance center at $t-1$, and encode it into (γ^f, β^f) for fine-grained motion-aware normalization using Eq. 5.

Future Forecasting with World Decoder. $\mathcal{W}_\mathcal{D}$ is an auto-regressive transformer that predicts the BEV embeddings $\boldsymbol{F}_{+t}^{bev}$ for the future frame $+t$ based on historical BEV features stored in $\mathcal{W}_\mathcal{M}$ and the expected action condition a_{+t}.

Specifically, $\mathcal{W}_\mathcal{D}$ takes learnable BEV queries as input and performs deformable self-attention, temporal cross-attention with historical embeddings, conditional cross-attention with action conditions, and a feedforward network to generate future BEV embeddings. The conditional layer performs cross-attention between BEV queries and action embeddings, which will be illustrated in the following section, injecting action-controllable information into the forecasting process. After obtaining the next BEV embeddings $\boldsymbol{F}_{+t}^{bev}$, prediction heads utilizing the channel-to-height operation (Yu et al. 2023) to predict semantic occupancy and 3D backward centripetal flow $(\mathcal{S}_{+t}, \mathcal{F}_{+t}) \in \mathbb{R}^{h\times w\times d}$.

In the training process, we employ multiple losses, including cross-entropy loss, Lovász loss (Berman, Rannen Triki, and Blaschko 2018), and binary occupancy loss, to constrain the semantics and geometries of occupancy predictions $\mathcal{S}_{1:f}$. The l_1 loss is used to supervise flow predictions $\mathcal{F}_{1:f}$.

3.3 Action-Controllable Generation

Due to the inherent complexity of the real world, the motion states of the ego vehicle are crucial for the world model to understand how the agent interacts with its environment. Therefore, to fully comprehend the environment, we propose leveraging diverse action conditions to empower DriveOccWorld with the capability for controllable generation.

Diverse Action Conditions include multiple formats: **(1)** ***Velocity*** is defined at a given time step as (v_x, v_y), representing the speeds of the ego vehicle decomposed along the x and y axes in m/s. **(2)** ***Steering Angle*** is collected from the steering feedback sensor. Following VAD, we convert it into curvature in m^{-1}, indicating the reciprocal of the turning radius while considering the geometric structure of the ego car. **(3)** ***Trajectory*** represents the movement of the ego vehicle's location to the next timestamp, formulated as $(\triangle x, \triangle y)$ in meters. It is widely used as the output of end-to-end methods, including our planner $\mathcal{P}$. **(4)** ***Commands*** consist of go forward, turn left, and turn right, which represent the highest-level intentions for controlling the vehicle.

Unified Conditioning Interface is designed to incorporate heterogeneous action conditions into a coherent embedding, inspired by (Gao et al. 2024; Wang et al. 2024b). We first encode the required actions via Fourier embeddings (Tancik et al. 2020), which are then concatenated and fused via learned projections to align with the dimensions of the conditional cross-attention layers in $\mathcal{W}_\mathcal{D}$. This method enables efficient integration of flexible action conditions into controllable generation, with experiments demonstrating that the unified interface with conditional cross-attention provides superior controllability compared to other approaches such as additive embeddings.

3.4 End-to-End Planning with World Model

Leveraging the future forecasting capabilities of our world model, as illustrated in Figure 2, we introduce a planner that employs an occupancy-based cost function to enforce safety constraints during the planning process.

Method	mIoU_c	mIoU_f	$\tilde{\text{mIoU}}_f$	VPQ_f	VPQ^*_f
SPC	1.3	failed	failed	–	–
CONet-C (Wang et al. 2023b)	12.2	11.5	11.7	–	–
PowerBEV-3D (Li et al. 2023a)	23.1	21.3	21.9	–	–
Cam4DOcc (Ma et al. 2024b)	**31.3**	26.8	28.0	21.4	–
Drive-OccWorld$^{\mathcal{A}}$ (Ours)	29.4	<u>28.6</u>	<u>28.7</u>	<u>22.6</u>	<u>32.0</u>
Drive-OccWorld$^{\mathcal{P}}$ (Ours)	<u>29.6</u>	**28.8**	**29.0**	**23.3**	**33.2**

SPC: SurroundDepth (Wei et al. 2023) + PCPNet (Luo et al. 2023) + Cylinder3D (Zhu et al. 2021)

Table 1: Comparisons of Inflated Occupancy and Flow Forecasting on the nuScenes validation set.

