2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Vision-Centric 4D Occupancy Forecasting and Planning via Implicit Residual World Models

Jianbiao Mei*, Yu Yang*, Xuemeng Yang†, Licheng Wen, Jiajun Lv†, Botian Shi, Yong Liu

***Abstract*— End-to-end autonomous driving systems increasingly rely on vision-centric world models to understand and predict their environment. However, a common ineffectiveness in these models is the full reconstruction of future scenes, which expends significant capacity on redundantly modeling static backgrounds. To address this, we propose IR-WM, an Implicit Residual World Model that focuses on modeling the current state and evolution of the world. IR-WM first establishes a robust bird's-eye-view representation of the current state from the visual observation. It then leverages the BEV features from the previous timestep as a strong temporal prior and predicts only the "residual", i.e., the changes conditioned on the ego-vehicle's actions and scene context. To alleviate error accumulation over time, we further apply an alignment module to calibrate semantic and dynamic misalignments. Moreover, we investigate different forecasting–planning coupling schemes and demonstrate that the implicit future state generated by world models substantially improves planning accuracy. On the nuScenes benchmark, IR-WM achieves top performance in both 4D occupancy forecasting and trajectory planning. Codes are available at https://github.com/yuyang-cloud/Drive-OccWorld**

I. Introduction

End-to-end autonomous driving models, which directly map raw sensor data to planning outputs, have become a dominant paradigm [1], [2]. At its core, safe decision-making in autonomous driving hinges on the precise understanding of current and past status, along with the reliable anticipation of future events. Consequently, action-driven world models have emerged as a promising approach that bridges agent intentions and environmental dynamics through action-guided world generation and forecast-driven planning, thereby enabling more reliable decision-making.

However, existing vision-centric world models for end-to-end planning often cast forecasting as full-scene reconstruction. Whether they are implicit models that produce low-dimension latents [3], [4] or explicit models that predict high-dimension representations like videos and 4D occupancy [5], [6], they tend to regenerate the entire scene from scratch for both current and future states, wasting capacity on largely static backgrounds and constraining scene context encoding from past and current visual observation. This motivates a shift from regenerating the world to modeling the current state and its future changes, as shown in Fig. 1, better allocating model capacity.

Therefore, we introduce the Implicit Residual World Model, **IR-WM**, guided by auxiliary vision-centric 4D occupancy forecasting to enable action-guided world generation and forecast-driven planning. IR-WM first constructs an implicit bird's-eye-view (BEV) representation of the current state from raw images. Instead of reconstructing full futures, it treats the previous BEV as a spatiotemporal prior and predicts only residual changes conditioned on ego trajectories. This residual design yields: (1) modeling efficiency by focusing capacity on scene context encoding from visual observation and salient context changes, e.g., dynamics and background changes, for forecasting; (2) improved dynamics understanding by emphasizing temporal differences; and (3) temporal coherence by grounding each prediction in the preceding state. We recover future states by adding residuals to the prior and applying an alignment module to correct semantic and dynamic misalignments, alleviating potential error accumulation during rollout. All BEV representations are supervised with semantic occupancy grids, enhancing semantic richness and geometric fidelity for reliable planning.

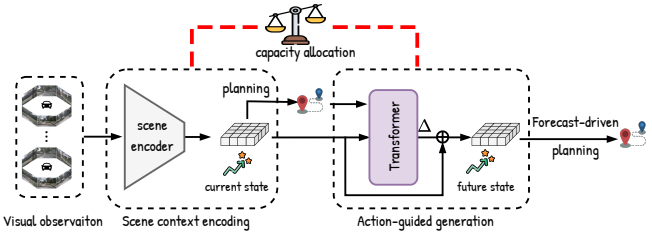

Fig. 1. We model the current state and predict residual future changes, yielding a more effective allocation of model capacity.

Moreover, we design different forecasting–planning coupling schemes to study how forecasting influences planning. Our insights are: (1) using occupancy forecasting to filter candidate trajectories is largely unnecessary, offering only marginal planning gains while adding latency; (2) implicit future state generated by world models substantially improves planning accuracy; and (3) action-conditioned world models can serve as effective observers, providing strong supervisory signals for trajectory generation.

