SparseWorld: A Flexible, Adaptive, and Efficient 4D Occupancy World Model Powered by Sparse and Dynamic Queries

Chenxu Dang[1,2*], **Haiyan Liu**[3], **Jason Bao**[3], **Pei An**[1], **Xinyue Tang**[3], **PanAn**[4], **Jie Ma**[1], **Bingchuan Sun**[3†], **Yan Wang**[2†]

[1] Huazhong University of Science and Technology
[2] Institute for AI Industry Research (AIR), Tsinghua University
[3] Lenovo Group Limited
[4] AIR Wuxi Innovation Center, Tsinghua University (AIRIC)

Abstract

Semantic occupancy has emerged as a powerful representation in world models for its ability to capture rich spatial semantics. However, most existing occupancy world models rely on static and fixed embeddings or grids, which inherently limit the flexibility of perception. Moreover, their "in-place classification" over grids exhibits a potential misalignment with the dynamic and continuous nature of real scenarios. In this paper, we propose **SparseWorld**, a novel 4D occupancy world model that is flexible, adaptive, and efficient, powered by **sparse** and **dynamic** queries. We propose a Range-Adaptive Perception module, in which learnable queries are modulated by the ego vehicle states and enriched with temporal-spatial associations to enable extended-range perception. To effectively capture the dynamics of the scene, we design a State-Conditioned Forecasting module, which replaces classification-based forecasting with regression-guided formulation, precisely aligning the dynamic queries with the continuity of the 4D environment. In addition, We specifically devise a Temporal-Aware Self-Scheduling training strategy to enable smooth and efficient training. Extensive experiments demonstrate that SparseWorld achieves state-of-the-art performance across perception, forecasting, and planning tasks. Comprehensive visualizations and ablation studies further validate the advantages of SparseWorld in terms of flexibility, adaptability, and efficiency.

Code — https://github.com/MSunDYY/SparseWorld

Introduction

In recent years, vision-centric end-to-end autonomous driving, which predicts the ego vehicle's future trajectory from monocular or multi-view images, has gained significant attention. Among them, occupancy-based world models (Zheng et al. 2024; Wei et al. 2024; Li et al. 2025) leverage semantic occupancy representations for rich spatial understanding and have shown superior planning performance.

Early occupancy world models (Zheng et al. 2024; Gu et al. 2024), as shown in Fig. 1(a), independently encode each frame's occupancy into embeddings, which are later fused and decoded by a future-oriented world model. Such decoupled designs separate forecasting from perception, hindering gradient flow and end-to-end optimization, while the repeated encode–decode process inevitably loses fine-grained information, no matter how carefully it is designed.

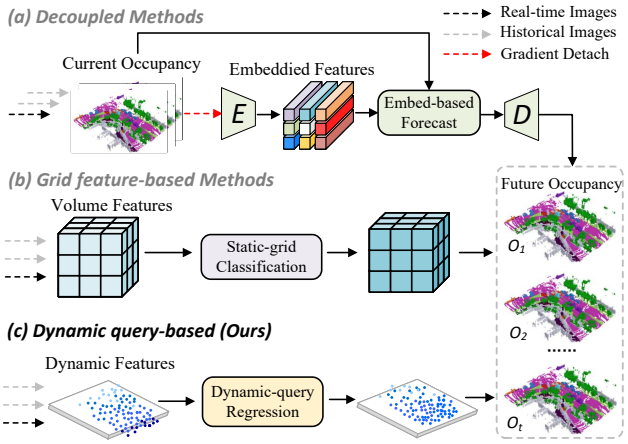


Figure 1: **(a)** Perception-forecasting-decoupled methods. **(b)** Grid feature-based methods. **(c)** We adopt dynamic query representations, which facilitate continuous and coherent 4D scene forecasting and planning.

Recently, a series of grid feature-based works (Li et al. 2025; Yang et al. 2025) have adopted perception features as intermediate representations for per-grid forecasting, as show in Figure 1(b), thus enabling end-to-end optimization. However, their static "in-place classification" operation misaligns with the continuity of ego-motion and scene dynamics, causing temporal inconsistency, feature drift, and cumulative errors.

Both aforementioned world models rely on manually predefined spatial ranges, limiting perceptional flexibility and adaptivity. In real-world driving scenarios, where vehicle speed varies drastically, adaptive perception ranges are critical for accurate forecasting and planning. Moreover, dense grids incur high computational and memory costs while ig-

*Working done during the internship at AIR.
†Corresponding Authors.

noring the inherently sparse nature of the physical world.

To overcome these limitations, inspired by recent sparse occupancy perception models (Wang et al. 2024a; Liu et al. 2024), we propose SparseWorld, a fully sparse 4D occupancy world model built on sparse and dynamic queries, as show in Figure 1(c). Following a "perceive-then-forecast" paradigm, SparseWorld adaptively constructs extended-range occupancy queries, and regresses the future motion of scene elements relative to the ego vehicle, moving beyond static grid classification.

We introduce a **Range-Adaptive Perception** module that takes learnable queries as input and employs stacked decoders featuring temporal-spatial fusion. To fully enhance the perception adaptability, an Adaptive Scaling sub-module encodes ego vehicle's historical trajectory to modulate the initial distribution of queries. Leveraging the continuity and dynamics of queries, we further design a **State-Conditioned Forecasting** module, where the ego query interacts with scene queries via spatial modulation,and regression-guided migration replaces traditional "in-place classification" to for continuous and plausible motion forecasting.

We further introduce a targeted **Temporal-Aware Self-Scheduling** training strategy that implicitly partitions query timestamps, allowing the model to autonomously learn timestamp assignments during training, which significantly improves training efficiency and autonomy.

