IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

OccTENS: 3D Occupancy World Model via Temporal Next-Scale Prediction

Bu Jin*, Songen Gu*, Xiaotao Hu*, Yupeng Zheng, Xiaoyang Guo, Qian Zhang, Xiaoxiao Long, Wei Yin

***Abstract*—In this paper, we propose OccTENS, a generative occupancy world model that enables controllable, high-fidelity long-term occupancy generation while maintaining computational efficiency. Different from visual generation, the occupancy world model must capture the fine-grained 3D geometry and dynamic evolution of the 3D scenes, posing great challenges for the generative models. Recent approaches based on autoregression (AR) have demonstrated the potential to predict vehicle movement and future occupancy scenes simultaneously from historical observations, but they typically suffer from inefficiency, temporal degradation in long-term generation and lack of controllability. To holistically address these issues, we reformulate the occupancy world model as a temporal next-scale prediction (TENS) task, which decomposes the temporal sequence modeling problem into the modeling of spatial scale-by-scale generation and temporal scene-by-scene prediction. With a TensFormer, OccTENS can effectively manage the temporal causality and spatial relationships of occupancy sequences in a flexible and scalable way. To enhance the pose controllability, we further propose a holistic pose aggregation strategy, which features a unified sequence modeling for occupancy and ego-motion. Experiments show that OccTENS outperforms the state-of-the-art method with both higher occupancy quality and faster inference time.**

***Index Terms*—Occupancy Generation, World Model, Autonomous Driving**

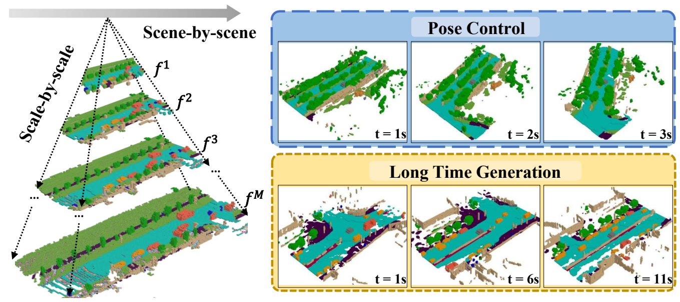

Fig. 1: We propose OccTENS, a coarse-to-fine occupancy world model that enables controllable, high-fidelity long-time occupancy generation while maintaining computational efficiency.

I. Introduction

RECENT years have seen significant advancements in the development of autonomous driving (AD) systems. While existing AD methods [1]–[5] have demonstrated excellent results across a range of driving scenarios, there are still challenges when dealing with long-tail distributions or out-of-distribution situations. A promising direction to address these challenges is the world model [6]–[12], which simulates and comprehends the surrounding environment by learning a comprehensive representation of the external world.

Occupancy world model, as a specialized type of world model, has gained significant attention for its expressiveness of the 3D geometry. Several occupancy world models [13]–[20] have been developed in recent years. For example, some works [13], [14], [16] have attempted to tailor GPT-like architecture to generate occupancy in an autoregressive manner. However, when generating long-time sequences, they exhibit **temporal degradation** and **inefficiency** issues due to the increasing accumulation of tokens over time. Moreover, they lack the ability to control generation results based on a specified camera pose or trajectory, which limits the model's predictive capacity for varying viewpoints and its reasoning ability regarding the agent's location and orientation.

To address these issues, we introduce **OccTENS**, a novel autoregressive occupancy world model designed to generate controllable, high-fidelity long-term occupancy scenes while maintaining computational efficiency. Our key innovation lies in reformulating the occupancy world model as a temporal next-scale prediction (TENS) task, which decomposes the temporal sequence modeling problem to the modeling of spatial scale-by-scale generation and temporal scene-by-scene prediction. This decomposition enables OccTENS with both computational efficiency and high-fidelity long-term occupancy scene generation through a principled balance of parallelizable spatial refinement and sequential temporal reasoning. Further, we propose a multi-modal camera pose aggregation module tailored for auto-regressive models, which facilitates pose controllability and motion planning simultaneously.

While the temporal next-scale prediction in OccTENS is inspired by next-scale prediction in VAR [21], it is **non-trivial** at all from VAR to OccTENS: (1) *Lack of temporal modeling*. While the original VAR was primarily designed for a class-to-image task, we explore its extension to temporal sequences, allowing it to capture dynamic, time-varying patterns beyond purely spatial next-scale prediction. (2) *Temporal degradation*. The multi-scale design naturally requires longer token sequences, which can overload attention mechanisms and result in temporal degradation during long-term genera-

Received 10 July 2025; accepted 28 December 2025. Date of publication 19 January 2026; date of current version 10 February 2026. This article was recommended for publication by Associate Editor Y. Sun and Editor H. Moon upon evaluation of the reviewers' comments. This work was supported by Project NSFC 62501261. (Bu Jin, Songen Gu, and Xiaotao Hu contributed equally to this work.) (Corresponding author: Wei Yin.)

Bu Jin is with The Hong Kong University of Science and Technology. Songen Gu and Yupeng Zheng are with the University of Chinese Academy of Sciences. Xiaoyang Guo, Qian Zhang, and Wei Yin are with Horizon Robotics. Xiaotao Hu is with The Hong Kong University of Science and Technology. Xiaoxiao Long is with the Nanjing University.

Bu Jin*, Songen Gu*, and Xiaotao Hu* contribute equally. Wei Yin✉ is the corresponding author. Email: yvanwy@outlook.com

Digital Object Identifier 10.1109/LRA.2026.3655202

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

tion. Instead, OccTENS decouples the frame regression from scale regression, enabling a more effective modeling of both temporal causality across different timestamps and spatial relationships across various scales. (3) *Lack of multi-modal interaction*. A world model should understand the trajectory or motion of the ego vehicle to plan a trajectory or control future occupancy generation. (4) *Representation discrepancy*. 3D occupancy involves a more complex topological structure than 2D images, such as the geometric continuity or sparse nature, highlighting the superiority of OccTENS as a critical advancement for the occupancy world model.

Our contributions are summarized as follows:

- We introduce **OccTENS**, a coarse-to-fine occupancy world model that enables controllable, high-fidelity long-term occupancy generation while maintaining computational efficiency.
- We reformulate the occupancy world model as a temporal next-scale prediction (TENS) task, which decomposes the temporal sequence modeling problem to the modeling of spatial scale-by-scale generation and temporal scene-by-scene prediction. With a **TensFormer**, OccTENS can effectively manage the temporal causality and spatial relationships of occupancy sequences.
- We propose a holistic camera pose aggregation strategy tailored for auto-regressive models. With a unified sequence modeling for occupancy and camera pose, OccTENS facilitates pose controllability (controlling occupancy with a given trajectory) and motion planning (planning a trajectory for the AD vehicle) simultaneously.
- Extensive experiments demonstrate the superior performance and efficiency of OccTENS on the occupancy prediction task and motion planning task.