We evaluate IR-WM on nuScenes [7] for vision-centric 4D occupancy and trajectory planning. Under the OpenOccupancy benchmark [8], IR-WM achieves consistent gains across 4D occupancy tasks and significantly reduces L2 error and collision rate relative to the baseline.

Our main contributions can be summarized as follows:

- We introduce IR-WM, a world model that models the current world state and its dynamics in a compact BEV space, providing a strong basis for vision-centric 4D occupancy forecasting and trajectory planning.
- We investigate the impact of forecasting on planning,

* Equal contributors. † Corresponding authors.
Jianbiao Mei, Yu Yang, Jiajun Lv, and Yong Liu are with Zhejiang University. Xuemeng Yang, Licheng Wen, and Botian Shi are with Shanghai AI Laboratory.

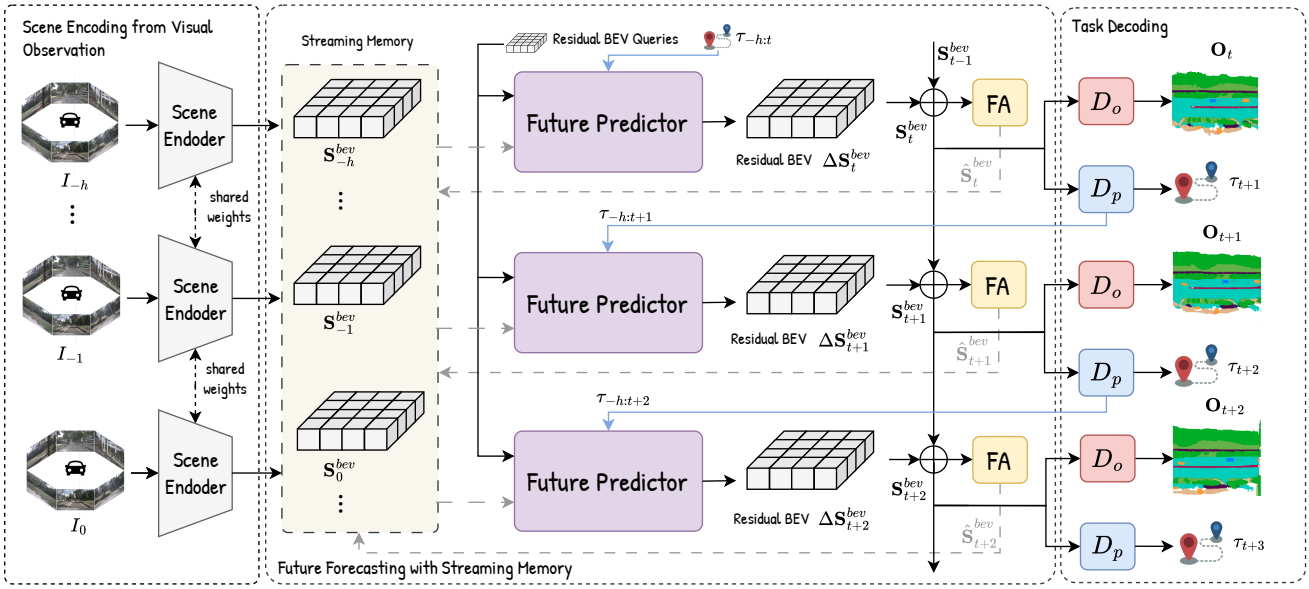

Fig. 2. Overview of IR-WM. IR-WM comprises three parts: (a) Scene encoding from visual observations via a scene encoder; (b) future forecasting, performed in an implicit residual prediction manner with a streaming memory using an autoregressive future predictor; (c) Task decoding via occupancy head D_o and planning head D_p. "FA" denotes the feature alignment for alleviating error accumulation.

and exploit a semi-coupled design separating 4D occupancy and planning to enhance flexibility while maintaining planning accuracy.

- Extensive experiments demonstrate the effectiveness of IR-WM. It achieves state-of-the-art performance in both 4D occupancy forecasting and trajectory planning.