To validate the effectiveness of SparseWorld, we conducted extensive experiments on the Occ3d-nuScenes (Tian et al. 2023) benchmark, comparing it with other state-of-the-art methods. The experimental results demonstrat that SparseWorld significantly outperforms dense models in both forecasting and planning tasks. Specifically, SparseWorld surpasses PreWorld (Li et al. 2025) by 20%–40% mIoU in future occupancy forecasting, and reduces collision rate in trajectory planning by half. Moreover, our SparseWorld achieves an approximate 7x speedup in inference compared to dense methods, greatly enhancing its practicality for real-world deployment.

Our main contributions can be summarized as follows:

- We propose a sparse 4D occupancy world model powered by sparse and dynamic queries for flexible, adaptive, and efficient modeling of autonomous driving scenarios.
- We propose a Range-Adaptive Perception module featuring the ego vehicle' state and introduce a regression-oriented State-Conditioned Forecasting paradigm that effectively exploits the spatiotemporal continuity to improve 4D scene evolution forecasting.
- We develop a novel Temporal-Aware Self-Scheduling training strategy to enable smooth and efficient training.
- Our SparseWorld significantly outperforms state-of-the-art methods in both effectiveness and efficiency. We design comprehensive ablation studies, complemented by visualizations, to support our claims.

Related Work

3D Occupancy Prediction

Occupancy perception methods can be broadly categorized into dense and sparse paradigms according to their modeling strategies. Dense ones (Ma et al. 2024; Yu et al. 2023; Huang and Huang 2022; Wu et al. 2024; Kim et al. 2025; Wang et al. 2024b; Zhang et al. 2024b; Kim et al. 2025; Ye et al. 2024; Li et al. 2023b) typically construct BEV or volume features that are conceptually straightforward but incur significant computations and limited flexibility. In contrast, sparse approaches (Liu et al. 2024; Wang et al. 2024a; Shi et al. 2024; Li et al. 2023a) eliminate the reliance on dense representations. Recently, some weakly- and self-supervised approaches (Pan et al. 2024; Huang et al. 2024; Boeder, Gigengack, and Risse 2025; Sun et al. 2024a) has emerged to alleviate the reliance on expensive 3D annotations.

4D Occupancy World Models

World models aim to predict future scenes and plan ego-agent trajectories based on historical observations and actions (Ha and Schmidhuber 2018; Gao et al. 2023; Yang et al. 2024; Gao et al. 2024). Occupancy world models are required to simultaneously forecast future occupancy scenes and plan trajectories. Early works such as OccWorld (Zheng et al. 2024) and its variants (Wei et al. 2024; Xu et al. 2025) decouple occupancy perception and prediction by first encoding the observed scene and then forecasting autoregressively followed by re-encoding and decoding. Recent grid-based methods attempt to unify perception and prediction by constructing volumetric (Li et al. 2025) or BEV (Yang et al. 2025) features to represent the spatiotemporal world consistently.

End to End Autonomous Driving

Planning-oriented end-to-end autonomous driving typically requires precise environmental perception. SP-T3 (Hu et al. 2022) and UniAD (Hu et al. 2023) represent the scene with a unified BEV representation, while VAD (Jiang et al. 2023) introduces a vectorized design. Zhang et al. (2024a) and Sun et al. (2024b) perform perception and planning sequentially based on sparse perception signals, whereas the more recent DriveTransformer (Jia et al. 2025) decouple and execute perception and planning in parallel, demonstrating superior performance in closed-loop inference. Some recent works have explored the use of anchor-based(Chen et al. 2024) and diffusion(Liao et al. 2025; Zheng et al. 2025) to encourage diverse trajectory outputs. In our work, we adopt autoregressive generation with L1 supervision for trajectory prediction for fair comparison.

Methodology

Preliminary

A practical and coherent AD world model is expected to take advantage of current and past p frames of the ego vehicle waypoints $\{w^i\}_{i=-p}^{0}$ and sensor inputs $\{s^i\}_{i=-p}^{0}$ to predict the current and future f frames of the semantic occupancy representations $\{o^i\}_{i=0}^{f}$ and the corresponding planning waypoints $\{w^i\}_{i=0}^{f}$. Decoupled methods embed historical occupancy observations $\{o^i\}_{i=-p}^{0}$ into compact latent embeddings and forecast the corresponding future latent codes, followed with occupancy decoders to reconstruct