Method	mIoU_c			mIoU_f (2 s)			$\tilde{\text{mIoU}}_f$
	GMO	GSO	mean	GMO	GSO	mean	GMO
SPC	5.9	3.3	4.6	1.1	1.4	1.2	1.1
PowerBEV-3D (Li et al. 2023a)	5.9	–	–	5.3	–	–	5.5
CONet-C (Wang et al. 2023b)	9.6	17.2	13.4	7.4	17.3	12.4	7.9
Cam4DOcc (Ma et al. 2024b)	11.0	**17.8**	14.4	9.2	<u>17.8</u>	13.5	9.7
DriveOccWorld$^{\mathcal{A}}$ (Ours)	<u>12.3</u>	16.8	<u>14.5</u>	<u>10.2</u>	16.7	<u>13.5</u>	<u>10.4</u>
DriveOccWorld$^{\mathcal{P}}$ (Ours)	**12.6**	<u>17.5</u>	**15.1**	**10.3**	**17.9**	**14.1**	**10.8**

SPC: SurroundDepth (Wei et al. 2023) + PCPNet (Luo et al. 2023) + Cylinder3D (Zhu et al. 2021)

Table 2: Comparisons of Fine-grained Occupancy Forecasting on nuScenes-Occupancy. GMO indicates general movable objects, while GSO refers to general static objects.

Occupancy-based Cost Function is designed to ensure the safe driving of the ego vehicle. It consists of multiple cost factors: **(1)** ***Agent-Safety Cost*** constrains the ego vehicle from colliding with other agents, such as pedestrians and vehicles. It penalizes trajectory candidates that overlap with grids occupied by other road users. Additionally, trajectories that are too close to other agents, in terms of lateral or longitudinal distance, are also restricted to avoid potential collisions. **(2)** ***Road-Safety Cost*** ensures the vehicle remains on the road. It extracts road layouts from occupancy predictions, penalizing trajectories that fall outside the drivable area. **(3)** ***Learned-Volume Cost*** is inspired by ST-P3 (Hu et al. 2022). It employs a learnable head based on $\boldsymbol{F}^{bev}_{+t}$ to generate a 2D cost map, enabling a more comprehensive evaluation of occupancy grids in complex environments.

The total cost function is the summation of the above cost factors. Following the approach of ST-P3, a trajectory sampler generates a set of candidate trajectories $\tau^*_{+t} \in \mathbb{R}^{N_\tau \times 2}$ distributed across the 2D grid map surrounding the ego vehicle. Subsequently, the trajectory planner $\mathcal{P}$ selects the optimal trajectory τ_{+t} by minimizing the total cost function, while simultaneously ensuring agent and road safety.

BEV Refinement is introduced to further refine the trajectory using the latent features of BEV embeddings $\boldsymbol{F}^{bev}_{+t}$. We encode τ_{+t} into an embedding and concatenate it with a command embedding to form an ego query, which performs cross-attention with $\boldsymbol{F}^{bev}_{+t}$ to extract fine-grained representations of the environment. The final trajectory is predicted based on the refined ego query through MLPs.

The planning loss $\mathcal{L}_{plan}$ consists of three components: a max-margin loss introduced by (Sadat et al. 2020) to constrain the safety of trajectory candidates τ^*_{+t}, a naive l_2 loss for imitation learning, and a collision loss that ensures the planned trajectory avoids grids occupied by obstacles.

No.	Action Condition traj	vel	angle	cmd	mIoU_c	mIoU_f (1 s)	$\tilde{\text{mIoU}}_f$	VPQ^*_f
1					28.7	26.4	26.8	33.5
2	✓				28.5	<u>27.6</u>$_{\uparrow \mathbf{1.2}}$	<u>27.8</u>$_{\uparrow \mathbf{1.0}}$	<u>33.7</u>$_{\uparrow 0.2}$
3		✓			28.9$_{\uparrow 0.2}$	27.5$_{\uparrow \mathbf{1.1}}$	<u>27.8</u>$_{\uparrow \mathbf{1.0}}$	33.9$_{\uparrow 0.4}$
4			✓		28.9$_{\uparrow 0.2}$	26.8$_{\uparrow 0.4}$	27.2$_{\uparrow 0.4}$	34.2$_{\uparrow 0.7}$
5				✓	**29.2**$_{\uparrow \mathbf{0.5}}$	26.8$_{\uparrow 0.4}$	27.3$_{\uparrow 0.5}$	34.7$_{\uparrow 1.2}$
6	✓	✓	✓	✓	<u>29.0</u>$_{\uparrow 0.3}$	<u>27.6</u>$_{\uparrow \mathbf{1.2}}$	<u>27.8</u>$_{\uparrow \mathbf{1.0}}$	<u>35.0</u>$_{\uparrow 1.5}$
7	✓$^{\mathcal{P}}$				**29.2**$_{\uparrow \mathbf{0.5}}$	**27.9**$_{\uparrow \mathbf{1.5}}$	**28.1**$_{\uparrow \mathbf{1.3}}$	**35.1**$_{\uparrow \mathbf{1.6}}$

Table 3: Comparisons of controllability under diverse action conditions. ✓$^{\mathcal{P}}$ denotes the predicted trajectory.