II. Related Works

A. Driving World Models

World models aim to simulate future environment states given past observations and actions. Image-based approaches [9], [10], [11], [12], [13], [14], [15] generate future driving videos using diffusion or transformer architectures, often conditioned on actions, BEV layouts, or HD maps to enhance realism. In contrast, volume-based methods [16], [17], [18], [5], [19], [20] directly model 3D structures such as occupancy or point clouds, enabling spatially consistent future prediction. Unlike recent approaches, we construct a residual BEV world model with feature alignment, guided by 4D occupancy forecasting, which balances model capacity to effectively capture both the current state and its future changes. This design alleviates error accumulation, leverages vision-centric state modeling as a foundation for accurate future prediction, and strengthens downstream planning.

B. End-to-end Driving

End-to-end models have emerged as a paradigm shift in autonomous driving, directly mapping sensor inputs to future trajectories [1], [2]. Early approaches such as P3 [21] and ST-P3 [22] incorporated differentiable occupancy to ensure safe planning. More recent frameworks leverage BEV perception as an intermediate representation, including UniAD [1], VAD [2], and GraphAD [23]. Further advances focus on disentangled scene representations [24], reducing annotation costs [25], [3], or improving efficiency through parallel or sparse architectures [26], [27]. Our work explores the impact of implicit representation generated by world models on end-to-end planning.

C. Occupancy Prediction and Forecasting

Occupancy prediction reconstructs the 3D state of the environment. LiDAR-based methods [28], [29], [30], [31] complete sparse point clouds into dense voxel grids, while camera-based approaches [32], [33], [8], [34] transform 2D features into 3D space, often enhanced with depth estimation [35] or multi-view aggregation [36]. Forecasting extends this task to the temporal domain, with LiDAR-driven approaches that predict future points or occupancy volumes [37], [38], and 4D occupancy forecasting of Cam4DOcc [39] and DriveOccWorld [6]. This work leverages occupancy forecasting as a fine-grained regularization to learn compact world state representations and to generate state-conditioned trajectories.

III. Method

A. Preliminary

Formally, letting $t = 0$ denote the current timestamp, the vision-centric end-to-end planner takes the past and current camera observations $\mathbf{I}_{-h:0}$ and ego trajectories $\tau_{-h:0} \in \mathbb{R}^{(h+1)\times 2}$ over past h and current timestamps, and predicts the ego trajectories $\tau_{1:f} \in \mathbb{R}^{f\times 2}$ over the next f steps:

$$\tau_{1:f} = \mathscr{F}(I_{-h:0}, \tau_{-h:0}) \tag{1}$$

End-to-end planning with world models typically decomposes this procedure into three stages: compact scene context encoding from camera observations, action-conditioned future state rollout, and downstream perception and planning:

$$\mathbf{S}_{-h:0} = \mathscr{E}(\mathbf{I}_{-h:0}, \tau_{-h:0}) \tag{2}$$

$$\mathbf{S}_t = \mathscr{W}(\mathbf{S}_{-h:t-1}, \tau_{-h:t}) \tag{3}$$

$$\tau_{t+1} = \mathscr{D}_p(\mathbf{S}_t) \tag{4}$$

where $\mathbf{S}$ denotes the compact state representations (e.g., BEV features). $\mathscr{E}$ is the scene encoder, $\mathscr{W}$ is the autoregressive future predictor, and $\mathscr{D}_*$ is the task decoder, such as the planning head. In this work, we also append an occupancy head to the model that (i) provides fine-grained perception outputs for interpretability and (ii) regularizes the latent state representations, thereby enhancing the semantic richness and geometric fidelity of both current and future states for reliable planning. The occupancy prediction for timestamp t is formulated as:

$$\mathbf{O}_t = \mathscr{D}_o(\mathbf{S}_t) \tag{5}$$

Accurate rollout hinges on informative encodings of both history and the current scene. Yet many methods rebuild the entire scene at each step, overloading $\mathscr{W}$ with static background prediction and limiting $\mathscr{E}$'s capacity for contextual encoding. We therefore propose IR-WM, which explicitly models the current world state and its residual evolution, providing a strong basis for 4D occupancy forecasting and trajectory planning.