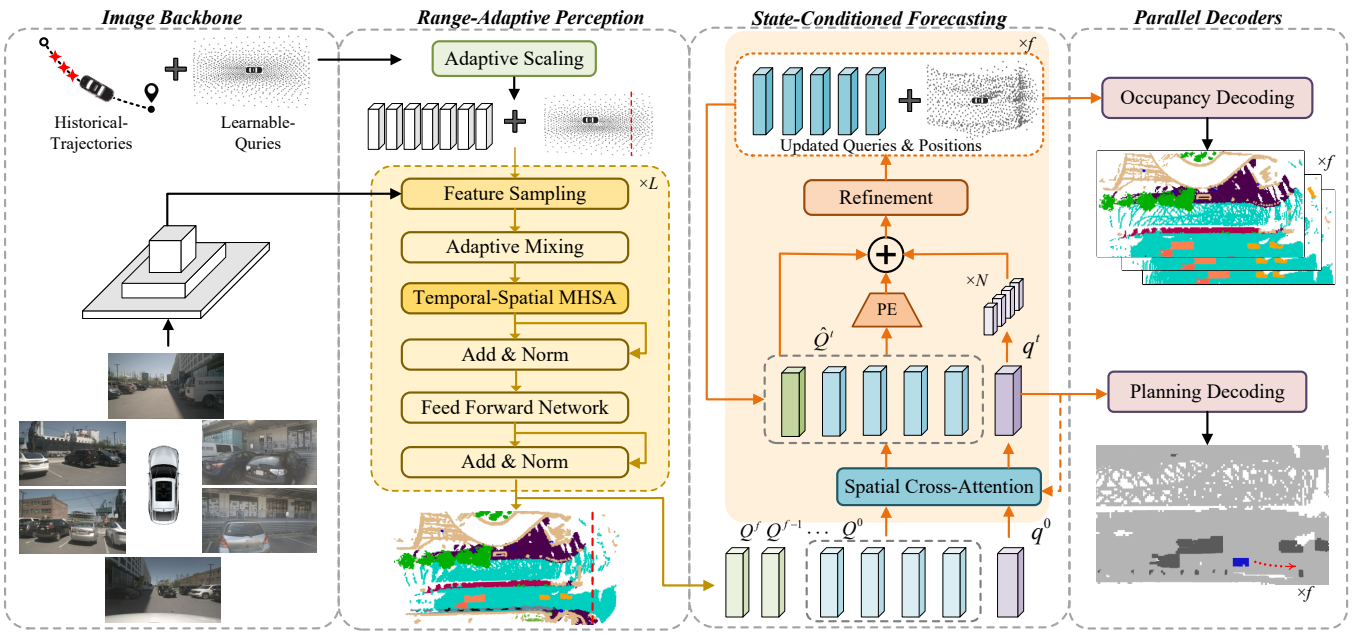


Figure 2: The overall architecture of **SparseWorld**. A set of learnable queries encoded with the historical trajectory through an Adaptive Scaling are then fed into stacked spatio-temporal decoders, which interacts with multi-frame, multi-view image features via a Temporal-Spatial MHSA. Subsequently, the extended-range queries are dynamically refined through a State-Conditional Forecasting module, which refines their positions while guiding the ego-vehicle's motion state in an autoregressive manner. Finally, two parallel decoders are employed for forecasting and planning, respectively.

future occupancy states:

$$z^t = \mathcal{E}(o^t), z^{t+1} = \mathcal{W}(z^t, z^{t-1}...), o^{t+1} = \mathcal{D}(z^{t+1}). \quad (1)$$

Here, $\mathcal{E}, \mathcal{D}, \mathcal{W}$ denote occupancy encoder, occupancy decoder, and latent embedding forecaster, respectively. In contrast, grid-based methods employ static volume features to perform per-voxel "in-place classification" forecasting:

$$F^0 = \mathcal{N}(s^{-p}, ..., s^0), F^t = \mathcal{F}(F^{t-1}), o^t = \mathcal{H}(F^t) \quad (2)$$

Here, F denotes the volume feature, while $\mathcal{N}$, $\mathcal{F}$, and $\mathcal{H}$ represent the perception, forecasting and occupancy head modules, respectively.

As illustrated in Figure 2, the overall architecture of SparseWorld consists of four components: (1) a generic image backbone that extracts multi-scale visual features over multiple frames; (2) a Range-Adaptive Perception module (**RAP**) that is composed of L decoder layers and detailed in Section 3.2; (3) a State-Conditioned Forecasting module (SCF) that is described in Section 3.3; (4) parallel decoding heads for 4D occupancy forecasting and motion planning. In Section 3.4, we will detail our Temporal-Aware Self-Scheduling training strategy.

Range-Adaptive Perception

The core of world models lies in forecasting the dynamics of the surrounding scene under the ego agent's motion conditions. However, existing grid-based perception models, constrained by fixed spatial resolution and truncated receptive fields, are inherently limited in handling dynamic scenarios.

To address these limitations, we adopt a more flexible dynamic query formulation. As shown in Figure 2, the input of RAP consists of learnable query embeddings and corresponding 4D coordinates (x, y, z, timestamp). Compared to dense grids, sparse queries offer multiple advantages: (1) **Flexible**: The adaptive spatial distribution enables ultra-range perception. (2) **Continuous**: It aligns with the spatiotemporal dynamics of real-world scenarios. (3) **Efficient**: It significantly reduces the storage and computational costs.

Following the driving intuition that faster speeds require longer perception ranges, the perception range should adapt to the ego vehicle's speed. We design an ego-guided Adaptive Scaling module to flexibly scale the perception range. Specifically, we encode historical ego waypoints $\{w^i\}_{i=-p}^0$, which implicitly reflect the current velocity, to modulate the initial coordinates $P_0 = \{\mathbf{p}_i\}_{i=1}^N$ of queries Q_0:

$$\begin{aligned} \boldsymbol{\gamma} &= [\gamma_x, \gamma_y, \gamma_z] = \text{MLP}([w^{-p}, ..., w^{-1}, w^0]) \\ \mathbf{p}_i' &= \boldsymbol{\gamma} \odot \mathbf{p_i} = [\gamma_x x_i, \gamma_y y_i, \gamma_z z_i] \end{aligned} \quad (3)$$

Here, N denotes the number of queries and P_0 is learnable, which will be explained later. The initial queries Q_0 combined with modulated positions $P_0' = \{\mathbf{p}_i'\}_{i=1}^N$, regardless of timestamps, are fed info L stacked decoders for extended-range occupancy perception for current moment.

The details of Decoder Our stacked decoders follow a coarse-to-fine paradigm. Each query samples semantic information from multi-view, multi-scale feature maps extracted by the image backbone, followed by adaptive mixing introduced in (Liu et al. 2023, 2024). The queries are then

fed into Temporal-Spatial Multi-Head Self-Attention (TS-MHSA) for spatiotemporal interaction.

The attention weights of Temporal-Spatial MHSA are composed of three components: **semantic similarity**, **spatial proximity**, and **temporal causality**. Formally, for any two queries q_i, q_j with 4D coordinates (x_i, y_i, z_i, t_i) and (x_j, y_j, z_j, t_j), the attention score A_{ij} is computed as:

$$A_{ij} = q_i^\top q_j - \tau_i ||\boldsymbol{p_i} - \boldsymbol{p_j}||^2 + M_{ij} \tag{4}$$

where $\boldsymbol{p_i} = (x_i, y_i, z_i)$, τ_i is a learnable scaling factor controlling spatial bias inspired by SparseBEV (Liu et al. 2023), and M_{ij} is a temporal masking term defined as:

$$M_{ij} = \begin{cases} 0, & \text{if } t_i \geq t_j \\ -\infty, & \text{otherwise} \end{cases} \tag{5}$$

This design ensures the temporal consistency by avoiding temporal interference.