II. Related Work

A. Occupancy Prediction and Generation

Recently, the 3D occupancy prediction task [22]–[31] has gained increasing attention, as it enhances scene understanding with both geometric and semantic information. MonoScene [24] is the pioneer in vision-based occupancy prediction. Subsequent research has advanced occupancy prediction accuracy and reduced computational overhead by improving scene representation [25], [26], enhancing view transformation modules [27], integrating object detection [28], introducing multi-view 2D information [30], [31], and employing privileged learning [29]. Recent studies [13]–[18] have adopted occupancy as the representation of world models in autonomous driving, aiming at generating future occupancy based on historical observation. Occworld [13] is the earliest approach for the occupancy world model. OccLlama [14] and Occllm [16] enhance the generation performance by incorporating large language models. OccSora [15] and Dome [17] taming diffusion models to generate future occupancy based on trajectory input. UnO [19] and DIO [20]leverage implicit formulations to model 3D scenes as continuous occupancy fields, enabling perception and forecasting at an infinite resolution. DIO builds upon this concept by introducing a decomposable 4D world model that jointly represents occupancy and scene flow over time. Uniscene [18] further introduces a unified scene prediction module including occupancy, video, and lidar prediction, requiring the BEV layouts as input. However, the demand for the ground truth of future trajectories or BEV layouts makes it impossible for the downstream motion planning task. Moreover, although these works are well-designed architectures, they suffer from either low-fidelity or low-efficiency problems. It remains challenging to *simultaneously* achieve high-fidelity occupancy estimation and fast inference time.

B. World Model

World model [32] can predict the consequences of various actions, which is crucial for autonomous driving. Traditional models emphasize visual prediction [7], [9], [10], [33]–[40], potentially overlooking the essential 3D information needed for AD vehicles. Some approaches attempt to forecast point clouds using unannotated LiDAR scans [41], [42], but these methods neglect semantic information and are not suitable for vision-based or fusion-based autonomous driving. Occupancy world models [13]–[15], [19], [20] create a world model in 3D occupancy space, providing a more comprehensive understanding of the evolvement of 3D scenes, which is the main focus of our work.

III. Method

In this work, we propose **OccTENS**, a novel occupancy world model designed to comprehend historical observations and forecast the future 3D scenarios. As illustrated in Fig. 2, our proposed OccTENS consists of two components: a robust tokenizer that encodes 3D occupancy and ego motion into discrete tokens (see Sec. III-A), and a generative world model using next-scale prediction for future scene forecasting and motion planning (see Sec. III-B).

A. Tokenizer

The goal of the tokenizer is to model the 3D occupancy scene and the ego motion as discrete tokens.

1) Scene Tokenizer: The goal of the scene tokenizer is to model the 3D occupancy scene as discrete tokens. To achieve this, a common practice for scene tokenizer is to employ a quantized autoencoder like VQVAE [13], [14], which quantizes the occupancy feature map with discrete feature vectors. However, unlike natural language sentences with an inherent left-to-right ordering, the occupancy feature maps are interdependent, resulting in the bidirectional correlations of the quantized token sequence. This contradicts the unidirectional dependency assumption of autoregressive models, where each token can only depend on its prefix, as illustrated in [21]. Thus, we propose a multi-scale tokenizer specifically designed for next-scale prediction.

Firstly, we employ an occupancy encoder to encode the occupancy scene into a feature map. Given a scene $\mathbf{S}$, we convert it to a BEV representation with a series of 2D convolution layers, resulting in a latent feature $\mathbf{F} \in \mathbb{R}^{H\times W\times C}$. We then utilize a quantizer to tokenize the latent feature $\mathbf{F}$ into multi-scale discrete tokens. Previous approaches [13], [14] attempt

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

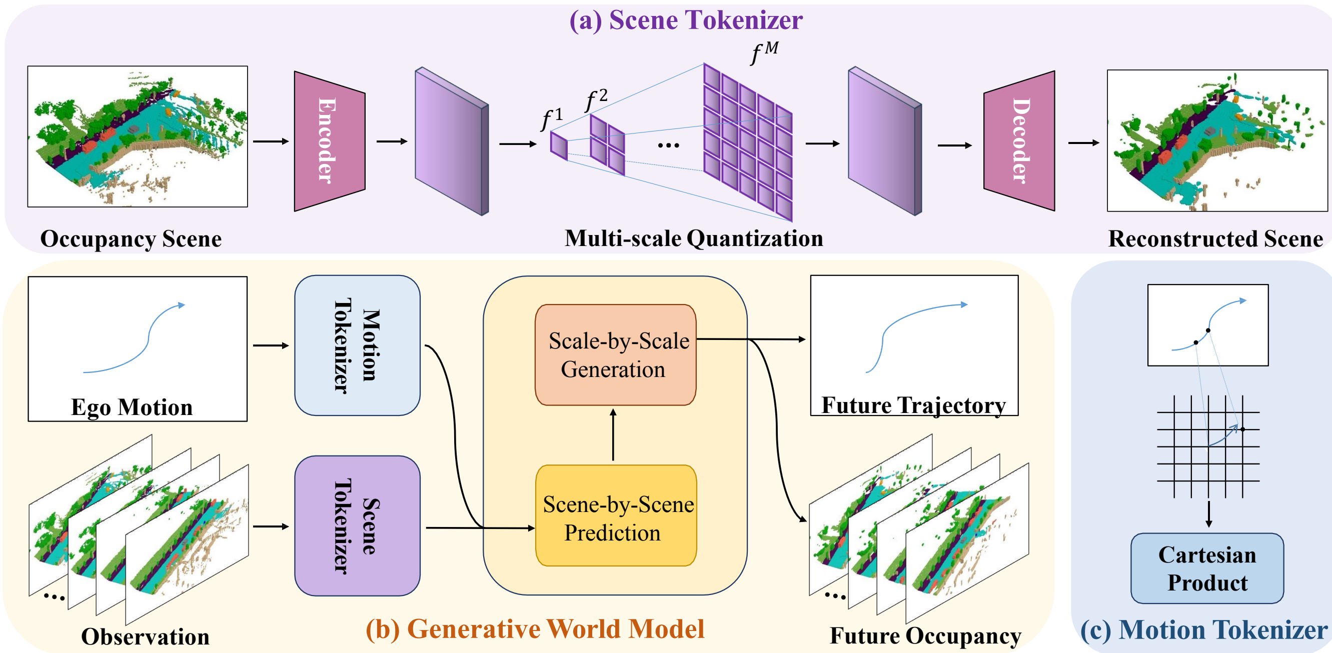

Fig. 2: **Overview of OccTENS**. OccTENS consists of two components: two tokenizers (a and c) that encode 3D occupancy and ego motion into discrete tokens, and a generative world model (b) using temporal next-scale prediction for future 3D scene forecasting.