B. Implicit Residual World Model

As depicted in Figure 2, IR-WM comprises three parts: **(1)** Scene encoding from visual observations via a scene encoder $\mathscr{E}$. **(2)** Future forecasting with streaming memory via autoregressive future predictor $\mathscr{W}$. **(3)** Task decoding via occupancy head $\mathscr{D}_o$ and planning head $\mathscr{D}_p$.

1) Scene Encoding from Visual Observation: Following previous works [40], [5], we utilize the BEVFormer [41] as our scene encoder $\mathscr{E}$ to take historical and current camera images as input, extract multi-view scene context, and transform them into BEV features $\mathbf{S}^{bev}_{-h:0}$, which are push into the streaming memory for future forecasting.

2) Future Forecasting with Streaming Memory: We build an autoregressive predictor $\mathscr{W}$ that models context changes from streaming memory and ego trajectories, and use an alignment module to correct semantic and dynamic misalignments in future BEV features. $\mathscr{W}$ consists of a stack of transformer layers and takes learnable BEV queries as input. At timestamp t, each layer performs deformable self-attention, deformable temporal cross-attention to memory $\mathbf{S}^{bev}_{t-m:t-1}$ (with m memory frames), conditioning via ego-trajectory embeddings $\tau_{t-m:t}$, and an FFN to output the residual $\Delta\mathbf{S}^{bev}_t$ w.r.t. the previous BEV. For efficiency, temporal cross-attention uses deformable sampling; reference points are computed from ego-motion transforms $\delta\tau_t$ derived from $\tau_{t-1:t}$ to pre-align features across time before aggregation.

Feature Alignment. To obtain the final BEV features $\mathbf{S}^{bev}_t$, we add the predicted feature differences $\Delta\mathbf{S}^{bev}_t$ with previous BEV features $\mathbf{S}^{bev}_{t-1}$. Then we apply a feature alignment module further to enhance both the semantic richness and dynamic understanding, providing aligned features for the next rollout to alleviate potential error accumulation. Inspired by AdaNorm [42] and [6], instead of just adding the information, we modulate the features by dynamically generating the scale and shift parameters (γ, β) of layer normalization.

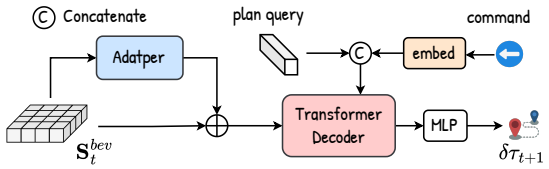

Fig. 3. The detailed architecture of the planning head.

The scale and shift parameters (γ_s, β_s) for semantics and (γ_e, β_e) for dynamics are generated by separate MLPs from the predicted semantic occupancy maps $\mathbf{O}^{align}_t$ and ego-motion transformations $\delta\tau_t$, respectively. The procedure can be formulated as:

$$\mathbf{S}^{bev}_t = \Delta\mathbf{S}^{bev}_t + \mathbf{S}^{bev}_{t-1} \tag{6}$$

$$\hat{\mathbf{S}}^{bev}_t = \gamma_s(\mathbf{O}^{align}_t) \cdot \mathrm{LN}(\mathbf{S}^{bev}_t) + \beta_s(\mathbf{O}^{align}_t) + \tag{7}$$

$$\gamma_e(\delta\tau_t) \cdot \mathrm{LN}(\mathbf{S}^{bev}_t) + \beta_e(\delta\tau_t) \tag{8}$$

where LN denotes the layer normalization.

We append a lightweight head consisting of MLPs to BEV features $\mathbf{S}^{bev}_t$ to generate a semantic occupancy map $\mathbf{O}^{align}_t$, which is supervised by cross-entropy loss $\mathscr{L}_{align}$. After obtaining aligned BEV features $\hat{\mathbf{S}}^{bev}_t$, we push them into the streaming memory queue for the next rollout.