Each decoder layer is followed by an occupancy head that outputs multiple points per query. Across decoder layers, we gradually increase the number of output points and update the query positions based on the mean of the output points. From the final decoder layer, we obtain the extended-range scene queries $Q = \{q_i\}_{i=1}^N$ and their updated 3D positions $P = \{p_i\}_{i=1}^N$ at the current timestamp. We supervise the outputs of each decoder layer individually.

State-Conditioned Continuous Forecasting

Leveraging the continuous and dynamic nature of queries, we design a regression-guided continuous forecasting strategy, conditioned on the ego state. The extended-range perception queries Q, are temporally partitioned into $Q^0, Q^1, ..., Q^f$ according to their timestamps.

Following existing methods, we encode ego state as query. At any time step t, the ego query q^t interacts with the scene query Q^t through spatial cross-attention for next state q^{t+1}. The attention weights between q^t and any scene query q_i^t consist of two components:

$$A_i^t = (q^t)^\top q_i^t - \tau_i^t ||p_i^t||^2, \quad \tau_i^t = \text{MLP}(q_i^t) \tag{6}$$

Then, the dynamic scene of next frame is forecasted as:

$$\hat{Q}^{t+1} = [\hat{Q}^t, Q^{t+1}] + \text{PE}([P^t, P^{t+1}]) + \text{repeat}(q^{t+1}) \tag{7}$$

Here, $[\cdot, \cdot]$ denotes concatenation for next scene augmentation, PE represents the 4D position encoding, repeat denotes broadcasting the ego query to match $\hat{Q}^{t+1}$. This design allows the model to capture both the global motion of the ego vehicle and the local state of each query.

We recursively repeat the above process. At each time step t, $\hat{Q}^t$ is decoded for the dynamic offset for the next frame while undergoing dynamic spatial refinement. q^t is synchronously decoded for planning.

In contrast to grid-based methods that formulate occupancy forecasting as recursive per-voxel classification, we reformulate the problem as a regression task to better capture the continuous evolution of both the ego vehicle and the surroundings. This paradigm shift enables smoother and more coherent spatiotemporal modeling. We will empirically demonstrate that SparseWorld effectively mitigates feature misalignment and temporal confusion commonly observed in grid-based methods.

Timestamp-Aware Self-Scheduling Training

To equip the model with extended-range perception capability, we supervise the training of RAP using a mixture of ground truths from multiple frames. Specifically, we sparsify the occupancy ground truth $\{\hat{\mathcal{G}}^t\}_{t=0}^f$ of the next f frames to extract the 3D occupied voxel coordinates as point clouds. These point clouds are then united into the current timestamp through coordinate transformation. The merged point clouds are re-voxelized to obtain the ground-truth $\hat{\mathcal{G}}$.

The 3D coordinates P_0' of queries can be learned via Chamfer distance to $\hat{\mathcal{G}}$:

$$\text{CD}(P_0', \mathcal{G}) = \sum_{p \in P_0'} \min_{g \in \hat{\mathcal{G}}} \|p - g\|_2^2 + \sum_{g \in \hat{\mathcal{G}}} \min_{p \in P_0'} \|p - g\|_2^2, \tag{8}$$

However, the timestamps of initial queries are discrete and can not be directly learned. There are two straightforward solutions: (1) Manually assigning timestamps to queries and applying explicit supervision for each frame. However, this approach fails to provide effective supervision signals, due to the inherent spatial overlap between adjacent frames. (2) Ignoring temporal distinctions and using all queries to forecast all future frames. Yet this leads to convergence conflicts during training and ultimately degrades performance.

To address this challenge, we craft a Timestamp-Aware Self-Scheduling Training strategy. Specifically, we first pretrain RAP without explicitly assigning query timestamps while temporarily removing the temporal component in Eq. 4. The loss during pretraining is defined as:

$$\mathcal{L}_{pretrain} = \text{CD}(P_0', \hat{\mathcal{G}}) + \sum_{l=1}^{L} (\text{CD}(P_l, \hat{\mathcal{G}}) + \mathcal{L}_{\text{focal}}(C_l, C_g)), \tag{9}$$

where C_g denotes semantic labels, P_l denotes point set output by the l-th decoder. $\mathcal{L}_{\text{focal}}$ denotes the Focal Loss.

We construct an statistical matrix $M \in \mathbb{R}^{N \times (f+1)}$ that records the count of output points from each of N queries corresponding to each timestamp over the entire dataset.

The ground truth timestamps are generated alongside $\hat{\mathcal{G}}$. As described in Eq. 8, if a predicted point p is matched to a ground-truth point g at time step t, the statistical counter of the source query of p corresponding to t is incremented by 1. Note that during the re-voxelization process, a single voxel g may correspond to multiple timestamps.

Based on M, we further design a max-proportion prioritized assigning algorithm to selectively assign query timestamps, with details in the Appendix.

After 6 epochs of pretraining, the 3D positions and timestamps of initial queries stabilize. We then perform end-to-end training, during which the statistical matrix M and query timestamps are updated dynamically each epoch. We employ the Chamfer distance and L2 loss to supervise the forecasted occupancy and trajectories, respectively.