to transform the feature into a collection of codebook entries through vector quantization, where each entry is responsible for a small area. However, operating tokenization solely on local information may result in the loss of global context. We utilize the multi-scale quantizer [21] to discretize the occupancy feature $\mathbf{F} \in \mathbb{R}^{H\times W\times C}$ to M multi-scale discrete token maps: $\mathbf{F} = (\mathbf{f}^1, \mathbf{f}^2, ..., \mathbf{f}^M)$.

The occupancy decoder takes the multi-scale quantized tokens as input and outputs the reconstructed 3D occupancy scenes. To reconstruct the 3D occupancy scene from the quantized tokens $\mathbf{F}$, we utilize another series of convolution layers to upsample the BEV feature map to the initial resolution, and then split the height dimension from the channel dimension, resulting in the reconstructed scene $\hat{\mathbf{S}}$.

2) Motion Tokenizer: The motion tokenizer is utilized to discretize the pose of the ego vehicle to better integrate it into our sequence prediction model. We utilize the position x, y and orientation θ relative to the previous frame to represent the motion of the vehicle. We discard the information in the z-axis because the vehicle's speed in the z-axis is nearly zero most of the time. We apply a vanilla uniform quantization of the motion information, resulting in V_x, V_y and V_θ tokens in the vocabulary. Then we map the relative motion with three discrete tokens to a motion token $\mathbf{P}$ by Cartesian product:

$$\mathbf{P} = \mathcal{E}(x + y \times V_x + \theta \times V_x \times V_y), \quad (1)$$

where $\mathcal{E}$ is an embedding layer. The discretized poses are then embedded into the sequence prediction model.

B. Generative World Model

In this section, we introduce our generative world model via temporal next-token prediction, as shown in Fig. 2.

1) Rethinking Autoregressive Models for Occupancy Generation: Some previous occupancy world models [14], [16] utilize a vanilla next-token autoregressive modeling on occupancy prediction. Considering the feature map with resolution $(n \times n)$, the likelihood of the sequence $x = \{x_1, x_2, ..., x_{n\times n}\}$ can be decomposed to the product of $n \times n$ conditional probabilities:

$$p(x) = \prod_{i=1}^{n\times n} p\left(x_t \mid x_1, x_2, \ldots, x_{i-1}\right) \quad (2)$$

However, such a vanilla modeling paradigm introduces several issues like inefficiency over longer autoregressive steps. A naive solution to address the inefficiency issue is to utilize next-scale prediction [21], which has demonstrated successful practice in class-to-image generation. However, the multi-scale nature makes it require longer token sequences. For example, assuming the scales are even distributed, given a feature map with $(n \times n)$ resolution and M scales, the total sequence number is:

$$n^2 \sum_{m=1}^{M} (\frac{m}{M})^2 \approx o(Mn^2) \quad (3)$$

When we adapt next-scale prediction for temporal modeling, the increasing token number can overload attention mechanisms, distributing their focus across expanding frame numbers. This results in progressive fragmentation of long-range dependencies, undermining global temporal coherence for long-term generation.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

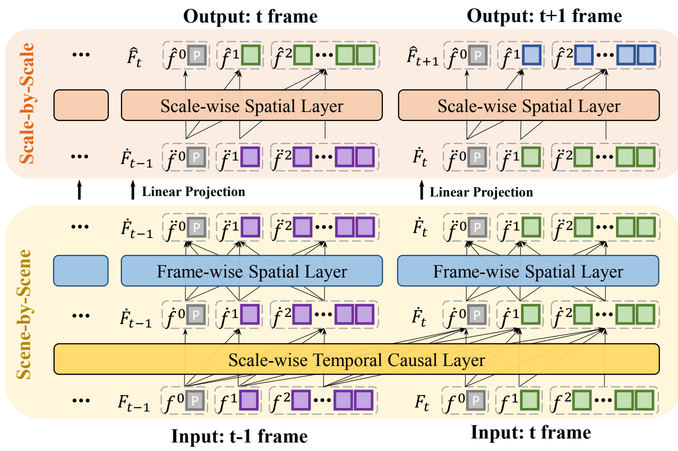

Fig. 3: **Temporal Next-scale Prediction.** The proposed TENSFormer decomposes the sequential occupancy generation into two distinct components: a scene-by-scene prediction and a scale-by-scale generation. The $\dot{F}$, $\ddot{F}$, $\hat{F}$ denote intermediate representations after the scale-wise temporal causal layer, after the frame-wise spatial layer and predicted logits for the next frame, respectively.

Moreover, the dependencies associated with scale also necessitate a comprehensive understanding of scale interaction over time. When generating a low-scale token, the model primarily needs to consider the previous low-scale information, as it encapsulates the global structural context like scene layouts that persists coherently over time. In contrast, higher-scale tokens, which capture finer-grained details such as textures, demand a global interaction between their own scale and the lower scales across frames. Thus, a trivial adaptation approach for next-scale prediction to temporal next-scale prediction is never sufficient.

To address the issues, OccTENS introduces a novel architecture, named **TENSFormer**, by decoupling the next-scale prediction within a single occupancy scene generation from the per-frame prediction across the occupancy sequences. This decoupling enables OccTENS to more effectively manage both temporal causality across different timestamps and spatial relationships across various scales, resulting in a more flexible and scalable framework for temporal occupancy modeling.

C. Architecture of TENSFormer

The proposed TENSFormer decomposes the sequential occupancy generation into two distinct components: a scene-by-scene prediction and a scale-by-scale generation, as detailed in Fig. 3. The scene-by-scene prediction is designed to model the dependencies between scales across consecutive frames, ensuring that temporal dynamics are accurately represented and integrated. Meanwhile, the scale-by-scale generation focuses on capturing the spatial composition of the occupancy scene within a single frame. To simultaneously facilitate pose control (controlling occupancy with given trajectory) and motion planning (planning a trajectory for the AD vehicle), we further introduce a multi-modal camera pose aggregation module tailored for auto-regressive generative models.