3) 4D Occupancy Forecasting and Planning: After obtaining the BEV features $\mathbf{S}^{bev}_t$, we employ two parallel task-specific heads to support both perception and decision making: an *occupancy head* for 4D occupancy forecasting and a *planning head* for trajectory generation.

Occupancy Head. Following [43], [31], an MLP-based occupancy head applies a channel-to-height mapping to lift BEV features into a vertical volume, producing dense semantic occupancy $\mathbf{O}_t$, where each voxel encodes geometry and class. This explicit representation enhances interpretability and geometric fidelity. Rolling out over multiple steps yields $\mathbf{O}_{0:f-1}$ for current and future frames, capturing the scene's 4D spatiotemporal evolution.

Planning Head. As shown in Fig. 3, for planning, a learnable plan query predicts the ego translation $\delta\tau_{t+1}$, updating the pose as $\tau_{t+1} = \tau_t + \delta\tau_{t+1}$. At timestamp t, the plan query is concatenated with the command embedding and passed to a transformer decoder that performs cross-attention over $\mathbf{S}^{bev}_t + \mathrm{Adapter}(\mathbf{S}^{\mathrm{bev}}_{\mathrm{t}})$ to capture fine-grained scene context and attend to potential dynamic obstacles, followed by a feed-forward network; an MLP then outputs $\delta\tau_{t+1}$. The adapter is a stack of convolutions for feature enhancement.

4) Interaction between Forecasting and Planning: Moreover, we design several variants to examine how forecasting affects planning, as shown in Fig. 4:

Tightly Coupled: Planning relies not only on generated BEV features $\mathbf{S}^{bev}_{0:f-1}$ but also on predicted semantic occupancy $\mathbf{O}_{0:f-1}$. Following ST-P3 [22] and Drive-OccWorld [6], at timestamp t, the predicted semantic occupancy $\mathbf{O}_t$ is used to filter initial candidate trajectories τ^*_{t+1}, which are guided by high-level commands and optimized via a safety-oriented cost function. The planner then predicts the final

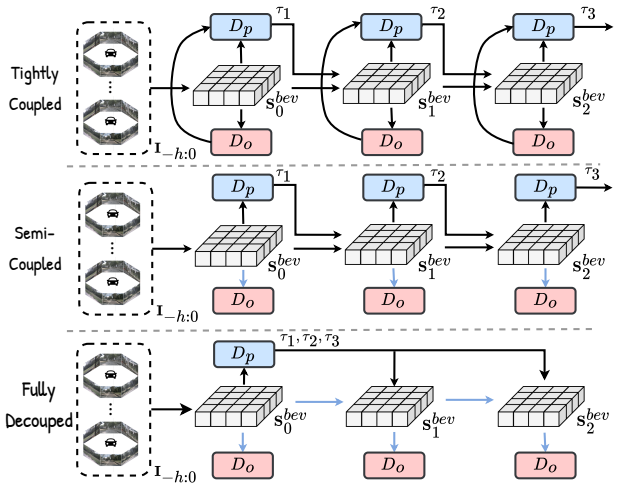

Fig. 4. Three variants illustrating how forecasting affects planning; the blue line marks components that can be removed during inference.

trajectory τ_{t+1} conditioned on the filtered candidates and the BEV features $\mathbf{S}_t^{bev}$, which can be formulated as:

$$\tau_{t+1}^* = \text{argmin}_{\tau_{\text{sample}}} \mathscr{C}(\tau_{\text{sample}}, \mathbf{O}_t, \mathbf{c}) \tag{9}$$

$$\tau_{t+1} = \mathscr{D}_p(\tau_{t+1}^*, \mathbf{S}_t^{bev}, c) \tag{10}$$

where $\mathscr{C}$ is the cost function in ST-P3 and c denotes the high-level command. τ_{sample} is the sampled initial trajectories.

Semi-Coupled: Planning only depends on the generated BEV features $\mathbf{S}_{0:f-1}^{bev}$. Occupancy head is used only as auxiliary supervision and regularization, enhancing the learned BEV feature and providing optional interpretable perception results. Our IR-WM adopts this strategy by default as suggested by experiments in Table V.