Action Condition	L2 (m) ↓				Collision (%) ↓			
	1s	2s	3s	Avg.	1s	2s	3s	Avg.
GT trajectory	**0.26**	**0.52**	**0.89**	**0.56**	**0.02**	**0.11**	**0.36**	**0.16**
Pred trajectory	0.32	0.75	1.49	0.85	0.05	0.17	0.64	0.29

Table 4: Planning upper bound of the Drive-OccWorld when using GT trajectory as action condition.

4 Experiments

4.1 Setup

Tasks Definition. We validate the effectiveness of Drive-OccWorld on three types of tasks: **(1)** ***Inflated Occupancy and Flow Forecasting*** is introduced in Cam4DOcc (Ma et al. 2024b), predicting the future states of movable objects with dilated occupancy patterns, where voxel-wise semantic and instance labels are assigned using bounding box annotations. The 3D backward centripetal flow points from the voxel at time t to its corresponding 3D instance center at timestamp $t-1$. **(2)** ***Fine-grained Occupancy Forecasting*** utilizes voxel-level semantic occupancy annotations provided by nuScenes-Occupancy, which include both movable objects and static environments. **(3)** ***End-to-end Planning*** follows the open-loop evaluation on nuScenes.

Metrics. **(1)** ***Occupancy forecasting*** is evaluated using the mIoU metric. Following Cam4DOcc, we assess the current moment ($t = 0$) with mIoU_c and the future timestamps ($t \in [1, f]$) with mIoU_f, along with a quantitative indicator $\tilde{\text{mIoU}}_f$ weighted by timestamp. **(2)** ***Flow predictions*** are evaluated through instance association using the video panoptic quality VPQ_f metric. We further report the flow forecasting results denoted as VPQ^*_f, utilizing a simple yet efficient center clustering technique, where the predicted object centers are clustered based on their relative distances. **(3)** ***End-to-end planning*** is evaluated using the L2 distance from ground truth trajectories and the object collision rate.

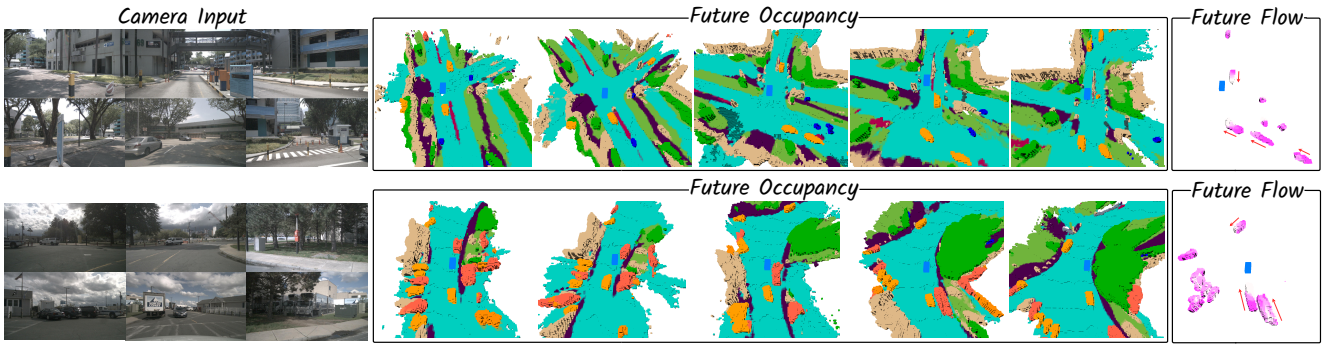

Figure 4: Qualitative results of 4D occupancy and flow forecasting. The results are presented at various future timestamps.

