

Gaussian-based World Model: Gaussian Priors for Voxel-Based Occupancy Prediction and Future Motion Prediction

Tuo Feng, Wenguan Wang*, Yi Yang

ReLER, CCAI, Zhejiang University

<https://github.com/FengZicai/GWM>

Abstract

In autonomous driving, accurately predicting occupancy and motion is crucial for safe navigation within dynamic environments. However, existing methods often suffer from difficulties in handling complex scenes and uncertainty arising from sensor data. To address these issues, we propose a new Gaussian-based World Model (GWM), seamlessly integrating raw multi-modal sensor inputs. In 1st stage, Gaussian representation learner utilizes self-supervised pre-training to learn robust Gaussian representation. Gaussian representation integrates semantic and geometric information and establishes a robust probabilistic understanding of the environment. In 2nd stage, GWM seamlessly integrates learning, simulation, and planning into a unified framework, empowering the uncertainty-aware simulator & planner to jointly forecast future scene evolutions and vehicle trajectories. Simulator generates future scene predictions by modeling both static and dynamic elements, while planner calculates optimal paths to minimize collision risks, thus enhancing navigation safety. Overall, GWM employs a sensor-to-planning world model that directly processes raw sensor data, setting it apart from previous methods. Experiments show that GWM outperforms state-of-the-art approaches by **1.46%** in semantic comprehension and **0.07m** in motion prediction. Moreover, we provide an in-depth analysis of Gaussian representations under complex scenarios.

1. Introduction

The safety of autonomous driving (AD) systems critically relies on accurately perceiving and predicting dynamic environments [3, 43, 64]. This involves not only reconstructing the driving scene but also forecasting the motion of objects, enabling vehicles to anticipate potential hazards and make informed trajectory decisions [12, 39, 51]. However, despite advancements in sensor technologies and compu-

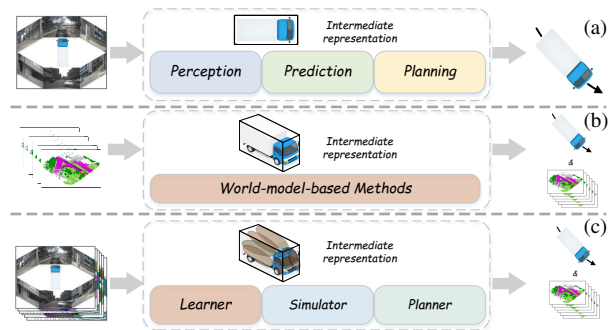


Figure 1. **Illustration of comparisons (§ 1).** (a) Sequential perception-prediction-planning pipeline in end-to-end methods like ST-P3 [23]. (b) World-model-based methods like OccWorld [87], which predict both occupancy and vehicle trajectories. (c) GWM directly utilizes raw multi-modal sensor data through a unified learning, simulation, and planning pipeline.

tational methods, current approaches still face significant challenges [13, 36, 68]. Earlier end-to-end AD frameworks (e.g., ST-P3 [23]) rely on Bird’s-Eye View (BEV) for scene prediction and planning. They lack the ability to capture the vertical dimension of the 3D scene and fail to achieve fine-grained 3D scene reconstruction (see Fig. 1(a,c)). Moreover, former scene reconstruction struggles to capture sufficient detail, especially in complex and rapidly changing settings [4, 13, 54]. Additionally, sensor uncertainties (e.g., LiDAR noise and image occlusions) further complicate accurate 3D occupancy estimation, ultimately reducing the precision of motion predictions, a key factor for effective driving decisions [4, 16, 29].

Occupancy-based approaches [12, 38, 41, 47, 57, 58, 64, 65, 68, 79, 85, 89] offer more granular 3D representations of the environment than BEV-based methods [25, 36, 40, 43, 84], capturing spatial structures at higher resolutions. For instance, OccNeRF [81] derives 3D occupancy grids from multi-view images, and significantly improves occupancy prediction accuracy. While these methods enhance static scene understanding, they often fall short with dynamic objects (e.g., moving vehicles or pedestrians) due to insufficient temporal modeling. Without the ability to predict future states, it remains challenging to handle environmental

*Corresponding author.

changes that are critical for reliable path planning. Recent world-model-based frameworks (e.g., OccWorld [87] and OccLLaMA [67]), as shown in Fig. 1(b), use temporal modeling and integrate both the vehicle’s motion planning and future environment forecasting simultaneously.

Our Gaussian-based World Model (GWM) directly utilizes raw multi-modal sensor data (i.e., camera and LiDAR) as inputs and has two stages. In *1st stage*, we introduce a *Gaussian representation learner*, and leverage self-supervised pre-training [9, 10] to map multi-modal visual features into a physically interpretable 3D Gaussian representation space, achieving the integrated encoding of semantic and 3D geometric information. Subsequently, by differentially rendering these Gaussian representations back into multi-view images, the model learns cross-modal perceptual priors and physically constrained embeddings from the visual data. This process ultimately provides a high-quality, robust Gaussian representation for the *2nd stage*.

By combining LiDAR prior with Gaussian [29], *Gaussian representation learner* effectively combines geometric depth information with semantic visual understanding. Gaussian representation achieves a more detailed comprehension of static and dynamic components within the environment [71, 88, 90]. The *learner* provides a probabilistic foundation for the system to effectively model uncertainties and variabilities within the scene. This approach offers advantages over former voxel-based methods, especially in rapidly changing and complex environments [18, 41, 64]. Moreover, the probabilistic nature of Gaussian representations enables GWM to model sensor uncertainties and enhance robustness in real-world applications.

In *2nd stage*, GWM has a unified learning, simulation, and planning pipeline. Based on learned Gaussian representation, GWM employs an *uncertainty-aware simulator* and an *uncertainty-aware planner* to model the temporal evolution of the scene and objects. The *simulator* propagates uncertainties and generates future scene predictions by modeling both static and dynamic elements, while the *planner* with uncertainty loss calculates optimal vehicle trajectories to minimize collision risks, thus enhancing navigation safety (see Fig. 1(c)). By accounting for both spatial and temporal factors, GWM can anticipate future changes in the environment, enabling more informed decision-making and improving the safety of AD systems.

Our method is validated on Occ3D [57] and nuScenes [2] datasets against state-of-the-art approaches like OccLLaMA [67] and OccWorld [87]. Specifically, GWM achieves better performance than the baselines in both 4D occupancy forecasting (10.12% vs 8.66% Avg. mIoU) and motion planning (1.13m vs 1.20m Avg. L2 Error and 0.59% vs 0.70% Avg. Collision Rate). Moreover, we provide a comprehensive analysis of Gaussian representations under complex scenarios. These findings underscore the effectiveness of

our GWM designs, providing a more reliable and predictive framework for AD systems.

2. Related Work

End-to-end AD. AD algorithms have evolved significantly in recent decades, transitioning from modular pipelines [19, 34] to end-to-end models [24, 28] that predict planning trajectories directly from raw sensor data. End-to-end methods unify perception, prediction, and planning within a single system, simplifying the traditional multi-stage pipeline [23, 26, 84]. One prominent direction in end-to-end AD is leveraging BEV representations [24, 28]. Detection [25, 37, 40, 62, 78] and segmentation [48] tasks in BEV have laid the groundwork for trajectory planning by extracting spatially structured information. Another line focuses on improving representation [8, 56, 82] and connectivity [70]. Despite the advancements, traditional methods often emphasize object motion and overlook intricate environmental details and contextual understanding [17, 21, 28, 84]. Recent works aim to bridge this gap by incorporating richer contextual data. For instance, P3 [53] and ST-P3 [23] learn differentiable occupancy representations as cost factors for safe maneuvering. Fusion-based approaches [77, 78] integrate multi-modal data to provide more comprehensive input for the planner, achieving remarkable improvements.