The total loss during end-to-end training is:

$$\begin{aligned} \mathcal{L}_{\text{total}} = & \mathcal{L}_{pretrain} + \sum_{t=1}^{f} (\lambda_1 \text{CD}(P^t, \hat{\mathcal{G}}^t) \\ & + \lambda_2 \mathcal{L}_{\text{focal}}(C^t, \hat{C}^t) + \lambda_3 \mathcal{L}_2(w^t, \hat{w}^t)) \end{aligned} \tag{10}$$

Method	Aux. Sup.	mIoU ↑				IoU ↑				FPS↑
		1s	2s	3s	Avg.	1s	2s	3s	Avg.	
OccWorld-T	Semantic LiDAR	4.68	3.36	2.63	3.56	9.32	8.23	7.47	8.34	-
OccWorld-D	3D Occ	11.55	8.66	6.98	8.66	18.90	16.26	14.43	16.53	-
OccLLaMA-F	3D Occ	10.34	8.66	6.98	8.66	**25.81**	**23.19**	19.97	**22.99**	-
PreWorld	3D Occ	11.69	8.72	6.77	9.06	23.01	20.79	18.84	20.88	1.0
+Pre-training	2D & 3D Occ	12.27	9.24	7.15	9.55	23.62	21.76	19.63	21.62	1.0
SparseWorld (Ours)	3D Occ	**14.93**	**13.15**	**11.51**	**13.20**	22.96	22.10	**21.05**	22.03	**8.0**

Table 1: 4D occupancy forecasting performance of Occ3D-nuScenes dataset. The best results are highlighted **bolded**, while the second-best results are underlined. All comparative results are copied from the original papers for fairness.

Method	Aux. Sup.	L2 (m) ↓				Collision Rate (%) ↓			
		1s	2s	3s	Avg.	1s	2s	3s	Avg.
UniAD	Map& Box& Motion& Track& Occ	0.48	0.96	1.65	1.03	**0.05**	**0.17**	0.71	0.31
OccNet	Map & Box & 3D Occ	1.29	2.13	2.99	2.14	0.21	0.59	1.37	0.72
OccWorld-D	3D Occ	0.52	1.27	2.41	1.40	0.12	0.40	2.08	0.87
OccLLaMA-F †	3D Occ	0.38	1.07	2.15	1.20	0.06	0.39	1.65	0.70
PreWorld	2D & 3D Occ	0.41	1.16	2.32	1.30	0.50	0.88	2.42	1.27
PreWorld †	2D & 3D Occ	0.22	0.30	0.40	0.31	0.21	0.66	0.71	0.53
SparseWorld (Ours)	3D Occ	0.49	0.94	1.47	0.97	0.18	0.78	1.88	0.95
SparseWorld† (Ours)	3D Occ	**0.19**	**0.25**	**0.36**	**0.27**	0.11	0.29	**0.46**	**0.29**

Table 2: Motion planning performance on the Occ3D-nuScenes dataset. † indicates that ego state is applied during training and inference. The best results are in **bold**, second-best are underlined.

Here, $\hat{\mathcal{G}}^t$ and $\hat{w}^t$ denote the occupancy and trajectory ground truths at frame t, respectively.

Notably, during inference, the query timestamps remain fixed, eliminating the query assigning process and ensuring the efficiency of SparseWorld.

Experiments

Experiment Settings

Dataset and Metrics We employ the widely adopted Occ3d-nuScenes (Tian et al. 2023) benchmark, which is built upon the nuScenes dataset (Caesar et al. 2020). It provides 700 training scenes and 150 validation scenes, each lasting 20 seconds, with labels provided every 0.5 seconds. The Occ3d-nuScenes dataset offers dense labels with a resolution of 200×200×16, comprising 17 semantic categories and one free category. Each occupancy grid cell has a size of 0.4m×0.4m×0.4m. Following established methodologies, we assess occupancy perception and forecasting using Intersection over Union (IoU) and mean IoU (mIoU) as metrics. IoU evaluates overleap considering only foreground and background, whereas mIoU computes the mean IoU across all 17 classes. L2 error and collision rate are applied as indicators for ego-vehicle trajectory planning.

Implementation Details In the RAP module, We employ 6 decoding layers, where the number of output points for each query across successive layers is (1, 4, 16, 24, 32, 48). The number of queries for the 7 time steps (the current and 6 future frames) is divided as (720, 60, 60, 60, 60, 40, 40), resulting in a total of 1040 initial queries. In our implementation, we utilize ResNet-50 (He et al. 2016) with 256×704 images to extract multi-scale features.

We train our model using Temporal-Aware Self-Scheduling strategy and the AdamW optimizer, with a standard learning rate set to 2e-4. A warm-up strategy and cosine annealing mechanism are applied. The model undergoes 6 epochs of pre-training, followed by 48 epochs of end-to-end training. The entire training is carried out on 4 A100 GPUs, with a total batch size of 8. The inference speed is measured on a 4090 GPU. For further implementation details, please refer to the appendix. Notably, the visible masks are **not** utilized during both training and inference.

Main Results

Following established methodologies, we take the current and past 2 seconds of video frames as input to forecast the occupancy and ego-vehicle trajectory for the next 3 seconds.

4D Occupancy Forecasting Table 1 compares the performance of SparseWorld with other excellent occupancy world models in terms of 4D forecasting. SparseWorld demonstrates exceptional performance, achieving a 20%-40% improvement in mIoU over PreWorld (Li et al. 2025), along with a 7× increase in inference speed. Notably, SparseWorld exhibits the smallest score degradation during autoregres-