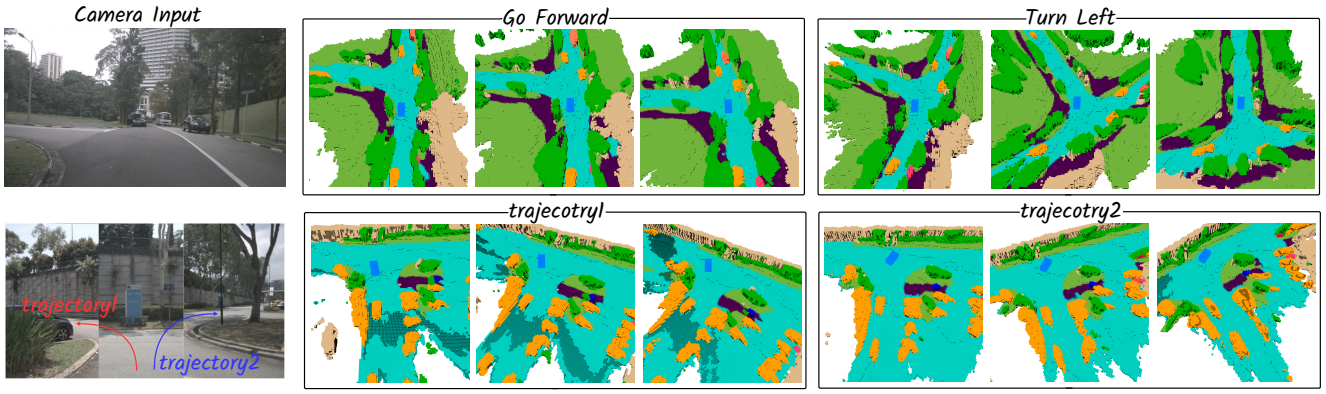

Figure 5: Qualitative results of controllable generation, using the high-level *command* or low-level *trajectory* conditions.

4.2 Main Results of 4D Occupancy Forecasting

We verify the quality of 4D occupancy forecasting and the controllable generation capabilities, reporting performance conditioned on GT actions as Drive-OccWorld$^{\mathcal{A}}$, and results conditioned on predicted trajectories as Drive-OccWorld$^{\mathcal{P}}$.

Inflated Occupancy and Flow Forecasting. Table 1 presents comparisons of inflated occupancy and flow forecasting on the nuScenes dataset. Drive-OccWorld outperforms Cam4DOcc on mIoU$_f$ by 2.0%, demonstrating a stronger ability to forecast future states. For future flow predictions, Drive-OccWorld$^{\mathcal{P}}$ consistently outperforms previous methods by 1.9% on VPQ$_f$, indicating a superior capability for modeling the motions of dynamic objects. Furthermore, by integrating the center clustering technique, Drive-OccWorld achieves remarkable results on VPQ$^*_f$, showcasing the effectiveness of instance proposal associations.

Fine-grained Occupancy Forecasting. Table 2 presents comparisons of fine-grained occupancy forecasting on the nuScenes-Occupancy. The results demonstrate that Drive-OccWorld achieves the best performance compared to all other approaches. Notably, Drive-OccWorld$^{\mathcal{P}}$ outperforms Cam4DOcc by 1.6% and 1.1% on mIoU for general movable objects at current and future timestamps, respectively, illustrating its ability to accurately locate movable objects for safe planning. Figure 4 provides qualitative results of the occupancy forecasting and flow predictions across frames.

Controllability. Table 3 examines controllability under various action conditions. Injecting any condition improves results compared to the baseline. Low-level conditions, such as trajectory and velocity, significantly enhance future forecasting (mIoU$_f$), while high-level conditions, such as commands, boost results at the current moment (mIoU$_c$). Incorporating more low-level conditions helps the world model better understand how the ego vehicle interacts with the environment, thereby improving forecasting performance.

Table 4 shows that using ground-truth trajectories as action conditions yields better planning results than predicted trajectories. However, using predicted trajectories slightly improves occupancy and flow forecasting quality, as indicated by comparisons in Table 3 (line 2 vs. line 7) and supported by Tables 1 and 2, where Drive-OccWorld$^{\mathcal{P}}$ outperforms Drive-OccWorld$^{\mathcal{A}}$. This performance gain may stem from planning constraints associated with predicted trajectories, allowing the planner to perform cross-attention between trajectories and BEV features. This process constrains the BEV features to account for ego-motion, thereby enhancing perception performance. Additionally, using predicted trajectories during training improves model learning to boost performance during testing.

Additionally, in Figure 5, we demonstrate the capability of Drive-OccWorld to simulate various future occupancies based on specific ego motions, showcasing the potential of generating plausible occupancy for autonomous driving.