1) Temporal Scene-by-scene Prediction: The scene-by-scene prediction module aims to capture the temporal dependencies in occupancy sequences. Occupancy frames are naturally causal sequences, with each frame depending on its preceding frames. Therefore, we utilize a causal attention mechanism to ensure that each frame can only attend to its preceding frames. Given occupancy tokens $\{\mathbf{F}_1, ..., \mathbf{F}_{T-1}\}$ with T frame, the process can be formulated as:

$$p(\mathbf{F}_1, ..., \mathbf{F}_T) = \prod_{t=1}^{T} p(\mathbf{F}_t \mid \mathbf{F}_0, ..., \mathbf{F}_{t-1}) \tag{4}$$

Note that we add three 1-D sine-cosine embeddings to indicate the position information, scale information, and time information respectively, contributing to the spatial and temporal modeling. However, we find that this frame-wise causal attention may result in structure collapse when generating occupancy tokens with large scales, which we attribute to an inherent ambiguity in separating temporal dynamics from spatial dependencies, as the frame-wise causal attention conflates inter-frame causality with intra-frame bidirectional dependency. Thus we further decompose the frame-wise causal attention into two parts: scale-wise temporal causal attention and frame-wise spatial attention. For scale-wise causal attention, the tokens $\mathbf{f}_t^m$ of m-th scale in t-th frame can only attend to its prefix and the tokens in the same scale: $\{\mathbf{F}_1, \mathbf{F}_2, \ldots, \mathbf{F}_{t-1}, \mathbf{f}_t^1, \mathbf{f}_t^2, \ldots, \mathbf{f}_t^{m-1}\}$. For frame-wise spatial attention, we utilize a full attention to model the scale interaction within each frame. In this way, the causal attention and spatial attention are fully decoupled in scene-by-scene prediction, making it possible to synchronously model the temporal dependency for each occupancy scale. This enhances the training efficiency and allows the progressive training strategy [21] for effective scale modeling, where hierarchical scale-specific dependencies are incrementally optimized.

2) Spatial Scale-by-scale Generation: The scale-by-scale generation module focuses on generating the occupancy given the historical features. In this period, we inherit the outputs of the previous scene-by-scene prediction module as the **guidance features** to navigate spatial feature generation. When generating the m-th scale at t-th frame, the autoregressive likelihood is formulated as:

$$p(\mathbf{F'}_{t-1}, \hat{\mathbf{f}}_t^1, \ldots, \hat{\mathbf{f}}_t^M) = \prod_{m=1}^{M} p(\hat{\mathbf{f}}_t^m \mid \mathbf{F'}_t, \hat{\mathbf{f}}_t^1, \ldots, \hat{\mathbf{f}}_t^{m-1}) \tag{5}$$

where the $\mathbf{F'}_{t-1}$ is the guidance features and $\hat{\mathbf{f}}_t^m$ denotes the k-th scale tokens at t-th frame. Note that we utilize a block-wise causal attention mask to ensure that each token $\hat{\mathbf{f}}_t^m$ can only attend to its prefix $\{\mathbf{F}_{t-1}, \hat{\mathbf{f}}_t^1, \ldots, \hat{\mathbf{f}}_t^{m-1}\}$.

3) Multi-modal Camera Pose Aggregation: A key challenge for diffusion-based methods [15], [17] is that they can not conduct the motion planning task because they require the ground-truth future trajectory as a fixed conditioning input throughout diffusion sampling. At the same time, the existing auto-regressive solutions [13], [14], [16] suffer from controlling the occupancy generation with given camera pose (or trajectory). In OccTENS, we introduce a multi-modal camera pose aggregation module tailored for auto-regressive