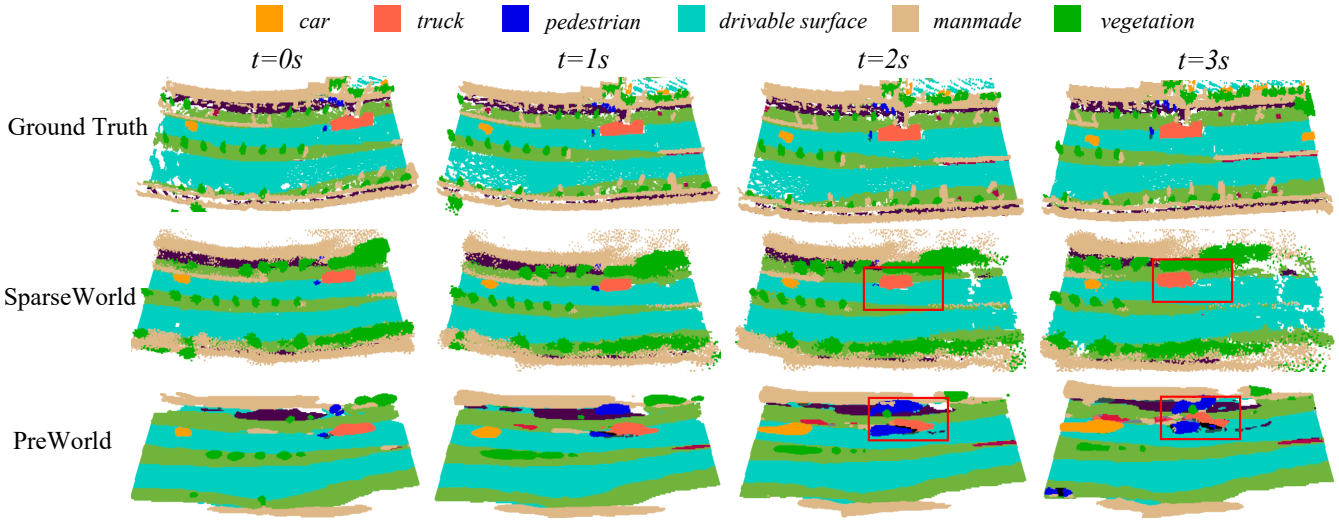


Figure 3: Visualization of the current and future 3-second ground truth and forecasts. As highlighted in red boxes, PreWorld (Li et al. 2025), which relies on grid-based modeling and voxel-level classification, exhibits severe distortions in long horizon, whereas SparseWorld effectively mitigates such issues.

sive forecasting among all models, highlighting the substantial advantages of dynamic and continuous queries in forecasting tasks. However, SparseWorld does not show a significant advantage on the IoU metric. We guess this is because IoU is dominated by background voxels (accounting for roughly 95%), most of which are not visible, whereas SparseWorld excels at recognizing foreground objects.

Our visualization of the inference results in Figure 3 further supports this conclusion. We observed that grid feature-based methods encounter feature distortion during frame-by-frame forecasting, leading to significant accumulative errors, especially in foreground categories that require particular attention. Our SparseWorld effectively avoids these issues. More visual examples can be found in the appendix.

Motion Planning Table 2 presents the motion planning results of SparseWorld. Clearly, our model also demonstrates excellent trajectory planning capabilities, particularly in terms of collision rates, where SparseWorld consistently achieves only half the collision rate of PreWorld. To ensure a fair comparison, we do not use the vehicle's state when generating the ego token, yet the model's performance remains impressive. We attribute this to the inherent nature of dynamic and continuous regression, which contributes to superior perception and 4D forecasting abilities, laying a solid foundation for safe and reasonable path planning.

Ablation Studies

In this section, we conduct detailed ablation experiments to investigate the impact of various designs on model and further validate our arguments. To expedite the verification process, we assume the use of SparseWorld, pre-training for 6 epochs and fully training the model 12 epochs, respectively.

Model Components Table 3 presents the ablation study evaluating the contribution of core modules. We report the average IoU and average mIoU within the future 3 seconds.

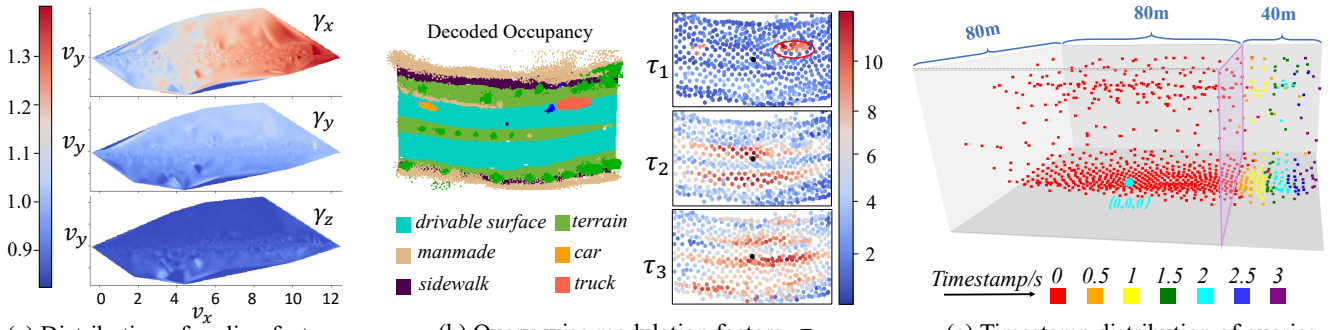
Module	Avg. mIoU	Avg. IoU
SparseWorld	**11.82**	**21.17**
w/o Adaptive Scaling	11.45 (-0.37)	20.88 (-0.29)
w/o Temporal mask	11.52 (-0.3)	20.89 (-0.28)
w/o 4D PE	11.58 (-0.24)	20.99 (-0.18)
w/o State Condition	11.31 (-0.51)	20.62 (-0.55)

Table 3: Ablation studies of core model components.

As shown in Row 3 of Table 3, the Adaptive Scaling improves mIoU by 0.37, demonstrating the significance of adaptive perception range for motion-aware scene forecasting. In Figure 4(a), we further visualize the heatmaps of learned scaling factors $\boldsymbol{\gamma}$ under different ego-motion states across the nuScenes validation set. It is observed that the longitudinal velocity v_x exhibits a larger value range than v_y and remains positive (indicating no reverse motion). The scaling factor γ_x is more sensitive to v_x, while γ_y and γ_z exhibit no significant differences across ego states. This confirms that larger perception ranges are required in the longitudinal direction, which aligns with the motion prior in autonomous driving.