Fully Decoupled: Trajectory prediction is performed solely on the current BEV features $\mathbf{S}_0^{bev}$, with future BEV features and semantic occupancy serving only as auxiliary supervision. Under this setting, no explicit forecasting of future features is required during inference:

$$\tau_{1:f} = \mathscr{D}_p(\mathbf{S}_0^{bev}, c) \tag{11}$$

C. Training Objective

We jointly optimize occupancy forecasting and trajectory planning with a combination of perception-oriented and planning-oriented losses.

Semantic Occupancy Supervision. For occupancy prediction, we combine complementary objectives to balance semantic accuracy and geometric consistency:

$$\mathscr{L}_{occ} = \frac{1}{f}\sum_{t=0}^{f-1}\left[\mathscr{L}_{ce}(\mathbf{O}_t, \hat{\mathbf{O}}_t) + \mathscr{L}_{lovasz}(\mathbf{O}_t, \hat{\mathbf{O}}_t) + \mathscr{L}_{bce}(\mathbf{O}_t, \hat{\mathbf{O}}_t)\right], \tag{12}$$

where f is the number of future steps, $\hat{\mathbf{O}}_t$ denotes the ground-truth semantic occupancy. The cross-entropy loss supervises voxel-wise semantic classification, the Lovász loss [44] directly optimizes IoU to improve segmentation quality, and binary occupancy loss emphasizes free/occupied space to preserve geometric fidelity.

TABLE I

COMPARISONS OF **INFLATED GMO FORECASTING** ON THE NUSCENES DATASET, AND **FINE-GRAINED GMO FORECASTING** ON THE NUSCENES-OCCUPANCY DATASET.

Method	nuScenes			nuScene-Occupancy		
	IoU_c	IoU_f (2 s)	IoU_f	IoU_c	IoU_f (2 s)	IoU_f
OpenOccupancy-C [8]	12.17	11.45	11.74	10.82	8.02	8.53
SPC†	1.27	-	-	5.85	1.08	1.12
PowerBEV-3D [45]	23.08	21.25	21.86	5.91	5.25	5.49
BEVDet4D [46]	31.60	24.87	26.87	-	-	-
OCFNet (Cam4DOcc) [39]	31.30	26.82	27.98	11.45	9.68	10.10
OccProphet [47]	34.36	26.94	29.15	15.38	10.69	11.98
Drive-OccWorld [6]	39.80	36.30	37.40	13.60	12.00	12.40
IR-WM (ours)	**40.80**	**37.20**	**38.20**	**16.20**	**14.50**	**15.00**

SPC†: SurroundDepth [48] + PCPNet [38] + Cylinder3D [49]

TABLE II

COMPARISONS OF **FINE-GRAINED GMO AND GSO FORECASTING** ON NUSCENES-OCCUPANCY DATASET.

Method	IoU_c			IoU_f (2 s)			IoU_f
	GMO	GSO	mean	GMO	GSO	mean	GMO
OpenOccupancy-C [8]	9.62	17.21	13.42	7.41	17.30	12.36	7.86
SPC†	5.85	3.29	4.57	1.08	1.40	1.24	1.12
PowerBEV-3D [45]	5.91	-	-	5.25	-	-	5.49
OCFNet (Cam4DOcc) [39]	11.02	17.79	14.41	9.20	17.83	13.52	9.66
OccProphet [47]	13.71	24.42	19.06	9.34	24.56	16.95	10.33
Drive-OccWorld [6]	16.90	20.20	18.50	14.30	21.20	17.80	14.90
IR-WM (ours)	**17.50**	**23.10**	**20.30**	**15.20**	**24.50**	**19.90**	**15.70**

SPC†: SurroundDepth [48] + PCPNet [38] + Cylinder3D [49]

Planning Supervision. For trajectory planning, we employ a hybrid objective that integrates imitation learning and collision avoidance:

$$\mathscr{L}_{plan} = \frac{1}{f}\sum_{t=1}^{f}\left[|\tau_t - \hat{\tau}_t|_2^2 + \lambda_{coll}\cdot\mathscr{L}_{coll}(\hat{\tau}_t, \hat{\mathbf{O}}_t)\right], \tag{13}$$

where $|\cdot|_2^2$ is the trajectory regression loss, and $\mathscr{L}_{coll}$ [22] penalizes predicted waypoints that overlap with occupied voxels in $\hat{\mathbf{O}}_t$. The hyperparameter λ_{coll} balances fidelity of imitation learning with safety constraints.