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

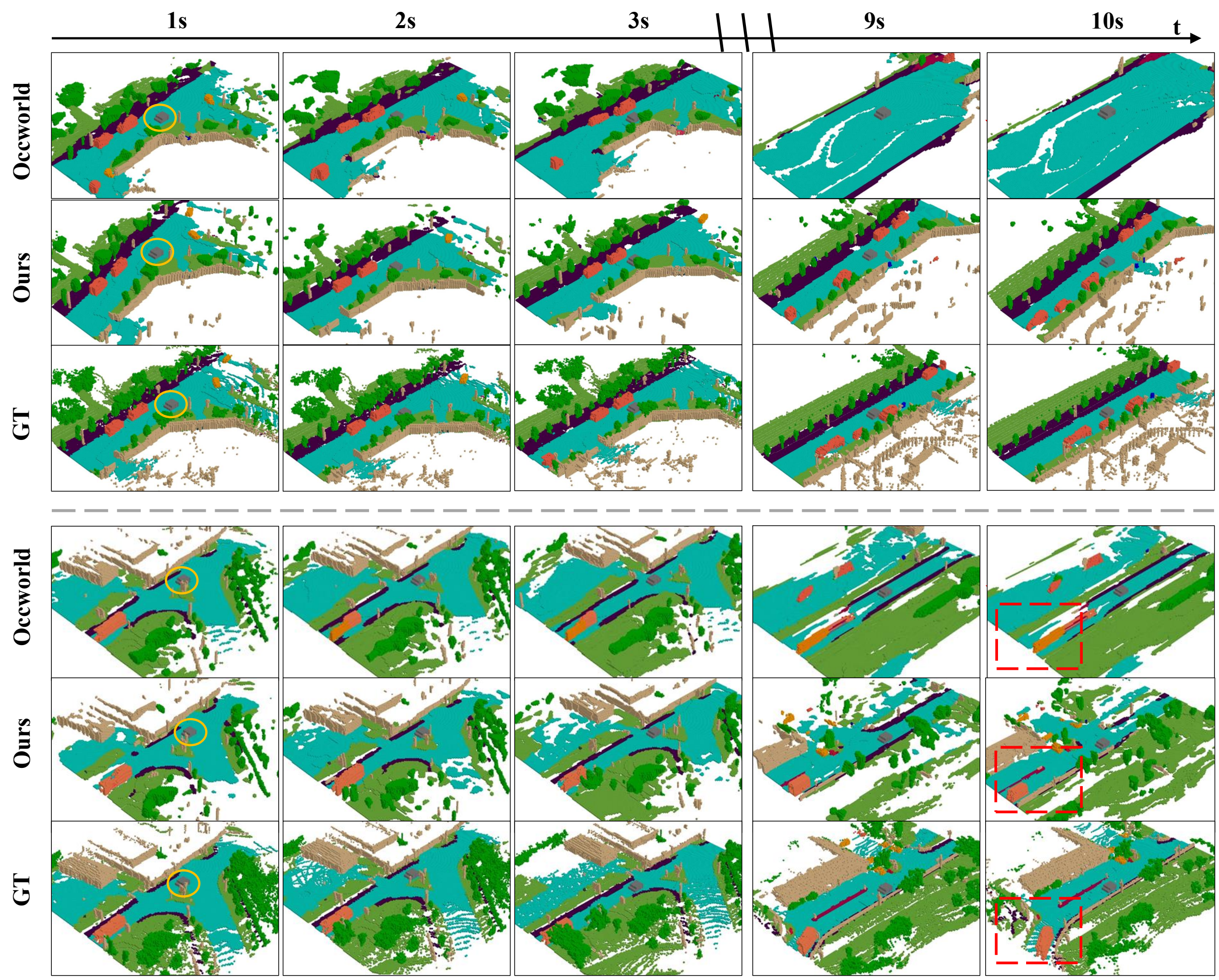

Fig. 4: **Qualitative results for long-term generation.** We compare OccTENS with OccWorld [13] in generating long sequences. OccWorld exhibits repetition artifacts. In contrast, OccTENS produces more diverse and realistic occupancy scenes. We mark the ego vehicle with an orange circle in the first column.

generative models, which enables motion planning and pose control simultaneously.

For auto-regressive models, we would add a BOS token before occupancy tokens, indicating the beginning of an occupancy frame sequence. To better integrate the ego-vehicle motion into our prediction paradigm, we treat the motion token as 0-th scale token $\mathbf{f}_0$ and splice it before the multi-scale occupancy tokens, resulting in $M+1$ tokens $\{\mathbf{f}_t^0, \mathbf{f}_t^1, ..., \mathbf{f}_t^M\}$ for each frame. Thus, the autoregressive likelihood is reformulated as

$$p(\mathbf{F'}_{t-1}, \hat{\mathbf{f}}_t^0, \hat{\mathbf{f}}_t^1, \ldots, \hat{\mathbf{f}}_t^M) = \prod_{m=0}^{M} p(\hat{\mathbf{f}}_t^m \mid \mathbf{F'}_t, \hat{\mathbf{f}}_t^0, \hat{\mathbf{f}}_t^1, \ldots, \hat{\mathbf{f}}_t^{m-1}) \tag{6}$$

D. Loss Function

When training the scene tokenizer, we utilize cross-entropy loss and lovasz-softmax loss [43]. To enhance the global occupancy reconstruction performance, we also utilize geoscal loss and semscal loss illustrated in [24], which optimize the class-wise derivable precision, recall and specificity for semantics and geometry. In general, our loss function is defined as: $\mathcal{L} = \lambda_1\mathcal{L}_{ce} + \lambda_2\mathcal{L}_{lovasz} + \lambda_3\mathcal{L}_{geoscal} + \lambda_4\mathcal{L}_{semscal}$, where the factors $\lambda_{1,2,3,4}$ are used to balance the losses.

When training the world model, we utilize cross-entropy loss for the generation of occupancy tokens and pose tokens, defined as: $\mathcal{L} = \beta_1\mathcal{L}_{occ} + \beta_2\mathcal{L}_{pose}$.

IV. Experiments

A. Experimental Setup

Datasets and Metrics. We evaluate our method on the nuScenes [44] dataset and Waymo dataset [45]. We employ the occupancy annotation in Occ3D [46] based on nuScenes and Waymo. Following common practices, we utilize a 2-second historical context (4 frames) and forecast the subsequent 3-second scenes (6 frames) unless specified.

Implementation Details. Our training period consists of 2 stages: tokenization and generation. For tokenization, we downsample the occupancy with a factor of 8. The codebook comprises 4096 nodes, and the channel dimension of the codebook entry is 128. We utilize 6 scales with [1,5,10,15,20,25] for multi-scale settings. In the tokenizer loss function, the λ_1, λ_2, λ_3, λ_4 are 10.0, 1.0, 0.3, 0.5 respectively. For generation, we utilize 4 layers each for three blocks of our methods.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

Method	Dataset	Input	MIOU(%)↑				IOU(%)↑			
			1s	2s	3s	Avg.	1s	2s	3s	Avg.
OccWorld-F [13]	nuScenes	Cam	8.03	6.91	3.54	6.16	23.62	18.13	15.22	18.99
OccLLaMA-F [14]	nuScenes	Cam	10.34	8.66	6.98	8.66	25.81	23.19	19.97	22.99
OccTENS-F (Ours)	nuScenes	Cam	**17.17**	**10.38**	**7.82**	**11.79**	**27.60**	**25.14**	**20.33**	**24.35**
OccWorld-O [13]	nuScenes	Occ	25.78	15.14	10.51	17.14	34.63	25.07	20.18	26.63
OccLLaMA-O [14]	nuScenes	Occ	25.05	19.49	15.26	19.93	34.56	28.53	24.41	29.17
OccTENS-O (Ours)	nuScenes	Occ	**27.96**	**21.75**	**16.47**	**22.06**	**38.73**	**29.50**	**24.86**	**31.03**
OccWorld-F [13]	Waymo	Cam	17.46	14.39	11.72	14.52	20.94	17.33	12.27	17.18
OccTENS-F	Waymo	Cam	**25.64**	**22.16**	**17.73**	**21.84**	**29.63**	**25.40**	**23.52**	**26.18**
OccWorld-O [13]	Waymo	Occ	32.04	25.77	23.76	27.19	36.04	30.48	27.96	31.49
OccTENS-O	Waymo	Occ	**34.17**	**27.04**	**25.47**	**28.89**	**38.72**	**31.85**	**30.28**	**33.62**

TABLE I: **Quantitative results of 4D occupancy forecasting on nuScenes and Waymo.** The "-O" represents the results utilizing ground truth occupancy as input. The "-F" represents that the input is multi-view camera images and we use FBOCC [27] to predict the occupancy from images.

The hidden dimension and head number are 128 and 4, respectively. The β_1 and β_2 are 1.0 and 1.0 respectively.

B. Main Results

The evaluation of OccTENS includes two tasks: 4D occupancy forecasting task and the motion planning task. The 4D occupancy forecasting task aims to forecast the future observation of 3D occupancy scenes. The motion planning task plans a trajectory for the ego vehicle.