Method	L2 (m) ↓ 1s	2s	3s	Avg.	Collision (%) ↓ 1s	2s	3s	Avg.
NMP (Zeng et al. 2019)	-	-	2.31	-	-	-	1.92	-
SA-NMP (Zeng et al. 2019)	-	-	2.05	-	-	-	1.59	-
FF (Hu et al. 2021)	0.55	1.20	2.54	1.43	0.06	0.17	1.07	0.43
EO (Khurana et al. 2022)	0.67	1.36	2.78	1.60	0.04	0.09	0.88	0.33
ST-P3† (Hu et al. 2022)	1.72	3.26	4.86	3.28	0.44	1.08	3.01	1.51
UniAD† (Hu et al. 2023b)	0.48	0.96	1.65	1.03	0.05	0.17	0.71	0.31
VAD-B† (Jiang et al. 2023)	0.54	1.15	1.98	1.22	0.10	0.24	0.96	0.43
OccNet† (Tong et al. 2023)	1.29	2.13	2.99	2.14	0.21	0.59	1.37	0.72
Drive-OccWorld$^{\mathcal{P}\dagger}$ (Ours)	**0.32**	**0.75**	**1.49**	**0.85**	**0.05**	**0.17**	**0.64**	**0.29**
ST-P3‡ (Hu et al. 2022)	1.33	2.11	2.90	2.11	0.23	0.62	1.27	0.71
UniAD‡ (Hu et al. 2023b)	0.44	0.67	0.96	0.69	0.04	0.08	0.23	0.12
VAD-B‡ (Jiang et al. 2023)	0.41	0.70	1.05	0.72	0.07	0.17	0.41	0.22
DriveWM‡ (Wang et al. 2024b)	0.43	0.77	1.20	0.80	0.10	0.21	0.48	0.26
Drive-OccWorld$^{\mathcal{P}\ddagger}$ (Ours)	**0.25**	**0.44**	**0.72**	**0.47**	**0.03**	**0.08**	**0.22**	**0.11**
UniAD‡* (Hu et al. 2023b)	0.20	0.42	0.75	0.46	0.02	0.25	0.84	0.37
VAD-B‡* (Jiang et al. 2023)	0.17	0.34	0.60	0.37	0.04	0.27	0.67	0.33
BEV-Planner‡* (Li et al. 2024)	0.16	0.32	0.57	0.35	0.00	0.29	0.73	0.34
Drive-OccWorld$^{\mathcal{P}\ddagger *}$ (Ours)	0.17	**0.31**	**0.49**	**0.32**	0.02	**0.24**	**0.62**	**0.29**

Table 5: End-to-end planning performance on nuScenes. † denotes the NoAvg protocol, ‡ represents the TemAvg protocol, and * indicates the use of ego status in the planner.

Conditional Norm semantic	ego	agent	mIoU_c	mIoU_f (1 s)	$\widetilde{\text{mIoU}}_f$	VPQ^*_f
			28.7	26.4	26.8	33.5
✓			$29.0_{\uparrow 0.3}$	$26.6_{\uparrow 0.2}$	$27.0_{\uparrow 0.2}$	33.2
	✓		$29.4_{\uparrow \mathbf{0.7}}$	$28.3_{\uparrow \mathbf{1.9}}$	$28.5_{\uparrow \mathbf{1.7}}$	32.6
		✓	$29.3_{\uparrow \mathbf{0.6}}$	$27.1_{\uparrow 0.7}$	$27.5_{\uparrow 0.7}$	$34.4_{\uparrow \mathbf{0.9}}$
✓	✓	✓	$\mathbf{29.4}_{\uparrow \mathbf{0.7}}$	$\mathbf{28.3}_{\uparrow \mathbf{1.9}}$	$\mathbf{28.6}_{\uparrow \mathbf{1.8}}$	$\mathbf{34.5}_{\uparrow \mathbf{1.0}}$

Table 6: Ablations on the conditional normalization.

4.3 End-to-end Planning with Drive-OccWorld

Table 5 presents the planning performance compared to existing end-to-end methods in terms of L2 error and collision rate. We provide results under different evaluation protocol settings from ST-P3 and UniAD. Specifically, NoAvg denotes the result at the corresponding timestamp, while TemAvg calculates metrics by averaging performances from 0.5s to the corresponding timestamp.

As shown in Table 5, Drive-OccWorld$^{\mathcal{P}}$ achieves superior planning performance compared to existing methods. For instance, Drive-OccWorld$^{\mathcal{P}\dagger}$ obtains relative improvements of 33%, 22%, and 9.7% on L2@1s, L2@2s, and L2@3s, respectively, compared to UniAD†. We attribute this improvement to the world model's capacity to accumulate world knowledge and envision future states. It effectively enhances the planning results for future timestamps and improves the safety and robustness of end-to-end planning.