Joint Optimization. The final training objective is a weighted sum of perception and planning losses:

$$\mathscr{L} = \mathscr{L}_{align} + \mathscr{L}_{occ} + \lambda_{plan}\cdot\mathscr{L}_{plan}, \tag{14}$$

where λ_{plan} controls the trade-off between occupancy forecasting and trajectory planning.

IV. EXPERIMENTS

A. Experimental Setups

1) Datasets: We evaluate our method on the nuScenes [7] and nuScenes-Occupancy [8] datasets , targeting both occupancy forecasting and planning. Among the 850 annotated scenes, 700 are allocated for training and the remainder for evaluation. Following the protocol in [39], [6], the model takes as input the images from two past frames and the current frame to predict future states across four timestamps for 4D occupancy forecasting. The occupancy labels cover a spatial range of $[-51.2m, 51.2m]$ along the x and y axes and $[-5m, 3m]$ along the z axis, with a voxel resolution of 0.2 m, yielding a grid size of $(512, 512, 40)$.

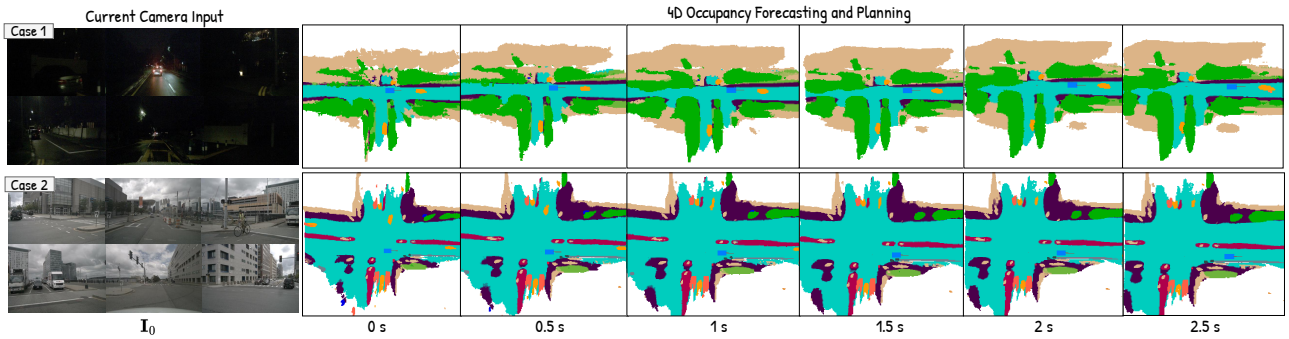

Fig. 5. **Qualitative results for 4D occupancy forecasting and planning** show that our method yields reliable predictions for static and dynamic objects at different times of day.

2) Tasks and Metrics.: We evaluate IR-WM on vison-centric 4D occupancy forecasting tasks and planning task: **(1) Inflated Occupancy Forecasting** [39]: predicting future states of general movable objects (**GMO**) represented by dilated occupancy grids derived from bounding-box annotations in nuScenes. **(2) Fine-grained Occupancy Forecasting**: using voxel-level annotations from nuScenes-Occupancy, covering both general movable objects (**GMO**) and general static objects (**GSO**). We also provide evaluation results on multiple movable objects (**MMO**) and multiple static objects (**MSO**). **(3) End-to-end Planning**: open-loop trajectory planning on nuScenes.