4D Occupancy Forecasting In this experiment, we compare our method with state-of-the-art approaches on the 4D occupancy forecasting task. In fairness, we only report the occupancy generation with **only** historical observation as input, with no additional auxiliary inputs. Following common practice, we conduct our evaluation in two highlights: (1) using ground-truth 3D occupancy data (-O); and (2) using predicted results from FBOCC [27] based on camera data (-F). The results are shown in Tab. I. We observe that our OccTENS achieves significant performance gain over existing methods in short-time forecasting within 3 seconds. These results highlight the strong predictive performance of OccTENS, which sets the state-of-the-art on the nuScenes dataset and Waymo dataset. Moreover, as shown in Fig. 4, OccTENS demonstrates significantly improved performance in long-term sequence generation, producing coherent, high-fidelity occupancy scenes that closely mirror 3D dynamics.

Motion Planning We compare the motion planning performance of OccTENS with several strong baselines that utilize the same inputs and supervision methods. The results are shown in Tab. II. We observe that overall OccTENS surpasses existing occupancy world models in the motion planning task. Interestingly, OccTENS achieves improvements in planning results, especially at 2s and 3s. We attribute this to the fact that OccTENS explicitly models the dynamics during occupancy generation, enabling robust long-horizon trajectory optimization through accurate anticipation of future occupancy changes, which is critical for navigating complex, unstructured scenarios. We also include the results of several SOTA planning methods for reference. While our results are lower than methods like UniAD, this is expected since UniAD benefits from extensive multi-task supervision (e.g., detection, tracking, and map annotation). In contrast, OccTENS focuses on a more self-contained and generative world modeling setup, relying solely on occupancy-based representations without such auxiliary labels. This comparison highlights the complementary nature of these approaches and underscores the potential of occupancy-centric modeling even under reduced supervision.

Method	L2(m)↓				Coll.(%)↓			
	1s	2s	3s	Avg.	1s	2s	3s	Avg.
OccWorld [13]	0.43	1.08	1.99	1.17	0.07	0.38	1.35	0.60
OccLlama [14]	0.37	1.02	2.03	1.14	0.04	0.24	1.20	0.49
Ours	0.39	**1.02**	**1.96**	**1.12**	0.08	0.25	**1.12**	**0.48**
UniAD [1]	0.48	0.96	1.65	1.03	0.05	0.17	0.71	0.31
VAD [2]	0.41	0.70	1.05	0.72	0.07	0.17	0.41	0.22
PPAD [47]	0.31	0.56	0.87	0.58	0.08	0.12	0.38	0.19
GenAD [37]	0.28	0.49	0.78	0.52	0.08	0.14	0.34	0.19
LAW [48]	0.24	0.46	0.76	0.49	0.08	0.10	0.39	0.19

TABLE II: **Motion planning results.** We can see that OccTENS gets a better planning performance. We include SOTA end-to-end planning methods for reference.

Scale Num.	mIoU(%)	Avg. Gen. IoU(%)	Latency(s)
OccWorld	17.14	26.63	0.35
OccLlama	19.93	29.17	-
OccSora	-	-	$\tilde{2}0$
Ours-2-scales	18.35	27.31	**0.21**
Ours-4-scales	21.24	29.72	0.34
Ours-6-scales	22.06	31.03	0.56
Ours-8-scales	**22.23**	**31.24**	0.93

TABLE III: **Efficiency analysis**. OccTENS can get a higher performance with a faster inference time. The "-" represents the unreported results. The gray square represents the setting we used in the previous experiment.

C. Ablation Results

To delve into the effect of each module, we conduct a comprehensive ablation study on OccTENS on efficiency and

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

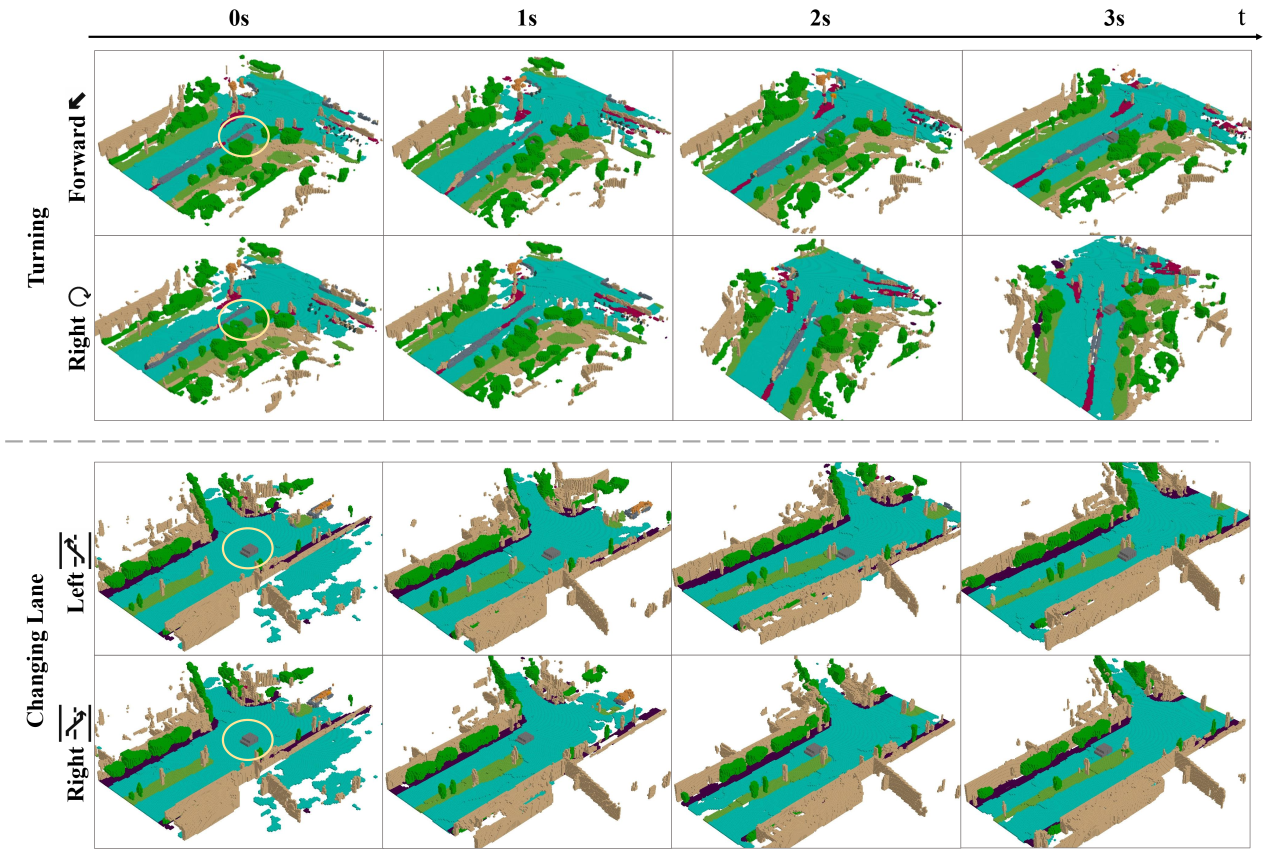

Fig. 5: **Qualitative results of pose controllability**. OccTENS successfully generates results aligned with the pose input like turning (top) or changing lane (bottom), indicating the superior controllability of OccTENS.