Unlike previous methods that focus on dynamic elements, our approach introduces a world model that predicts the progression of both dynamic and static components in the environment. By modeling the joint evolution of surrounding agents and static structures, our method captures fine-grained spatial and semantic information, enabling more accurate planning and safer AD. This holistic perspective addresses the limitations of prior works.

World Models in AD. World models in AD are designed to understand and predict dynamic environments, enabling autonomous systems to navigate safely and efficiently [11, 60]. They aim to predict future scenes based on actions and observations [20, 67, 72]. Different models represent the scene in various representations spaces, which can be divided into 2D image representations [15, 22, 35, 63, 66, 73, 86], 3D point clouds representations [30, 31, 46, 69, 83], and 3D occupancy representations [1, 61, 67, 87]. Visual world models using 2D image representations leverage StableDiffusion [52] to generate diverse driving sequences [15, 22, 35, 63, 73, 74], lacking a deeper understanding of the 3D driving environment. 3D point cloud representations miss critical semantic information and are not suitable for vision-based or fusion-based systems. Thus, combining 3D scene representation with semantic understanding offers a promising approach to modeling scene evolution [1, 86].

Several concurrent studies (e.g., OccWorld [87] and OccLLaMA [67]) attempt to integrate world models into AD systems. However, they exhibit several limitations: (1) re-

liance on precomputed occupancy inputs, (2) degraded accuracy in long-term predictions, and (3) inadequate evaluation under complex driving scenarios. Specifically, OccWorld fails to effectively utilize multimodal sensor data, while Oc-cLLaMA demonstrates information degradation in its action representation space. In contrast, our GWM uniquely integrates raw multi-modal sensor inputs through a unified learning, simulation, and planning framework. Although GaussianWorld [90] recently introduced 3D Gaussian splatting [6, 14, 29], it has differences from GWM: insufficient analysis of Gaussian representations in complex scenarios, and incomplete design and experiment for motion planning.

3. Methodology

We propose GWM, a two-stage framework that directly processes raw multi-modal sensor inputs, as illustrated in Fig. 2. In §3.1, we present *1st stage* and *Gaussian representation learner*, which harness self-supervised pre-training to learn robust Gaussian representations. Subsequently, §3.2 details *2nd stage* and *uncertainty-aware simulator & planner*, which leverage occupancy uncertainty to jointly forecast scene evolution and optimize trajectories.

3.1. 1st Stage: Gaussian Representation Learner

In *1st stage*, GWM fuses multi-modal sensor data (*i.e.*, LiDAR points P^T , multi-view images and semantic maps I^T) to construct a robust 3D Gaussian representation that captures both semantic and geometric cues while modeling inherent sensor uncertainties. As shown in the Fig. 2, the 3D and 2D encoders extract complementary features, which are then aggregated by the *Gaussian representation learner* ϕ . To refine this representation, GWM differentially renders the 3D Gaussians back into 2D space [29] and compares them against the original RGB and semantic maps. Through this self-supervised cross-modal alignment, the model acquires physically constrained embeddings and perceptual priors, leading to a more reliable understanding of scene structure and object dynamics.

Gaussian Representation Learner. GWM adopts 3D Gaussian splatting (3DGS) [29], incorporating a Gaussian-based 3D representation into the world model. By representing scenes as sparse 3D semantic Gaussians (each defining a flexible region of interest along with semantic features) our approach surpasses the limitations of fixed-grid representations. Specifically, a scene is modeled as a set of 3D Gaussians, each defined by parameters such as mean, covariance, RGB, and semantics. We aggregate information from images and point clouds, progressively optimizing these parameters through the *Gaussian representation learner* ϕ . While structure-from-motion (SfM) priors [29, 55] yield only partial reconstructions from sparse viewpoints [88], aligning LiDAR priors with multi-camera images provides more accurate geometric constraints [88]. Accordingly, our

Gaussian representation integrates multimodal sensor data by incorporating the LiDAR prior into the 3D Gaussians, ensuring robust and high-fidelity reconstructions.

LiDAR prior P is colorized via calibrated camera projection and used to initialize the Gaussians. We describe each scene as a collection of 3D Gaussians $\mathcal{G} = \{\mathbf{G}_i \in \mathbb{R}^d \mid i = 1, \dots, M\}$, where \mathcal{G} contains M Gaussians. Each 3D Gaussian \mathbf{G} is expressed as a d -dimensional vector in the form $(\mathbf{m} \in \mathbb{R}^3, \mathbf{s} \in \mathbb{R}^3, \mathbf{r} \in \mathbb{R}^4, \mathbf{c} \in \mathbb{R}^6)$, where $d = 16$, and \mathbf{m} , \mathbf{s} , \mathbf{r} denote the mean, scale, rotation vectors, respectively. \mathbf{c} is a tensor that contains a 3-channel RGB image and a 3-channel semantic map (predicted by SAM [32]). The value of a semantic Gaussian distribution \mathbf{g} evaluated at point \mathbf{p} is

$$\mathbf{g}(\mathbf{p}; \mathbf{m}, \mathbf{s}, \mathbf{r}, \mathbf{c}) = \exp\left(-\frac{1}{2}(\mathbf{p} - \mathbf{m})^\top \Sigma^{-1}(\mathbf{p} - \mathbf{m})\right) \mathbf{c}, \quad (1)$$

$$\Sigma = \mathbf{R} \mathbf{S} \mathbf{S}^\top \mathbf{R}^\top, \quad \mathbf{S} = \text{diag}(\mathbf{s}), \quad \mathbf{R} = \text{q2r}(\mathbf{r}),$$

where \mathbf{p} is the position of the LiDAR prior; Σ represents the covariance matrix; $\text{diag}(\cdot)$ constructs a diagonal matrix from a vector; and $\text{q2r}(\cdot)$ transforms a quaternion into a rotation matrix. GWM enhances the world model by integrating Gaussian representation into voxel-based scene representations, and further enriches by high-resolution LiDAR points and multi-view images. Gaussian representations play a pivotal role in modeling uncertainties inherent in data. By assigning the Gaussian distribution to each point in space, GWM encapsulates not only the mean position but also the variance, effectively modeling the confidence in measurements. Gaussian representations establish a probabilistic foundation, enhancing the model’s capacity to navigate the uncertainties and variabilities of dynamic environments.

3.2. 2nd Stage: Learner-Simulator-Planner Pipeline

In *2nd stage*, GWM adopts a unified learner, simulator, and planner pipeline. Specifically, We introduce a simulator, ψ , to model environment evolution and a planner, θ , to generate trajectories for the ego vehicle.