Recent studies (Li et al. 2024) have explored the effect of ego status in planning modules. We compare our model, which incorporates ego status, to previous works and find that Drive-OccWorld maintains the highest performance at distant future timestamps. This highlights the effectiveness of continuous forecasting and planning.

Cond Interface addition	cross-attn	Fourier Embed	mIoU_c	mIoU_f (1 s)	$\widetilde{\text{mIoU}}_f$	VPQ^*_f
			28.7	26.4	26.8	33.5
✓		✓	$28.9_{\uparrow 0.2}$	$27.4_{\uparrow 1.0}$	$\mathbf{28.0}_{\uparrow 1.2}$	$34.2_{\uparrow 0.7}$
	✓		28.5	$27.1_{\uparrow 0.7}$	$27.4_{\uparrow 0.6}$	$33.9_{\uparrow 0.4}$
	✓	✓	$\mathbf{29.0}_{\uparrow 0.3}$	$\mathbf{27.6}_{\uparrow 1.2}$	$27.8_{\uparrow 1.0}$	$\mathbf{35.0}_{\uparrow 1.5}$

Table 7: Ablations on the action conditioning interface.

Cost Factors Agent	Road	Volume	BEV Refine	L2 (m) ↓ 0.5s	1s	1.5s	Avg.	Collision (%) ↓ 0.5s	1s	1.5s	Avg.
✗	✓	✓	✓	0.15	0.30	0.50	0.32	0.14	0.16	0.18	0.16
✓	✗	✓	✓	0.14	0.28	0.46	0.29	0.09	0.11	0.13	0.11
✓	✓	✗	✓	0.14	0.27	**0.44**	0.28	0.09	0.14	0.18	0.14
✓	✓	✓	✗	0.22	0.36	0.52	0.37	0.14	0.20	0.27	0.20
✓	✓	✓	✓	**0.11**	**0.26**	0.46	**0.28**	**0.04**	**0.11**	**0.13**	**0.09**

Table 8: Contributions of occupancy-based cost factors.

4.4 Ablation Study

The default configuration for the ablation experiments involves using one historical and the current images as input to predict the inflated occupancy over two future timestamps.

Conditional Normalization. In Table 6, we ablate the conditional normalization method while discarding the action conditions in Sec. 3.3 to avoid potential influence. The results indicate that each pattern yields gains, particularly with ego-motion aware normalization achieving a 1.9% increase in mIoU_f, highlighting the importance of ego status for future state forecasting. Additionally, agent-motion aware normalization enhances VQP^*_f by 0.9% by compensating for the movements of other agents.

Action Conditioning Interface. In Table 7, we investigate the method of injecting action conditions into the world decoder. Compared to adding conditions to BEV queries, cross-attention is a more effective approach for integrating prior knowledge into the generation process. Furthermore, Fourier embedding provides additional improvement by encoding conditions into latent space at high frequencies.

Occupany-based Costs. Table 8 ablates the occupancy-based cost function, and the results indicate that each cost factor contributes to safe planning, particularly highlighting that the absence of agent constraints results in a higher collision rate. Additionally, BEV refinement is vital as it provides more comprehensive 3D information about the environment.

5 Conclusion

We propose Drive-OccWorld, a 4D occupancy forecasting and planning world model for autonomous driving. Flexible action conditions can be injected into the world model for action-controllable generation. An occupancy-based planner is integrated with the world model for motion planning, considering both safety and the 3D structure of the environment. Experiments demonstrate that our method exhibits remarkable performance in occupancy and flow forecasting. Planning results are improved by leveraging the world model's capacity to accumulate world knowledge and envision future states, enhancing the safety of end-to-end planning.

Acknowledgments

We thank Jiangning Zhang and Jiang He for helpful discussions and valuable support on the paper. We thank all authors for their contributions. This work was supported by a Grant from The National Natural Science Foundation of China (No. 62103363).

References

Agro, B.; Sykora, Q.; Casas, S.; Gilles, T.; and Urtasun, R. 2024. UnO: Unsupervised Occupancy Fields for Perception and Forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14487–14496.

Ayoub, J.; Zhou, F.; Bao, S.; and Yang, X. J. 2019. From manual driving to automated driving: A review of 10 years of autoui. In *Proceedings of the 11th international conference on automotive user interfaces and interactive vehicular applications*, 70–90.

Berman, M.; Rannen Triki, A.; and Blaschko, M. B. 2018. The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4413–4421.

Blattmann, A.; Dockhorn, T.; Kulal, S.; Mendelevitch, D.; Kilian, M.; Lorenz, D.; Levi, Y.; English, Z.; Voleti, V.; Letts, A.; et al. 2023. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*.