Decomp	TNSP	Pose Agg	Latency(s)	mIoU(%)	IoU(%)
	✓	✓	2.47	17.14	26.63
✓		✓	5.13	19.20	27.32
✓	✓		0.59	19.93	29.17
✓	✓	✓	**0.56**	**22.06**	**31.03**

TABLE IV: **Component-wise ablation.** Each component contributes to the improvement of the performance.

pose controllability.

Efficiency. The latency is of great significance for the deployment of the autonomous driving system. In this experiment, we compare OccTENS with existing works. It should be emphasized that OccTENS can indeed achieve competitive speed if we reduce the number of scales. Our choice to use 6 scales aims to strike an optimal trade-off between performance and efficiency. However, we can observe that the OccTENS with 2 scales outperforms OccWorld in both generation performance and inference time, demonstrating the effectiveness and efficiency of the proposed architecture.

Our findings show that increasing the number of scales consistently improves both reconstruction and generation performance. However, this improvement comes at the cost of increased latency, indicating that a trade-off between performance and efficiency must be maintained.

Pose Controllability Controllability refers to the model's capacity to precisely adhere to these inputs, ensuring that the generated scenes reflect the specified conditions with high fidelity. Pose control is particularly critical, as it ensures that the model generates scenes from the correct viewpoint and perspective. We manipulate the camera pose inputs and measure the generation results in terms of alignment. As shown in Fig. 5, OccTENS can generate the corresponding results that are collaboratively aligned with the conditional motion input, indicating the powerful generalization ability of our method.

Component-wise ablation We conduct a component-wise ablation study to illustrate the effectiveness of our design, including: (1) decomposing Architecture (with/without), (2) next-token prediction (NTP) vs. temporal next-scale prediction (TNSP), (3) Multi-modal pose aggregation (with/without). The results are shown in Table IV. We can observe that each component of our architecture contributes to the improvement of the performance. We can also observe that our design of decomposing architecture (first row) and temporal next-scale prediction (second row) can reduce the inference latencies of the occupancy world model.

V. Conclusions

In this paper, we present OccTENS, an autoregressive occupancy world model designed to generate controllable, high-fidelity long-term occupancy generation while maintaining computational efficiency. By reformulating the occupancy world model as a temporal next-scale prediction (TENS) task, OccTENS is capable of effectively managing the temporal

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

causality and spatial relationships of occupancy sequences. Extensive evaluations demonstrate the superiority of OccTENS on fidelity, efficiency, and controllability, surpassing existing methods. These results highlight the potential of OccTENS to facilitate real-time applications in autonomous driving, paving the way for future advancements in world models.

REFERENCES

[1] Y. Hu, J. Yang, L. Chen, K. Li, C. Sima, X. Zhu, S. Chai, S. Du, T. Lin, W. Wang *et al.*, "Planning-oriented autonomous driving," in *CVPR*, 2023.

[2] B. Jiang, S. Chen, Q. Xu, B. Liao, J. Chen, H. Zhou, Q. Zhang, W. Liu, C. Huang, and X. Wang, "Vad: Vectorized scene representation for efficient autonomous driving," in *ICCV*, 2023.

[3] S. Hu, L. Chen, P. Wu, H. Li, J. Yan, and D. Tao, "St-p3: End-to-end vision-based autonomous driving via spatial-temporal feature learning," in *ECCV*, 2022.

[4] Z. Huang, H. Liu, J. Wu, and C. Lv, "Differentiable integrated motion prediction and planning with learnable cost function for autonomous driving," *IEEE TNNLS*, 2023.

[5] H. Yang, S. Zhang, D. Huang, X. Wu, H. Zhu, T. He, S. Tang, H. Zhao, Q. Qiu, B. Lin *et al.*, "Unipad: A universal pre-training paradigm for autonomous driving," in *CVPR*, 2024.

[6] X. Wang, Z. Zhu, G. Huang, X. Chen, J. Zhu, and J. Lu, "Drivedreamer: Towards real-world-drive world models for autonomous driving," in *ECCV*, 2024.

[7] S. Gao, J. Yang, L. Chen, K. Chitta, Y. Qiu, A. Geiger, J. Zhang, and H. Li, "Vista: A generalizable driving world model with high fidelity and versatile controllability," *Neurips*, 2024.

[8] R. Gao, K. Chen, E. Xie, L. Hong, Z. Li, D.-Y. Yeung, and Q. Xu, "Magicdrive: Street view generation with diverse 3d geometry control," *arXiv preprint arXiv:2310.02601*, 2023.

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

[10] Y. Wang, J. He, L. Fan, H. Li, Y. Chen, and Z. Zhang, "Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving," in *CVPR*, 2024.

[11] K. Yang, E. Ma, J. Peng, Q. Guo, D. Lin, and K. Yu, "Bevcontrol: Accurately controlling street-view elements with multi-perspective consistency via bev sketch layout," *arXiv preprint arXiv:2308.01661*, 2023.

[12] A. Swerdlow, R. Xu, and B. Zhou, "Street-view image generation from a bird's-eye view layout," *IEEE RAL*, 2024.

[13] W. Zheng, W. Chen, Y. Huang, B. Zhang, Y. Duan, and J. Lu, "Occworld: Learning a 3d occupancy world model for autonomous driving," *ECCV*, 2024.

[14] J. Wei, S. Yuan, P. Li, Q. Hu, Z. Gan, and W. Ding, "Occllama: An occupancy-language-action generative world model for autonomous driving," *arXiv preprint arXiv:2409.03272*, 2024.

[15] L. Wang, W. Zheng, Y. Ren, H. Jiang, Z. Cui, H. Yu, and J. Lu, "Occsora: 4d occupancy generation models as world simulators for autonomous driving," *arXiv preprint arXiv:2405.20337*, 2024.