Learner. *Gaussian representation learner* ϕ first learns the Gaussian representation, then occupancy uncertainty for a 3D voxel is computed by aggregating the Gaussian distribution values evaluated at that location. The occupancy uncertainty for voxel \mathbf{v} can be expressed as:

$$z(\mathbf{v}; \mathcal{G}) = \sum_{i \in \mathcal{N}(\mathbf{v})} \mathbf{g}_i(\mathbf{v}; \mathbf{m}_i, \mathbf{s}_i, \mathbf{r}_i, \mathbf{c}_i), \quad (2)$$

where $\mathcal{N}(\mathbf{v})$ is the set of neighboring Gaussians for the voxel at \mathbf{v} , and $\mathcal{N}(\mathbf{v})$ is obtained by considering the voxel position \mathbf{v} and Gaussian’s means \mathbf{m} and scale property \mathbf{s} [27]. In GWM, each voxel is augmented with probabilistic information derived from the Gaussian representations, creating a rich representation that encapsulates both occupancy and uncertainty. This is particularly beneficial in scenarios with noisy sensor data or when dealing with partial occlusions. Our GWM forms the foundation for further processing

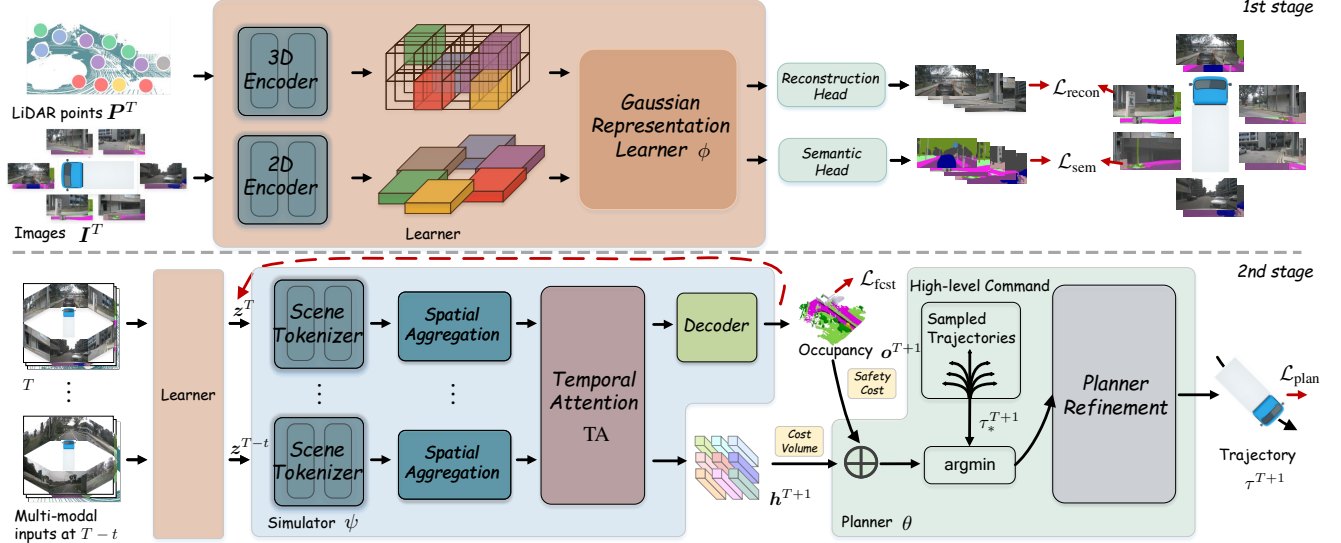


Figure 2. **Overview of our GWM w for 4D occupancy forecasting and motion planning in AD (§3).** Our framework includes three key components: 1) a Gaussian representation learner that encodes 3D scenes using Gaussian-based representations to capture spatial and semantic information; 2) a simulator that generates future scene predictions based on dynamic inputs; and 3) a planner, which produces safe and efficient driving trajectories. The semantic map is predicted by SAM [32].

and is essential for handling dynamic environments. Occlusions pose significant challenges in 3D occupancy prediction. GMW addresses this by maintaining visibility probabilities for each Gaussian. When a Gaussian is observed from multiple views, the system refines its occupancy probability, improving the accuracy of occlusion handling.

Uncertainty-Aware Simulator. Building upon occupancy uncertainty z , GWM enhances its capabilities for dynamic 4D scene modeling by integrating temporal dimensions. The simulator, ψ , forms a 4D occupancy uncertainty $\mathbf{Z} = \{z_i\}_{i=T-t}^T$, where each z_i corresponds to the occupancy uncertainty at timestamp $T-t$. T indicates the current time step, and t denotes the number of historical frames. This integration aims to capture the evolution of the environment. 4D occupancy uncertainty contains spatial information (position and occupancy probability) and reflects temporal dynamics. The *simulator* ψ takes as input the past 4D occupancy uncertainty \mathbf{Z} and predicts future scene evolutions after a certain time interval. The *simulator* ψ then operates on \mathbf{Z} :

$$\psi(z^T, \dots, z^{T-t}) = \mathbf{o}^{T+1}, \quad (3)$$

where \mathbf{o}^{T+1} is the predicted occupancy at timestamp $T+1$. To comprehensively model scene evolution, it is essential to consider both the spatial relationships within each past scene and the temporal relationships of scenes across different timestamps. Hence, we adopt a spatial-temporal transformer [87] to instantiate ψ . ψ produces scene tokens by applying a scene tokenizer. The scene tokens are then down-sampled with a factor of 4. We repeat this procedure K times to obtain high-level scene tokens \mathbf{H} . Next, we apply masked temporal attention TA to the set of high-level scene

tokens $\{h_v^T, \dots, h_v^{T-t}\}$ at each position v to predict the corresponding token h_v^{T+1} of the next scene:

$$h_v^{T+1} = \text{TA}(h_v^T, \dots, h_v^{T-t}). \quad (4)$$

TA blocks the influence of future tokens on past tokens. Here, h_v^t represents the v -th scene token at timestamp t .

The *simulator* 1) integrates multiple 3D occupancy uncertainty to construct a 4D representation and 2) employs a spatiotemporal Transformer with temporal attention mechanisms to enhance the accuracy of future scene predictions. Its advantages include effectively combining spatial and temporal information, enabling more precise modeling of scene evolution and allowing it to capture complex dynamic changes (see §C.1 in supplementary).

Uncertainty-Aware Planner. Building upon h^{T+1} and \mathbf{o}^{T+1} , the *planner* θ forecasts the ego vehicle’s trajectory. We design the *uncertainty-aware planner* by exploring a range of candidate trajectories and selecting the one that minimizes the learned cost function [5, 23, 53, 80]. We formulate the cost function to fully leverage the learned occupancy \mathbf{o}^{T+1} and rich prior knowledge, ensuring both the safety and smoothness of τ^{T+1} . The cost function includes: 1) Cost Volume: Inspired by ST-P3 [23], we employ a learnable module based on the BEV representation h_{bev}^{T+1} derived from h^{T+1} . h_{bev}^{T+1} represents the occupancy uncertainty (*i.e.*, occupancy probability field) in the BEV view. This module produces a cost volume that encapsulates detailed occupancy uncertainty of the environment, allowing for a thorough evaluation to support safe planning. 2) Safety Cost: Planned trajectories should avoid overlapping with regions used by other agents or road elements. This cost penalizes trajectory

options that intersect with regions occupied by other agents. Occupancy information from \mathbf{o}^{T+1} is used to assess collision risks with dynamic obstacles. A sampler [5, 23, 49] generates a set of trajectory candidates τ_*^{T+1} based on high-level commands¹. Then, θ selects the candidate trajectory τ^{T+1} by minimizing the total cost. Additionally, the candidate trajectory τ^{T+1} is encoded as an ego query. This query performs cross-attention with the future occupancy probability $\mathbf{h}_{\text{bev}}^{T+1}$, drawing detailed environmental information pertinent to the trajectory [23, 75]. The enhanced ego query is then used to predict the final trajectory.

Uncertainty Loss. To enhance the *planner*'s reliability, the uncertainty loss is constructed by computing *reconstruction inconsistency score* $r^{T+1} = \|\tau^{T+1} - \tau_*^{T+1}\|^2$ between predicted and expert trajectories. These scores are aggregated via conformal prediction [59] to derive an uncertainty metric \hat{C} . This quantile-based metric \hat{C} directly serves as the uncertainty loss $\mathcal{L}_{\text{unct}}$. We can form the uncertainty loss:

$$\mathcal{L}_{\text{unct}} = \text{Quantile}(\{r_i^{T+1}\}_{i=1}^k; [(k+1)(1-\alpha)]/k), \quad (5)$$

where k is the number of training samples, and α is the allowed failure probability. We use the differentiable ranking and sorting techniques [7] for the soft quantile function.