Caesar, H.; Bankiti, V.; Lang, A. H.; Vora, S.; Liong, V. E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; and Beijbom, O. 2020. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11621–11631.

Chen, L.; Li, Y.; Huang, C.; Xing, Y.; Tian, D.; Li, L.; Hu, Z.; Teng, S.; Lv, C.; Wang, J.; et al. 2023. Milestones in autonomous driving and intelligent vehicles—Part I: Control, computing system design, communication, HD map, testing, and human behaviors. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 53(9): 5831–5847.

Gao, S.; Yang, J.; Chen, L.; Chitta, K.; Qiu, Y.; Geiger, A.; Zhang, J.; and Li, H. 2024. Vista: A Generalizable Driving World Model with High Fidelity and Versatile Controllability. *arXiv preprint arXiv:2405.17398*.

Guo, Z.; Gao, X.; Zhou, J.; Cai, X.; and Shi, B. 2023. Scenedm: Scene-level multi-agent trajectory generation with consistent diffusion models. *arXiv preprint arXiv:2311.15736*.

Ha, D.; and Schmidhuber, J. 2018. World models. *arXiv preprint arXiv:1803.10122*.

Hu, A.; Russell, L.; Yeo, H.; Murez, Z.; Fedoseev, G.; Kendall, A.; Shotton, J.; and Corrado, G. 2023a. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*.

Hu, P.; Huang, A.; Dolan, J.; Held, D.; and Ramanan, D. 2021. Safe local motion planning with self-supervised freespace forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12732–12741.

Hu, S.; Chen, L.; Wu, P.; Li, H.; Yan, J.; and Tao, D. 2022. St-p3: End-to-end vision-based autonomous driving via spatial-temporal feature learning. In *European Conference on Computer Vision*, 533–549. Springer.

Hu, Y.; Yang, J.; Chen, L.; Li, K.; Sima, C.; Zhu, X.; Chai, S.; Du, S.; Lin, T.; Wang, W.; et al. 2023b. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17853–17862.

Jia, F.; Mao, W.; Liu, Y.; Zhao, Y.; Wen, Y.; Zhang, C.; Zhang, X.; and Wang, T. 2023. Adriver-i: A general world model for autonomous driving. *arXiv preprint arXiv:2311.13549*.

Jiang, B.; Chen, S.; Xu, Q.; Liao, B.; Chen, J.; Zhou, H.; Zhang, Q.; Liu, W.; Huang, C.; and Wang, X. 2023. Vad: Vectorized scene representation for efficient autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8340–8350.

Khurana, T.; Hu, P.; Dave, A.; Ziglar, J.; Held, D.; and Ramanan, D. 2022. Differentiable raycasting for self-supervised occupancy forecasting. In *European Conference on Computer Vision*, 353–369. Springer.

Li, P.; Ding, S.; Chen, X.; Hanselmann, N.; Cordts, M.; and Gall, J. 2023a. PowerBEV: a powerful yet lightweight framework for instance prediction in bird's-eye view. *arXiv preprint arXiv:2306.10761*.

Li, X.; Ma, T.; Hou, Y.; Shi, B.; Yang, Y.; Liu, Y.; Wu, X.; Chen, Q.; Li, Y.; Qiao, Y.; et al. 2023b. Logonet: Towards accurate 3d object detection with local-to-global cross-modal fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17524–17534.

Li, X.; Zhang, Y.; and Ye, X. 2023. DrivingDiffusion: Layout-Guided multi-view driving scene video generation with latent diffusion model. *arXiv preprint arXiv:2310.07771*.

Li, Z.; Wang, W.; Li, H.; Xie, E.; Sima, C.; Lu, T.; Qiao, Y.; and Dai, J. 2022. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *European conference on computer vision*, 1–18. Springer.

Li, Z.; Yu, Z.; Lan, S.; Li, J.; Kautz, J.; Lu, T.; and Alvarez, J. M. 2024. Is ego status all you need for open-loop end-to-end autonomous driving? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14864–14873.

Luo, Z.; Ma, J.; Zhou, Z.; and Xiong, G. 2023. Pcpnet: An efficient and semantic-enhanced transformer network for point cloud prediction. *IEEE Robotics and Automation Letters*.

Ma, E.; Zhou, L.; Tang, T.; Zhang, Z.; Han, D.; Jiang, J.; Zhan, K.; Jia, P.; Lang, X.; Sun, H.; et al. 2024a. Unleashing Generalization of End-to-End Autonomous Driving with Controllable Long Video Generation. *arXiv preprint arXiv:2406.01349*.