As shown in row 4 of Table 3, without the temporal mask of Temporal-Spatial MHSA, SparseWorld suffered a performance loss of 0.3. This is because, although adaptive perception is necessary, the range of ground truth for each frame is fixed ($\pm$40m in Occ3D-nuScenes). Future queries can "mislead" the current query, thus disrupting the completion of the current scene. Conversely, subsequent frames can attend to the previous frame's query, allowing later frames to supplement earlier ones .

As shown in row 5 of Table 3, the 4D position encoding in SCF resulted in a performance improvement of 0.24. We believe that the introduction of spatial position allows the model to learn to correct potential perceptual errors, while



(a) Distribution of scaling factors (b) Query-wise modulation factors τ (c) Timestamp distribution of queries

Figure 4: Visualizations of intermediate reasoning representations.

Ego	Cross-Attn	L2 (m) ↓	Col (%) ↓	mIoU ↑
✓	Spatial	0.29	0.37	11.82
	Spatial	1.01	0.65	10.64
✓	Common	0.31	0.44	11.71
	Common	1.25	0.77	10.28

Table 4: Ablation studies on ego-state and spatial cross-attention, we report the average mIoU over future 3 seconds.

Training Strategy	Avg. mIoU	Avg. IoU	Time
Tem-Aware Self-Sche	11.82	21.17	12h
No Tem Different	10.95	20.08	11h
Manually Specified	11.37	20.31	22h

Table 5: Ablation results of different training strategies.

the temporal information enables the model to consciously adjust the magnitude of refinement.

As shown in row 6 of Table 3, removing the Ego State Condition leads to a substantial drop of 0.51 mIoU, highlighting the critical role of ego motion in forecasting.

Ego status and Spatial Cross-attention In Table 4, we investigate the impact of ego-state spatial cross-attention on planning. Clearly, better planning consistently promotes more accurate occupancy forecast. This is intuitive: as larger trajectory errors hinder the model from capturing scene shifts caused by ego-motion. Moreover, compared to common cross-attention, introducing spatial modulation significantly improves planning performance, particularly in the absence of ego-state input.

To better understand this, we visualize heatmaps of several heads of τ values within a given scene in Figure 4(b), where higher values indicate stronger attention from the ego agent. We observe that τ_1 assigns higher weights to foreground queries, while τ_2 focuses more on queries corresponding to drivable surfaces. This suggests that adaptive modulation enables the model to integrate both foreground and background cues when making planning decisions.

Temporal-Aware Self-Scheduling Training Training SparseWorld, a multi-stage and multi-output model, is non-trivial. In Table 5, we compare different training strategies to investigate the effect of Temporal-Aware Self-Scheduling. Without temporal differentiation, where all queries are supervised by multiple frames of GT, the training suffers from convergence instability, leading to a catastrophic drop of 0.87 mIoU. When timestamps are manually assigned, each perception decoder must learn positions under f-time supervisions, requiring $L \times f$ Chamfer distance computations, which is computationally inefficient and yields suboptimal results. In contrast, our Temporal-Aware Self-Scheduling only requires $L + f$ Chamfer distance computations. And through pretraining, the model learns the temporal distribution of queries on its own, enabling more efficient and smoother convergence.

In Figure 4(c), we visualize the learned queries with timestamp distributions, and distinguish them by color. It obviously that the queries exhibit a clear hierarchical structure along the longitudinal direction, which benefits long-range perception and facilitates cross-temporal forecasting.

Conclusion

In this paper, we proposed SparseWorld, a fully sparse, flexible, adaptive, and efficient 4D occupancy world model. Guided by ego state, SparseWorld achieves adaptive-range perception and ego-conditioned forecasting through bidirectional query interaction, aligning with the continuous dynamics of 4D scenes. A dedicated Temporal-Aware Self-Scheduling strategy further ensures stable training.

Extensive experiments demonstrate SparseWorld's superior forecasting and planning capabilities, while ablation studies and visualizations confirm its interpretability.

Limitations and Future work

Despite its strong performance, SparseWorld's geometric reasoning (IoU) and generalization to unseen scenarios remain areas for improvement. Notably, it requires only LiDAR and 2D labels without dense occupancy annotations. Future work will explore weakly supervised training and integrate large language models to enhance reasoning in complex scenes. Overall, SparseWorld demonstrates that a small number of sparse queries can effectively represent 4D scenes, offering a new perspective for autonomous driving research.

Acknowledgements

This work is funded by Xiongan AI Institute, Lenovo Research and Wuxi Research Institute of Applied Technologies, Tsinghua University under Grant 20242001120. We sincerely appreciate their supports and contributions.

References

Boeder, S.; Gigengack, F.; and Risse, B. 2025. GaussianFlowOcc: Sparse and Weakly Supervised Occupancy Estimation using Gaussian Splatting and Temporal Flow. *arXiv preprint arXiv:2502.17288*.

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

Chen, S.; Jiang, B.; Gao, H.; Liao, B.; Xu, Q.; Zhang, Q.; Huang, C.; Liu, W.; and Wang, X. 2024. Vadv2: End-to-end vectorized autonomous driving via probabilistic planning. *arXiv preprint arXiv:2402.13243*.

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

Gao, S.; Yang, J.; Chen, L.; Chitta, K.; Qiu, Y.; Geiger, A.; Zhang, J.; and Li, H. 2024. Vista: A generalizable driving world model with high fidelity and versatile controllability. *Advances in Neural Information Processing Systems*, 37: 91560–91596.

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

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

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

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

Hu, Y.; Yang, J.; Chen, L.; Li, K.; Sima, C.; Zhu, X.; Chai, S.; Du, S.; Lin, T.; Wang, W.; et al. 2023. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 17853–17862.