3.3. Gaussian-based World Model

For AD systems, a critical aspect is accurately predicting the future states of both the ego vehicle and its surrounding environment. This section outlines how our model predicts future scenes and details the training losses employed. GWM w takes sensor inputs (including a set of multi-view images and semantic maps \mathbf{I} and a set of LiDAR points \mathbf{P}) collected from previous frames and infers the scene and trajectory for the next frames. Specifically, the ego trajectory at time $T+1$, denoted as τ^{T+1} , is predicted alongside the surrounding scene \mathbf{o}^{T+1} . GWM captures the joint distribution of the ego vehicle's movement and the surrounding scene's evolution, enabling the prediction of future states. Formally, the function w is given by:

$$\mathbf{o}^{T+1}, \tau^{T+1} = w((\mathbf{I}^T, \dots, \mathbf{I}^{T-t}), (\mathbf{P}^T, \dots, \mathbf{P}^{T-t})). \quad (6)$$

After obtaining the predicted scene \mathbf{o}^{T+1} and ego trajectory τ^{T+1} , these are fed back into the model to recursively predict subsequent frames in an auto-regressive manner. During training, ground-truth data is used as input for future predictions, while inferred results are used during inference. This world model effectively captures high-order interactions between the ego vehicle and its environment, leading to more accurate predictions of dynamic changes in the scene.

Training Losses. Our GWM is trained in two stages. In the first stage, we train the 2D/3D encoder and the *Gaussian*

¹The commands include go forward, turn left, and turn right, representing the vehicle's highest-level intentions.

representation learner ϕ to learn the Gaussian representation using the following loss function:

$$\mathcal{J}_\phi = \lambda_1 \mathcal{L}_{\text{sem}} + \lambda_2 \mathcal{L}_{\text{recon}}, \quad (7)$$

where \mathcal{L}_{sem} is the semantic loss, ensuring accurate semantic classification; $\mathcal{L}_{\text{recon}}$ is the reconstruction loss, enforcing consistency between projected images and raw images; λ_1 and λ_2 are coefficients. \mathcal{L}_{sem} measures the discrepancy between projected semantic images and raw images; $\mathcal{L}_{\text{recon}}$ enforces consistency between projected images from the Gaussian representation and raw input images (see §4.4 for details).

In the second stage, we train the *Simulator* ψ and the *Planner* θ to predict future occupancy states and plan trajectories. The total loss function is:

$$\mathcal{J}_{\psi, \theta} = \mathcal{L}_{\text{fcst}} + \mathcal{L}_{\text{plan}}, \quad (8)$$

where $\mathcal{L}_{\text{fcst}}$ is the forecasting loss, guiding ψ in learning how the world evolves; $\mathcal{L}_{\text{plan}}$ is the planning loss, encouraging the *planner* to generate safe and efficient trajectories. It includes an uncertainty loss, a max-margin loss that penalizes low-cost trajectories that deviate from expert demonstrations, an imitation learning loss that ensures the predicted trajectory closely follows the expert trajectory, and a collision loss that penalizes trajectories that collide with obstacles. By combining these losses, GWM effectively learns to represent the environment using Gaussian-based semantic and geometric features, anticipate future dynamics and plan trajectories for safe and precise decision-making.

4. Experiment

In this section, we evaluate our GWM on two tasks: forecasting on the Occ3D [57] dataset (in §4.1) and planning on the nuScenes [2] dataset (in §4.2). Furthermore, we conduct occlusions and noise analysis, and ablation experiments in §4.3 and §4.4 to verify GWM's effectiveness.

Implementation Details. Following [24, 28, 87], we use a 2-second historical context to predict the next 3 seconds. During training, we apply masking to temporal attention to prevent information leakage from future frames, ensuring the model learns to forecast based only on past data. Ground-truth data serves as input for future predictions during training, while an autoregressive prediction strategy is employed for inference, using inferred results as input. The model is trained with the AdamW optimizer [44] alongside a cosine annealing scheduler [45], starting with an initial learning rate of 1×10^{-3} and a weight decay of 0.01.

4.1. 4D Occupancy Forecasting

Task Description. We delve into 4D occupancy forecasting, which predicts future 3D occupancy scenes given a few historical scene inputs. Specifically, we follow existing works [67, 87] and use a 2-second historical frames to forecast the subsequent 3 seconds. We evaluate performance

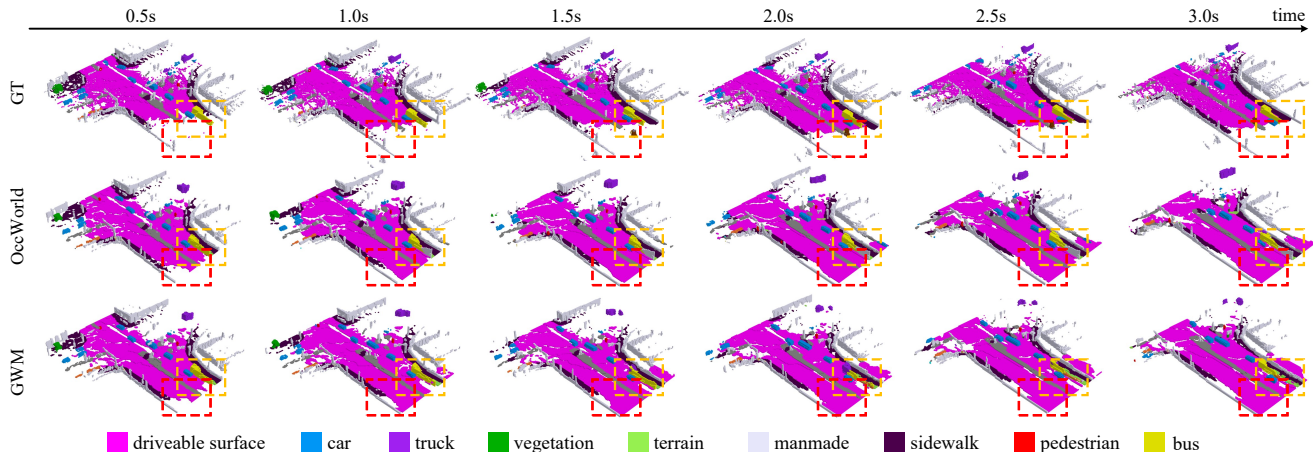


Figure 3. **Qualitative results of 4D occupancy forecasting on the Occ3D [57] dataset (§4.1).** The differences are highlighted with red and yellow boxes. (Best viewed with zoom-in.)

Table 1. **Quantitative results of 4D occupancy forecasting on the Occ3D [57] dataset (§4.1).** Aux. Sup. denotes auxiliary supervision apart from the ego trajectory. Recon. refers to reconstruction performance in 0s. Avg. denotes the average performance of that in 1s, 2s, and 3s. The best result is **bolded**. OccWorld-D uses TPVFormer predictions trained with dense ground-truth 3D occupancy; OccWorld-T uses TPVFormer predictions trained with LiDAR [87].