Ma, J.; Chen, X.; Huang, J.; Xu, J.; Luo, Z.; Xu, J.; Gu, W.; Ai, R.; and Wang, H. 2024b. Cam4docc: Benchmark for camera-only 4d occupancy forecasting in autonomous driving applications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21486–21495.

Min, C.; Zhao, D.; Xiao, L.; Nie, Y.; and Dai, B. 2023. Uniworld: Autonomous driving pre-training via world models. *arXiv preprint arXiv:2308.07234*.

Min, C.; Zhao, D.; Xiao, L.; Zhao, J.; Xu, X.; Zhu, Z.; Jin, L.; Li, J.; Guo, Y.; Xing, J.; et al. 2024. Driveworld: 4d pre-trained scene understanding via world models for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15522–15533.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Sadat, A.; Casas, S.; Ren, M.; Wu, X.; Dhawan, P.; and Urtasun, R. 2020. Perceive, predict, and plan: Safe motion planning through interpretable semantic representations. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII 16*, 414–430. Springer.

Tancik, M.; Srinivasan, P.; Mildenhall, B.; Fridovich-Keil, S.; Raghavan, N.; Singhal, U.; Ramamoorthi, R.; Barron, J.; and Ng, R. 2020. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in neural information processing systems*, 33: 7537–7547.

Tong, W.; Sima, C.; Wang, T.; Chen, L.; Wu, S.; Deng, H.; Gu, Y.; Lu, L.; Luo, P.; Lin, D.; et al. 2023. Scene as occupancy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8406–8415.

Van Den Oord, A.; Vinyals, O.; et al. 2017. Neural discrete representation learning. *Advances in neural information processing systems*, 30.

Wang, L.; Zheng, W.; Ren, Y.; Jiang, H.; Cui, Z.; Yu, H.; and Lu, J. 2024a. OccSora: 4D Occupancy Generation Models as World Simulators for Autonomous Driving. *arXiv preprint arXiv:2405.20337*.

Wang, X.; Zhu, Z.; Huang, G.; Chen, X.; and Lu, J. 2023a. Drivedreamer: Towards real-world-driven world models for autonomous driving. *arXiv preprint arXiv:2309.09777*.

Wang, X.; Zhu, Z.; Xu, W.; Zhang, Y.; Wei, Y.; Chi, X.; Ye, Y.; Du, D.; Lu, J.; and Wang, X. 2023b. Openoccupancy: A large scale benchmark for surrounding semantic occupancy perception. *arXiv preprint arXiv:2303.03991*.

Wang, Y.; He, J.; Fan, L.; Li, H.; Chen, Y.; and Zhang, Z. 2024b. Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14749–14759.

Wei, Y.; Zhao, L.; Zheng, W.; Zhu, Z.; Rao, Y.; Huang, G.; Lu, J.; and Zhou, J. 2023. Surrounddepth: Entangling surrounding views for self-supervised multi-camera depth estimation. In *Conference on robot learning*, 539–549. PMLR.

Yang, J.; Gao, S.; Qiu, Y.; Chen, L.; Li, T.; Dai, B.; Chitta, K.; Wu, P.; Zeng, J.; Luo, P.; et al. 2024a. Generalized predictive model for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14662–14672.

Yang, Z.; Chen, L.; Sun, Y.; and Li, H. 2024b. Visual point cloud forecasting enables scalable autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14673–14684.

Yu, Z.; Shu, C.; Deng, J.; Lu, K.; Liu, Z.; Yu, J.; Yang, D.; Li, H.; and Chen, Y. 2023. Flashocc: Fast and memory-efficient occupancy prediction via channel-to-height plugin. *arXiv preprint arXiv:2311.12058*.

Zeng, W.; Luo, W.; Suo, S.; Sadat, A.; Yang, B.; Casas, S.; and Urtasun, R. 2019. End-to-end interpretable neural motion planner. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8660–8669.

Zhang, L.; Xiong, Y.; Yang, Z.; Casas, S.; Hu, R.; and Urtasun, R. 2023. Learning unsupervised world models for autonomous driving via discrete diffusion. *arXiv preprint arXiv:2311.01017*.

Zheng, W.; Chen, W.; Huang, Y.; Zhang, B.; Duan, Y.; and Lu, J. 2023. Occworld: Learning a 3d occupancy world model for autonomous driving. *arXiv preprint arXiv:2311.16038*.

Zhu, X.; Zhou, H.; Wang, T.; Hong, F.; Ma, Y.; Li, W.; Li, H.; and Lin, D. 2021. Cylindrical and asymmetrical 3d convolution networks for lidar segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9939–9948.