[16] T. Xu, H. Lu, X. Yan, Y. Cai, B. Liu, and Y. Chen, "Occ-llm: Enhancing autonomous driving with occupancy-based large language models," *arXiv preprint arXiv:2502.06419*, 2025.

[17] S. Gu, W. Yin, B. Jin, X. Guo, J. Wang, H. Li, Q. Zhang, and X. Long, "Dome: Taming diffusion model into high-fidelity controllable occupancy world model," *arXiv preprint arXiv:2410.10429*, 2024.

[18] B. Li, J. Guo, H. Liu, Y. Zou, Y. Ding, X. Chen, H. Zhu, F. Tan, C. Zhang, T. Wang *et al.*, "Uniscene: Unified occupancy-centric driving scene generation," *arXiv preprint arXiv:2412.05435*, 2024.

[19] B. Agro, Q. Sykora, S. Casas, T. Gilles, and R. Urtasun, "Uno: Unsupervised occupancy fields for perception and forecasting," in *CVPR*, 2024.

[20] C. Diehl, Q. Sykora, B. Agro, T. Gilles, S. Casas, and R. Urtasun, "Dio: Decomposable implicit 4d occupancy-flow world model," in *CVPR*, 2025.

[21] K. Tian, Y. Jiang, Z. Yuan, B. Peng, and L. Wang, "Visual autoregressive modeling: Scalable image generation via next-scale prediction," *Neurips*, 2024.

[22] Y. Wei, L. Zhao, W. Zheng, Z. Zhu, J. Zhou, and J. Lu, "Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving," in *ICCV*, 2023.

[23] W. Tong, C. Sima, T. Wang, L. Chen, S. Wu, H. Deng, Y. Gu, L. Lu, P. Luo, D. Lin *et al.*, "Scene as occupancy," in *ICCV*, 2023.

[24] A.-Q. Cao and R. De Charette, "Monoscene: Monocular 3d semantic scene completion," in *CVPR*, 2022.

[25] Y. Huang, W. Zheng, Y. Zhang, J. Zhou, and J. Lu, "Tri-perspective view for vision-based 3d semantic occupancy prediction," *arXiv preprint arXiv:2302.07817*, 2023.

[26] Q. Ma, X. Tan, Y. Qu, L. Ma, Z. Zhang, and Y. Xie, "Cotr: Compact occupancy transformer for vision-based 3d occupancy prediction," in *CVPR*, 2024.

[27] Z. Li, Z. Yu, D. Austin, M. Fang, S. Lan, J. Kautz, and J. M. Alvarez, "Fb-occ: 3d occupancy prediction based on forward-backward view transformation," *arXiv preprint arXiv:2307.01492*, 2023.

[28] Y. Wang, Y. Chen, X. Liao, L. Fan, and Z. Zhang, "Panoocc: Unified occupancy representation for camera-based 3d panoptic segmentation," in *CVPR*, 2024.

[29] Y. Zheng, X. Li, P. Li, Y. Zheng, B. Jin, C. Zhong, X. Long, H. Zhao, and Q. Zhang, "Monoocc: Digging into monocular semantic occupancy prediction," *arXiv preprint arXiv:2403.08766*, 2024.

[30] M. Pan, J. Liu, R. Zhang, P. Huang, X. Li, H. Xie, B. Wang, L. Liu, and S. Zhang, "Renderocc: Vision-centric 3d occupancy prediction with 2d rendering supervision," in *ICRA*, 2024.

[31] Y. Huang, W. Zheng, B. Zhang, J. Zhou, and J. Lu, "Selfocc: Self-supervised vision-based 3d occupancy prediction," in *CVPR*, 2024.

[32] D. Ha and J. Schmidhuber, "World models," *arXiv preprint arXiv:1803.10122*, 2018.

[33] G. Zhao, X. Wang, Z. Zhu, X. Chen, G. Huang, X. Bao, and X. Wang, "Drivedreamer-2: Llm-enhanced world models for diverse driving video generation," *arXiv preprint arXiv:2403.06845*, 2024.

[34] J. Su, S. Gu, Y. Duan, X. Chen, and J. Luo, "Text2street: Controllable text-to-image generation for street views," *arXiv preprint arXiv:2402.04504*, 2024.

[35] X. Wang, Z. Zhu, G. Huang, X. Chen, J. Zhu, and J. Lu, "Drivedreamer: Towards real-world-driven world models for autonomous driving," *ECCV*, 2024.

[36] J. Lu, Z. Huang, J. Zhang, Z. Yang, and L. Zhang, "Wovogen: World volume-aware diffusion for controllable multi-camera driving scene generation," *ECCV*, 2024.

[37] W. Zheng, R. Song, X. Guo, and L. Chen, "Genad: Generative end-to-end autonomous driving," *ECCV*, 2024.

[38] J. Jiang, G. Hong, L. Zhou, E. Ma, H. Hu, X. Zhou, J. Xiang, F. Liu, K. Yu, H. Sun *et al.*, "Dive: Dit-based video generation with enhanced control," *arXiv preprint arXiv:2409.01595*, 2024.

[39] H. Deng, T. Pan, H. Diao, Z. Luo, Y. Cui, H. Lu, S. Shan, Y. Qi, and X. Wang, "Autoregressive video generation without vector quantization," *arXiv preprint arXiv:2412.14169*, 2024.

[40] T. Li, Y. Tian, H. Li, M. Deng, and K. He, "Autoregressive image generation without vector quantization," *Neurips*, 2024.

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

[42] V. Zyrianov, H. Che, Z. Liu, and S. Wang, "Lidardm: Generative lidar simulation in a generated world," *arXiv preprint arXiv:2404.02903*, 2024.

[43] M. Berman, A. R. Triki, and M. B. Blaschko, "The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks," in *CVPR*, 2018.

[44] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *CVPR*, 2020.

[45] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine *et al.*, "Scalability in perception for autonomous driving: Waymo open dataset," in *CVPR*, 2020.

[46] X. Tian, T. Jiang, L. Yun, Y. Mao, H. Yang, Y. Wang, Y. Wang, and H. Zhao, "Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving," *Neurips*, vol. 36, 2024.

[47] Z. Chen, M. Ye, S. Xu, T. Cao, and Q. Chen, "Ppad: Iterative interactions of prediction and planning for end-to-end autonomous driving," in *ECCV*. Springer, 2024, pp. 239–256.

[48] Y. Li, L. Fan, J. He, Y. Wang, Y. Chen, Z. Zhang, and T. Tan, "Enhancing end-to-end autonomous driving with latent world model," *arXiv preprint arXiv:2406.08481*, 2024.