Method	Input	Aux. Sup.	mIoU(%) \uparrow					IoU(%) \uparrow				
			Recon.	1s	2s	3s	Avg.	Recon.	1s	2s	3s	Avg.
RenderWorld [72]	Camera	Occupancy	-	2.83	2.55	2.37	2.58	-	14.61	13.61	12.98	13.73
OccWorld-T [87]	Camera	LiDAR	7.21	4.68	3.36	2.63	3.56	10.66	9.32	8.23	7.47	8.34
OccWorld-D [87]	Camera	Occupancy	18.63	11.55	8.10	6.22	8.62	22.88	18.90	16.26	14.43	16.53
OccWorld-F [67]	Camera	Occupancy	20.09	8.03	6.91	3.54	6.16	35.61	23.62	18.13	15.22	18.99
OccLLaMA-F [67]	Camera & Action & Language	Occupancy	37.38	10.34	8.66	6.98	8.66	38.92	25.81	23.19	19.97	22.99
GWM	Camera & LiDAR	Occupancy	39.44	11.63	10.07	8.17	10.12	40.22	26.22	24.97	22.13	24.60
OccWorld-O [67]	Occupancy	None	66.38	25.78	15.14	10.51	17.14	62.29	34.63	25.07	20.18	26.63
OccLLaMA-O [67]	Occ & Action & Language	None	75.20	25.05	19.49	15.26	19.93	63.76	34.56	28.53	24.41	29.17
RenderWorld [72]	Occupancy	None	-	28.69	18.89	14.83	20.80	-	37.74	28.41	24.08	30.08

with Mean Intersection over Union (mIoU) and Intersection over Union (IoU) across 3-second frames.

Quantitative Results. In the 4D occupancy forecasting task, we compare GWM with two SOTA methods: OccWorld and OccLLaMA. These two methods were evaluated under two different settings [67]: using ground-truth 3D occupancy maps (-O) and predicted 3D occupancy maps based on camera data (-F, using the FBOCC method [42]). The quantitative results in Tab. 1 highlight that GWM outperforms OccWorld-F [67] and OccLLaMA-F [67] across most future time steps, demonstrating a clear advantage in both mIoU and IoU metrics. Gaussian representations effectively manage uncertainties in sensor data, leading to more reliable predictions of dynamic object movements and environmental changes, particularly in complex scenarios where sensor noise and occlusions are common. Compared to OccWorld-F, GWM attains better gains over OccLLaMA-F (+3.96% vs +2.5% mIoU, +5.61% vs +4.0% IoU). OccWorld neglects multimodal information; OccLLaMA-F and GWM both attempt to extend to multimodal inputs. However, language priors can directly guide decision-making. Our improved

gains demonstrate the effectiveness of our design, even without language input. Moreover, GWM demonstrates better long-term prediction capabilities in comparison to OccWorld and OccLLaMA. Our performance is better at the 2-second and 3-second time steps, underscoring the model’s ability to better forecast over longer horizons. Our method is a *sensor-to-forecasting* pipeline that directly takes camera data as input. By integrating Gaussian representations into the model, GWM is able to handle the uncertainties inherent in sensor measurements, particularly in complex and dynamic environments. This allows our model to make more accurate and robust long-term predictions.

Qualitative Results. Fig. 3 compares the predicted 3D occupancy across future time steps for GWM and OccWorld. GWM not only excels at future occupancy prediction but also delivers more accurate and complete scene reconstructions, even under challenging conditions with occlusions and sensor noise. Furthermore, GWM more effectively forecasts the trajectories of moving objects (e.g., cars) and reconstructs previously unseen drivable areas with greater fidelity (see supplementary for more results).

Table 2. **Quantitative results of motion planning on the nuScenes [2] dataset (§4.2)**. Avg. denotes the average performance of that in 1s, 2s, and 3s. We use bold and underlined numbers to denote the highest-ranking and succeeding best results, respectively.

Method	Input	Auxiliary Supervision	L2(m)↓				Collision(%)↓			
			1s	2s	3s	Avg.	1s	2s	3s	Avg.
OccWorld-D [87]	Camera	Occupancy	0.52	1.27	2.41	1.40	0.12	0.40	2.08	0.87
RenderWorld [72]	Camera	Occupancy	0.48	1.30	2.67	1.48	0.14	0.55	2.23	0.97
OccWorld-F [67]	Camera	Occupancy	0.45	1.33	2.25	1.34	0.08	0.42	1.71	0.73
OccLLaMA-F [67]	Camera & Action & Language	Occupancy	<u>0.38</u>	<u>1.07</u>	<u>2.15</u>	<u>1.20</u>	0.06	0.39	1.65	0.70
GWM	Camera & LiDAR	Occupancy	0.34	1.01	2.05	1.13	<u>0.07</u>	0.26	1.45	0.59
OccWorld-O [87]	Occupancy	None	0.43	1.08	1.99	1.17	0.07	0.38	1.35	0.60
RenderWorld [72]	Occupancy	None	0.35	0.91	1.84	1.03	0.05	0.40	1.39	0.61
OccLLaMA-O [67]	Occ & Action & Language	None	0.37	1.02	2.03	1.14	0.04	0.24	1.20	0.49

Discussion. OccWorld and OccLLaMA suffer from: input dependence on provided occupancy (*-F), limited long-term prediction accuracy, and insufficient analysis under complex scenarios. In contrast, 1) we concentrate on world models within the 3D occupancy space with semantics. GWM seamlessly integrates raw multi-modal sensor inputs with the *sensor-to-forecasting* pipeline; 2) we achieve robust long-term predictions via Gaussian representations and uncertainty modeling; 3) we conduct a detailed analysis of complex scenarios to validate the effectiveness of GWM.

4.2. Motion Planning

Task Description. In motion planning, the goal is to generate safe and reliable trajectories for autonomous vehicles. We use the following key evaluation metrics: 1) L2 Error – this metric quantifies how closely the predicted trajectory aligns with the ground-truth trajectory by calculating their L2 distance; 2) Collision Rate – this evaluates the safety of the planned path by calculating the rate of collisions between the ego vehicle and surrounding obstacles.

Quantitative Results. In Tab. 2, GWM outperforms OccWorld and OccLLaMA in both trajectory accuracy and safety during motion planning. By leveraging the Gaussian Representation, GWM more effectively anticipates future environmental states, leading to precise and collision-free trajectories. Similar to the discussion under 4D occupancy forecasting task, when compared with OccWorld-F, GWM attains better gains over OccLLaMA-F (Avg. L2 error↓: -**0.21m** vs -0.14m, Avg. Collision Rate↓: -**0.14%** vs -0.03%). Furthermore, GWM maintains superior performance across all time frames. In terms of safety, GWM delivers a lower collision rate and better long-term planning capabilities than both OccWorld and OccLLaMA. At 1 second, it outperforms OccWorld and matches OccLLaMA’s performance. Over extended periods, GWM’s advantage becomes more pronounced, maintaining a better average collision rate (**0.59%** vs 0.73% and 0.70%). Notably, GWM achieves performance comparable to OccWorld-O, RenderWorld, and OccLLaMA-o (which require ground-truth 3D occupancy inputs) on both Avg. L2 errors and Avg. collision rates.

These results highlight GWM’s robustness and effec-

Table 3. **Quantitative results** on challenging urban environments with sensor errors and occlusions (§4.3).

Method	Forecasting		Planning	
	mIoU(%)↑	IoU(%)↑	L2 (m)↓	Collision (%)↓
OccWorld-F [67]	5.72	15.23	1.42	0.85
GWM (Ours)	9.14	21.37	1.18	0.66

tiveness, solidifying its superiority over OccWorld and OccLLaMA. The superiority of GWM can be attributed to its effective use of the Gaussian representation, which strengthens the model’s capability to interpret intricate spatiotemporal dynamics inherent in dynamic driving environments. Unlike prior voxel-based methods (*e.g.*, OccWorld and OccLLaMA), the Gaussian representation provides a probabilistic foundation that more accurately models uncertainties and variabilities within the scene. This allows GWM to achieve *sensor-to-planning* pipeline and generate more precise and collision-free trajectories, even in complex and rapidly changing scenarios. Moreover, although OccLLaMA employs multimodal inputs (including Action), it suffers from information loss regarding action vocabulary, one of the reasons that it is not outperforming GWM.