Huang, J.; and Huang, G. 2022. Bevdet4d: Exploit temporal cues in multi-camera 3d object detection. *arXiv preprint arXiv:2203.17054*.

Huang, Y.; Zheng, W.; Zhang, B.; Zhou, J.; and Lu, J. 2024. Selfocc: Self-supervised vision-based 3d occupancy prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19946–19956.

Jia, X.; You, J.; Zhang, Z.; and Yan, J. 2025. Drivetransformer: Unified transformer for scalable end-to-end autonomous driving. *arXiv preprint arXiv:2503.07656*.

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

Kim, J.; Kang, C.; Lee, D.; Choi, S.; and Choi, J. W. 2025. Protoocc: Accurate, efficient 3d occupancy prediction using dual branch encoder-prototype query decoder. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 4284–4292.

Li, X.; Li, P.; Zheng, Y.; Sun, W.; Wang, Y.; and Chen, Y. 2025. Semi-Supervised Vision-Centric 3D Occupancy World Model for Autonomous Driving. *arXiv preprint arXiv:2502.07309*.

Li, Y.; Yu, Z.; Choy, C.; and et al. 2023a. VoxFormer: Sparse Voxel Transformer for Camera-based 3D Semantic Scene Completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9087–9098.

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

Liao, B.; Chen, S.; Yin, H.; Jiang, B.; Wang, C.; Yan, S.; Zhang, X.; Li, X.; Zhang, Y.; Zhang, Q.; et al. 2025. Diffusiondrive: Truncated diffusion model for end-to-end autonomous driving. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 12037–12047.

Liu, H.; Chen, Y.; Wang, H.; Yang, Z.; Li, T.; Zeng, J.; Chen, L.; Li, H.; and Wang, L. 2024. Fully sparse 3d occupancy prediction. In *European Conference on Computer Vision*, 54–71. Springer.

Liu, H.; Teng, Y.; Lu, T.; Wang, H.; and Wang, L. 2023. SparseBEV: High-Performance Sparse 3D Object Detection from Multi-Camera Videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 18580–18590.

Ma, Q.; Tan, X.; Qu, Y.; Ma, L.; Zhang, Z.; and Xie, Y. 2024. Cotr: Compact occupancy transformer for vision-based 3d occupancy prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19936–19945.

Pan, M.; Liu, J.; Zhang, R.; Huang, P.; Li, X.; Xie, H.; Wang, B.; Liu, L.; and Zhang, S. 2024. Renderocc: Vision-centric 3d occupancy prediction with 2d rendering supervision. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 12404–12411. IEEE.

Shi, Y.; Cheng, T.; Zhang, Q.; Liu, W.; and Wang, X. 2024. Occupancy as set of points. In *European Conference on Computer Vision*, 72–87. Springer.

Sun, Q.; Shu, C.; Zhou, S.; Yu, Z.; Chen, Y.; Yang, D.; and Chun, Y. 2024a. Gsrender: Deduplicated occupancy prediction via weakly supervised 3d gaussian splatting. *arXiv preprint arXiv:2412.14579*.

Sun, W.; Lin, X.; Shi, Y.; Zhang, C.; Wu, H.; and Zheng, S. 2024b. Sparsedrive: End-to-end autonomous driving via sparse scene representation. *arXiv preprint arXiv:2405.19620*.

Tian, X.; Jiang, T.; Yun, L.; Mao, Y.; Yang, H.; Wang, Y.; Wang, Y.; and Zhao, H. 2023. Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving. *Advances in Neural Information Processing Systems*, 36: 64318–64330.

Wang, J.; Liu, Z.; Meng, Q.; Yan, L.; Wang, K.; Yang, J.; Liu, W.; Hou, Q.; and Cheng, M.-M. 2024a. Opus: occupancy prediction using a sparse set. *arXiv preprint arXiv:2409.09350*.

Wang, Y.; Chen, Y.; Liao, X.; Fan, L.; and Zhang, Z. 2024b. Panoocc: Unified occupancy representation for camera-based 3d panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 17158–17168.

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

Wu, Y.; Yan, Z.; Wang, Z.; Li, X.; Hui, L.; and Yang, J. 2024. Deep height decoupling for precise vision-based 3d occupancy prediction. *arXiv preprint arXiv:2409.07972*.

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

Yang, Y.; Mei, J.; Ma, Y.; Du, S.; Chen, W.; Qian, Y.; Feng, Y.; and Liu, Y. 2025. Driving in the occupancy world: Vision-centric 4d occupancy forecasting and planning via world models for autonomous driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 9327–9335.

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

Ye, Z.; Jiang, T.; Xu, C.; Li, Y.; and Zhao, H. 2024. Cvt-occ: Cost volume temporal fusion for 3d occupancy prediction. In *European Conference on Computer Vision*, 381–397. Springer.

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

Zhang, D.; Wang, G.; Zhu, R.; Zhao, J.; Chen, X.; Zhang, S.; Gong, J.; Zhou, Q.; Zhang, W.; Wang, N.; et al. 2024a. Sparsead: Sparse query-centric paradigm for efficient end-to-end autonomous driving. *arXiv preprint arXiv:2404.06892*.

Zhang, J.; Zhang, Y.; Liu, Q.; and Wang, Y. 2024b. Lightweight Spatial Embedding for Vision-based 3D Occupancy Prediction. *arXiv preprint arXiv:2412.05976*.

Zheng, W.; Chen, W.; Huang, Y.; Zhang, B.; Duan, Y.; and Lu, J. 2024. Occworld: Learning a 3d occupancy world model for autonomous driving. In *European conference on computer vision*, 55–72. Springer.

Zheng, Y.; Liang, R.; Zheng, K.; Zheng, J.; Mao, L.; Li, J.; Gu, W.; Ai, R.; Li, S. E.; Zhan, X.; et al. 2025. Diffusion-based planning for autonomous driving with flexible guidance. *arXiv preprint arXiv:2501.15564*.