4.3. Occlusions and Noise

To evaluate GWM in a challenging environment, we curate a challenging subset of the nuScenes [2] dataset.

Dataset Preparation. Noise from sensor errors and registration inaccuracies affects geometric consistency and reconstruction, while occlusion leads to missing data, depth estimation errors, and artifacts in reconstruction. Factors such as data distribution bias and dynamic interactions further amplify these uncertainties, making scene reconstruction and motion prediction more challenging. We select sequences from nuScenes [2] that exhibit significant occlusions and sensor noise. Additionally, to simulate sensor errors and additional occlusions, we inject LiDAR jitter as displacement noise into the LiDAR data [50] and apply random occlusions to the camera images. We then evaluate these processed scenes using uncertainty-aware 3DGS [33] and assess uncertainty quality via AUSE and NLL. Scenes with high uncertainty – resulting from sensor noise degrading structural

Table 4. **Comparison** of models on *1st stage* pre-training (§4.4).

Method	Gaussian Quality			Task Performance	
	$L_{\text{recon}}\downarrow$	PSNR \uparrow	SSIM \uparrow	mIoU($\%$) \uparrow	L2(m) \downarrow
GWM (<i>w/o 1st stage</i>)	0.213	24.58	0.723	8.25	1.28
GWM (<i>w/ 1st stage</i>)	0.146	27.62	0.791	10.12	1.13

Table 5. **Ablation study** on Gaussian representation, semantic map and LiDAR data (§4.4).

Method	Forecasting		Planning	
	mIoU($\%$) \uparrow	IoU($\%$) \uparrow	L2 (m) \downarrow	Collision ($\%$) \downarrow
OccWorld-F [67]	6.16	18.99	1.34	0.73
GWM (<i>w/ FBOCC</i>)	8.97	17.84	1.27	0.69
GWM (<i>w/ SfM priors</i>)	8.61	22.56	1.21	0.70
GWM (<i>w/o semantic map</i>)	9.84	23.37	1.14	0.61
GWM (Full Model)	10.12	24.60	1.13	0.59

modeling and occlusions causing data loss – are defined as challenging environments. This curated subset provides a comprehensive benchmark for evaluating the model’s performance under complex urban driving conditions.

Quantitative Results. As shown in Tab. 3, GWM outperforms OccWorld-F in both forecasting and planning on this challenging subset. GWM achieves a higher mIoU of **9.14%**, and reduces the average L2 error to **1.18m**. These improvements highlight GWM’s robustness in handling sensor errors and occlusions. Specifically, GWM’s use of Gaussian representations allows it to better handle uncertainties due to sensor noise, and the spatiotemporal Transformer effectively captures temporal dependencies, improving occupancy predictions in occluded regions. These findings underscore the practical applicability of GWM in real-world scenarios, where sensor errors and occlusions are commonplace.

4.4. Ablation Study

To evaluate our key designs, we perform comprehensive ablation studies on the Occ3D [57] and nuScenes [2] datasets. **1st Stage.** We investigate the role of the *1st stage* pre-training in learning a robust Gaussian Representation. We compare two variants: GWM (*w/ 1st stage*) and GWM (*w/o 1st stage*), and report: Gaussian quality (L_{recon} , PSNR and SSIM) and task performance (Avg. mIoU and Avg. L2 error). Gaussian quality is operated on learner ϕ of *2nd stage*. Tab. 4 summarizes the performance comparison. GWM (*w/ 1st stage*) shows a lower reconstruction loss and better reconstruction quality, indicating that the pre-training helps extract more robust Gaussian Representations. Consequently, the downstream forecasting and planning benefit from this improved representation. The results indicate that the *1st stage*, acting as a self-supervised pre-training step (similar to ViDAR [76]), significantly improves the quality of the Gaussian Representation. With a better Gaussian quality, GWM can better capture both spatial details and semantic context. The superior performance on the downstream tasks also confirms the effectiveness of the *1st stage*.

Gaussian Representation. As shown in Tab. 5, full GWM model outperforms the baseline (GWM *w/ FBOCC*, which uses a voxel-based representation), in both forecasting and planning (**10.12%** vs 8.97%, **1.13m** vs 1.27m). These gains underscore how a probabilistic Gaussian framework more effectively captures uncertainties and variabilities in the scene, thereby enhancing both occupancy prediction and motion planning. Moreover, Gaussian representation endows GWM with a robust ability to generate collision-free trajectories, even in rapidly changing scenarios, reaffirming the merit of modeling uncertainty through a Gaussian prior.

LiDAR Data. When replacing LiDAR inputs with SfM priors (Tab. 5), performance degrades across all metrics: the mIoU metric drops by **1.51%** (10.12%→8.61%) while L2 error increases by **0.08m** (1.13m→1.21m). This gap stems from LiDAR’s unique capability to preserve structural fidelity – its millimeter-level depth precision and dense spatial sampling enable robust 3D scene understanding that camera-derived priors fundamentally cannot match. SfM initialization proves constrained by its sparser points and cumulative structural errors from sequential image matching.

Semantic Map. In Tab. 5, the majority of performance gains stem from the introduction and design of the Gaussian representation. Semantic Map contributes only minor improvements, serving as an auxiliary component.

Additional Analyses. For ablation analyses of *simulator* and *planner*, please refer to §C.1 and §C.2 in the supplementary.

5. Conclusion

In conclusion, our study introduces GWM, a new method that integrates Gaussian priors with voxel-based world models to advance 3D reconstruction and motion prediction in autonomous driving. Utilizing LiDAR data and Gaussian splatting techniques, GWM enhances scene fidelity, dynamically modeling uncertainty and capturing elements. By incorporating a spatiotemporal model and an *uncertainty-aware planner*, our GWM accurately predicts future scene evolutions and vehicle trajectories. Tested extensively on the nuScenes dataset, GWM outperforms the baselines in both semantic understanding and trajectory forecasting. This method not only improves the precision and reliability of autonomous systems but also sets a new benchmark for world models in autonomous driving by providing more robust and accurate modeling of dynamic environments.

Acknowledgments: This work was supported by the National Science and Technology Major Project (No. 2023ZD0121300), National Natural Science Foundation of China (No. 62372405), Fundamental Research Funds for the Central Universities (226-2025-00057), CIE-Tencent Robotics X Rhino-Bird Focused Research Program, and the National Key Laboratory of Human-Machine Hybrid Augmented Intelligence, Xi’an Jiaotong University (No. HMHAI-202403).

References

- [1] Daniel Bogdoll, Yitian Yang, and J Marius Zöllner. Muvo: A multimodal generative world model for autonomous driving with geometric representations. *arXiv preprint arXiv:2311.11762*, 2023. 2
- [2] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020. 2, 5, 7, 8
- [3] Anh-Quan Cao and Raoul De Charette. Monoscene: Monocular 3d semantic scene completion. In *CVPR*, 2022. 1
- [4] Anh-Quan Cao and Raoul de Charette. Scenerf: Self-supervised monocular 3d scene reconstruction with radiance fields. In *ICCV*, 2023. 1
- [5] Sergio Casas, Abbas Sadat, and Raquel Urtasun. Mp3: A unified model to map, perceive, predict and plan. In *CVPR*, 2021. 4, 5
- [6] Guikun Chen and Wenguan Wang. A survey on 3d gaussian splatting. *arXiv preprint arXiv:2401.03890*, 2024. 3
- [7] Marco Cuturi, Olivier Teboul, and Jean-Philippe Vert. Differentiable ranking and sorting using optimal transport. In *NeurIPS*, 2019. 5
- [8] Simon Doll, Niklas Hanselmann, Lukas Schneider, Richard Schulz, Marius Cordts, Markus Enzweiler, and Hendrik Lensch. Dualad: Disentangling the dynamic and static world for end-to-end driving. In *CVPR*, 2024. 2
- [9] Tuo Feng, Wenguan Wang, Xiaohan Wang, Yi Yang, and Qinghua Zheng. Clustering based point cloud representation learning for 3d analysis. In *ICCV*, pages 8283–8294, 2023. 2
- [10] Tuo Feng, Wenguan Wang, Ruijie Quan, and Yi Yang. Shape2scene: 3d scene representation learning through pre-training on shape data. In *ECCV*, pages 73–91. Springer, 2024. 2
- [11] Tuo Feng, Wenguan Wang, and Yi Yang. A survey of world models for autonomous driving. *arXiv preprint arXiv:2501.11260*, 2025. 2
- [12] Wanshui Gan, Ningkai Mo, Hongbin Xu, and Naoto Yokoya. A simple attempt for 3d occupancy estimation in autonomous driving. *arXiv preprint arXiv:2303.10076*, 2023. 1
- [13] Chen Gao, Ayush Saraf, Johannes Kopf, and Jia-Bin Huang. Dynamic view synthesis from dynamic monocular video. In *ICCV*, 2021. 1
- [14] Jianzhe Gao, Rui Liu, and Wenguan Wang. 3d gaussian map with open-set semantic grouping for vision-language navigation. In *ICCV*, 2025. 3
- [15] Ruiyuan Gao, Kai Chen, Enze Xie, Lanqing Hong, Zhenguo Li, Dit-Yan Yeung, and Qiang Xu. Magicdrive: Street view generation with diverse 3d geometry control. *arXiv preprint arXiv:2310.02601*, 2023. 2
- [16] Clément Godard, Oisín Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, 2017. 1
- [17] Junru Gu, Chenxu Hu, Tianyuan Zhang, Xuanyao Chen, Yilun Wang, Yue Wang, and Hang Zhao. Vip3d: End-to-end visual trajectory prediction via 3d agent queries. In *CVPR*, 2023. 2
- [18] Vitor Guizilini, Igor Vasiljevic, Rares Ambrus, Greg Shakhnarovich, and Adrien Gaidon. Full surround monodepth from multiple cameras. *IEEE Robotics and Automation Letters*, 7(2):5397–5404, 2022. 2
- [19] Zhiming Guo, Xing Gao, Jianlan Zhou, Xinyu Cai, and Botian Shi. Scenedm: Scene-level multi-agent trajectory generation with consistent diffusion models. *arXiv preprint arXiv:2311.15736*, 2023. 2
- [20] David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018. 2
- [21] Anthony Hu, Zak Murez, Nikhil Mohan, Sofia Dudas, Jeffrey Hawke, Vijay Badrinarayanan, Roberto Cipolla, and Alex Kendall. Fiery: Future instance prediction in bird’s-eye view from surround monocular cameras. In *ICCV*, 2021. 2
- [22] Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023. 2
- [23] 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. 1, 2, 4, 5
- [24] Yihan Hu, Jiazhi 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*, 2023. 2, 5
- [25] Junjie Huang, Guan Huang, Zheng Zhu, Yun Ye, and Dalong Du. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*, 2021. 1, 2
- [26] Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Tri-perspective view for vision-based 3d semantic occupancy prediction. In *CVPR*, 2023. 2
- [27] 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. 3
- [28] 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. In *ICCV*, 2023. 2, 5
- [29] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM TOG*, 42(4):1–14, 2023. 1, 2, 3
- [30] Tarasha Khurana, Peiyun Hu, Achal Dave, Jason Ziglar, David Held, and Deva Ramanan. Differentiable raycasting for self-supervised occupancy forecasting. In *ECCV*, 2022. 2
- [31] Tarasha Khurana, Peiyun Hu, David Held, and Deva Ramanan. Point cloud forecasting as a proxy for 4d occupancy forecasting. In *CVPR*, 2023. 2
- [32] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, 2023. 3, 4
- [33] Ruiqi Li and Yiu-ming Cheung. Variational multi-scale representation for estimating uncertainty in 3d gaussian splatting. In *NeurIPS*, pages 87934–87958, 2025. 7
- [34] Xin Li, Tao Ma, Yuenan Hou, Botian Shi, Yuchen Yang, Youquan Liu, Xingjiao Wu, Qin Chen, Yikang Li, Yu Qiao,

- et al. Logonet: Towards accurate 3d object detection with local-to-global cross-modal fusion. In *CVPR*, 2023. 2
- [35] 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. 2
- [36] Yinhao Li, Han Bao, Zheng Ge, Jinrong Yang, Jianjian Sun, and Zeming Li. Bevstereo: Enhancing depth estimation in multi-view 3d object detection with temporal stereo. In *AAAI*, 2023. 1
- [37] 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. In *AAAI*, 2023. 2
- [38] Yiming Li, Zhiding Yu, Christopher Choy, Chaowei Xiao, Jose M Alvarez, Sanja Fidler, Chen Feng, and Anima Anandkumar. Voxformer: Sparse voxel transformer for camera-based 3d semantic scene completion. In *CVPR*, 2023. 1
- [39] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. In *CVPR*, 2021. 1
- [40] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird’s-eye-view representation from multi-camera images via spatiotemporal transformers. In *ECCV*, 2022. 1, 2
- [41] Zhiqi Li, Zhiding Yu, David Austin, Mingsheng Fang, Shiyi Lan, Jan Kautz, and Jose M Alvarez. Fb-occ: 3d occupancy prediction based on forward-backward view transformation. *arXiv preprint arXiv:2307.01492*, 2023. 1, 2
- [42] Zhiqi Li, Zhiding Yu, Wenhai Wang, Anima Anandkumar, Tong Lu, and Jose M Alvarez. Fb-bev: Bev representation from forward-backward view transformations. In *ICCV*, 2023. 6
- [43] Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huiyi Mao, Daniela L Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird’s-eye view representation. In *ICRA*, 2023. 1
- [44] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5
- [45] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 5
- [46] Benedikt Mersch, Xieyuanli Chen, Jens Behley, and Cyrill Stachniss. Self-supervised point cloud prediction using 3d spatio-temporal convolutional networks. In *CoRL*, 2022. 2
- [47] Ruihang Miao, Weizhou Liu, Mingrui Chen, Zheng Gong, Weixin Xu, Chen Hu, and Shuchang Zhou. Occdepth: A depth-aware method for 3d semantic scene completion. *arXiv preprint arXiv:2302.13540*, 2023. 1
- [48] Lang Peng, Zhirong Chen, Zhangjie Fu, Pengpeng Liang, and Erkang Cheng. Bevsegformer: Bird’s eye view semantic segmentation from arbitrary camera rigs. In *WACV*, 2023. 2
- [49] Philip Polack, Florent Alché, Brigitte d’Andréa Novel, and Arnaud de La Fortelle. The kinematic bicycle model: A consistent model for planning feasible trajectories for autonomous vehicles? In *IV*, 2017. 5
- [50] Guocheng Qian, Yuchen Li, Houwen Peng, Jinjie Mai, Hasan Hammoud, Mohamed Elhoseiny, and Bernard Ghanem. Pointnext: Revisiting pointnet++ with improved training and scaling strategies. In *NeurIPS*, pages 23192–23204, 2022. 7
- [51] Luis Roldao, Raoul de Charette, and Anne Verroust-Blondet. Lmscnet: Lightweight multiscale 3d semantic completion. In *3DV*, 2020. 1
- [52] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2
- [53] Abbas Sadat, Sergio Casas, Mengye Ren, Xinyu Wu, Pranaab Dhawan, and Raquel Urtasun. Perceive, predict, and plan: Safe motion planning through interpretable semantic representations. In *ECCV*, 2020. 2, 4
- [54] Aron Schmed, Tobias Fischer, Martin Danelljan, Marc Pollefeys, and Fisher Yu. R3d3: Dense 3d reconstruction of dynamic scenes from multiple cameras. In *ICCV*, 2023. 1
- [55] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 3
- [56] Wenchao Sun, Xuewu Lin, Yining Shi, Chuang Zhang, Hao-ran Wu, and Sifa Zheng. Sparsedrive: End-to-end autonomous driving via sparse scene representation. *arXiv preprint arXiv:2405.19620*, 2024. 2
- [57] Xiaoyu Tian, Tao Jiang, Longfei Yun, Yucheng Mao, Huitong Yang, Yue Wang, Yilun Wang, and Hang Zhao. Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving. In *NeurIPS*, 2024. 1, 2, 5, 6, 8
- [58] 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*, 2023. 1
- [59] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer, 2005. 5
- [60] Hanqing Wang, Wei Liang, Luc Van Gool, and Wenguan Wang. Dreamwalker: Mental planning for continuous vision-language navigation. In *ICCV*, pages 10873–10883, 2023. 2
- [61] 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. 2
- [62] Shihao Wang, Yingfei Liu, Tiancai Wang, Ying Li, and Xiangyu Zhang. Exploring object-centric temporal modeling for efficient multi-view 3d object detection. In *ICCV*, 2023. 2
- [63] Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, and Jiwen Lu. Drivedreamer: Towards real-world-driven world models for autonomous driving. *arXiv preprint arXiv:2309.09777*, 2023. 2
- [64] Xiaofeng Wang, Zheng Zhu, Wenbo Xu, Yunpeng Zhang, Yi Wei, Xu Chi, Yun Ye, Dalong Du, Jiwen Lu, and Xingang Wang. Openoccupancy: A large scale benchmark for surrounding semantic occupancy perception. In *ICCV*, 2023. 1, 2
- [65] Yuqi Wang, Yuntao Chen, Xingyu Liao, Lue Fan, and Zhaoxiang Zhang. Panoocc: Unified occupancy representation for camera-based 3d panoptic segmentation. In *CVPR*, 2024. 1
- [66] 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. 2
- [67] Julong Wei, Shanshuai Yuan, Pengfei Li, Qingda Hu, Zhongxue Gan, and Wenchao Ding. Occllama: An occupancy-language-action generative world model for autonomous driving. *arXiv preprint arXiv:2409.03272*, 2024. 2, 5, 6, 7, 8
- [68] Yi Wei, Linqing Zhao, Wenzhao Zheng, Zheng Zhu, Jie Zhou, and Jiwen Lu. Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving. In *ICCV*, 2023. 1
- [69] 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*, 2021. 2
- [70] Xinshuo Weng, Boris Ivanovic, Yan Wang, Yue Wang, and Marco Pavone. Para-drive: Parallelized architecture for real-time autonomous driving. In *CVPR*, 2024. 2
- [71] Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. In *CVPR*, 2024. 2
- [72] Ziyang Yan, Wenzhen Dong, Yihua Shao, Yuhang Lu, Liu Haiyang, Jingwen Liu, Haozhe Wang, Zhe Wang, Yan Wang, Fabio Remondino, et al. Renderworld: World model with self-supervised 3d label. *arXiv preprint arXiv:2409.11356*, 2024. 2, 6, 7
- [73] 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. 2
- [74] Xiangpeng Yang, Linchao Zhu, Hehe Fan, and Yi Yang. Videograin: Modulating space-time attention for multi-grained video editing. In *ICLR*, 2025. 2
- [75] 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. In *AAAI*, pages 9327–9335, 2025. 5
- [76] Zetong Yang, Li Chen, Yanan Sun, and Hongyang Li. Visual point cloud forecasting enables scalable autonomous driving. In *CVPR*, pages 14673–14684, 2024. 8
- [77] Tengju Ye, Wei Jing, Chunyong Hu, Shikun Huang, Lingping 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
- [78] Junbo Yin, Jianbing Shen, Runnan Chen, Wei Li, Ruigang Yang, Pascal Frossard, and Wenguan Wang. Is-fusion: Instance-scene collaborative fusion for multimodal 3d object detection. In *CVPR*, pages 14905–14915, 2024. 2
- [79] Zichen Yu, Changyong Shu, Jiajun Deng, Kangjie Lu, Zongdai Liu, Jiangyong Yu, Dawei Yang, Hui Li, and Yan Chen. Flashocc: Fast and memory-efficient occupancy prediction via channel-to-height plugin. *arXiv preprint arXiv:2311.12058*, 2023. 1
- [80] 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. 4
- [81] Chubin Zhang, Juncheng Yan, Yi Wei, Jiabin Li, Li Liu, Yansong Tang, Yueqi Duan, and Jiwen Lu. Occnerf: Self-supervised multi-camera occupancy prediction with neural radiance fields. *arXiv preprint arXiv:2312.09243*, 2023. 1
- [82] 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. 2
- [83] Lunjun Zhang, Yuwen Xiong, Ze Yang, Sergio Casas, Rui Hu, and Raquel Urtasun. Learning unsupervised world models for autonomous driving via discrete diffusion. *arXiv preprint arXiv:2311.01017*, 2023. 2
- [84] Yunpeng Zhang, Zheng Zhu, Wenzhao Zheng, Junjie Huang, Guan Huang, Jie Zhou, and Jiwen Lu. Beverse: Unified perception and prediction in birds-eye-view for vision-centric autonomous driving. *arXiv preprint arXiv:2205.09743*, 2022. 1, 2
- [85] Yunpeng Zhang, Zheng Zhu, and Dalong Du. Occformer: Dual-path transformer for vision-based 3d semantic occupancy prediction. In *ICCV*, 2023. 1
- [86] Yumeng Zhang, Shi Gong, Kaixin Xiong, Xiaoqing Ye, Xiao Tan, Fan Wang, Jizhou Huang, Hua Wu, and Haifeng Wang. Bevworld: A multimodal world model for autonomous driving via unified bev latent space. *arXiv preprint arXiv:2407.05679*, 2024. 2
- [87] Wenzhao Zheng, Weiliang Chen, Yuanhui Huang, Borui Zhang, Yueqi Duan, and Jiwen Lu. Occworld: Learning a 3d occupancy world model for autonomous driving. *arXiv preprint arXiv:2311.16038*, 2023. 1, 2, 4, 5, 6, 7
- [88] Xiaoyu Zhou, Zhiwei Lin, Xiaojun Shan, Yongtao Wang, Deqing Sun, and Ming-Hsuan Yang. Drivinggaussian: Composite gaussian splatting for surrounding dynamic autonomous driving scenes. In *CVPR*, 2024. 2, 3
- [89] 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. 1
- [90] Sicheng Zuo, Wenzhao Zheng, Yuanhui Huang, Jie Zhou, and Jiwen Lu. Gaussianworld: Gaussian world model for streaming 3d occupancy prediction. *arXiv preprint arXiv:2412.10373*, 2024. 2, 3