For occupancy forecasting, we adopt mean IoU (mIoU), reporting mIoU_c for the current frame ($t = 0$), mIoU_f for future timestamps ($t \in [1, f]$), and a time-weighted $\widetilde{\text{mIoU}}_f$ that emphasizes short-term prediction accuracy. Planning performance is measured by the L2 distance between predicted and GT trajectories as well as object collision rate.

3) Implementation Details: We take the BEVFormer-based encoder [41] as the scene encoder $\mathscr{E}$. The resolution of BEV features is set to 200×200. The future predictor $\mathscr{W}$ consists of three transformer layers with a hidden dimension of 256. Following previous work [22], [6], we set λ_{plan} and λ_{coll} to 1. We adopt an AdamW optimizer with an initial learning rate of 2e-4 and a cosine annealing scheduler to train our model on 8 NVIDIA A100 GPUs.

B. Main Results

1) Movable Objects 4D Occupancy Forecasting: Table I presents results on inflated and fine-grained GMO forecasting, focusing on dynamic objects prediction. Since dynamic agents are critical for driving safety, these results directly assess models' abilities to anticipate future motion patterns.

Inflated GMO Forecasting. As shown in the left part of Table I, IR-WM achieves the best performance with $\text{IoU}_c = 40.8$, $\text{IoU}_f = 37.2$, and $\widetilde{\text{IoU}}_f = 38.2$, surpassing the previous state-of-the-art Drive-OccWorld by about +1 point on all metrics. Compared with occupancy-only baselines like OpenOccupancy-C, our method delivers clear improvements, highlighting the benefit of richer spatiotemporal modeling. These results demonstrate that IR-WM effectively captures both the current distribution and future evolution of moving objects, which is crucial for downstream planning.

Fine-Grained GMO Forecasting. As shown in the right part of Table I, IR-WM surpasses previous methods at the voxel level, achieving $\text{IoU}_c = 16.2$ and $\text{IoU}_f = 14.5$, with notable gains over OccProphet (+0.8 and +3.8). Although voxel-level prediction is more challenging, these improvements demonstrate the robustness of our design. IR-WM captures not only coarse trajectories but also fine spatial details of dynamic agents, providing more reliable scene representations for safety-critical driving.

2) Movable & Static Objects 4D Occupancy Forecasting: Table II and III present results on fine-grained forecasting of both movable and static objects, a more comprehensive setting that requires jointly modeling dynamic agents and static structures for reliable scene understanding.

Fine-Grained GMO & GSO Forecasting. As shown in Table II, IR-WM achieves the best performance across all metrics. It attains $\text{IoU}_c = 20.3$ at the current frame, surpassing Drive-OccWorld by 1.8 points. For future predictions, it further improves to $\text{IoU}_f = 15.2$ (GMO) and 24.5 (GSO), outperforming OccProphet and Drive-OccWorld by clear margins. The weighted future score $\widetilde{\text{IoU}}_f$ also sets a new state-of-the-art, showing that IR-WM not only forecasts dynamic objects accurately but also captures fine-grained layouts of static structures.

Fine-Grained MMO & MSO Forecasting. As shown in Table III, IR-WM outperforms Drive-OccWorld across all metrics in forecasting fine-grained movable and static objects under the multi-class setting. It achieves $\text{mIoU}_c = 15.2$ and $\text{mIoU}_f = 14.4$, surpassing the baseline by +2.3 and +2.9, while the weighted score $\widetilde{\text{mIoU}}_f = 14.6$ sets a new state-of-the-art. Per-class analysis confirms broad gains across movable (e.g., Car, Truck, Bus) and static classes (e.g., Drivable surface, Sidewalk, Vegetation), with notable improvements on small and dynamic categories such as Pedestrian (+1.4) and Traffic cone (+3.2), demonstrating IR-WM captures both heterogeneous agent motion and fine-grained static context.

3) End-to-End Planning: Table IV presents results on open-loop end-to-end planning on the nuScenes validation. The results demonstrate that IR-WM achieves the best overall results, reaching an average L2 error of 0.53 m, substantially lower than Drive-OccWorld (0.85 m) and other camera-based baselines such as PARA-Drive (0.83 m) and GaussianAD (0.64 m). In particular, IR-WM reduces long-horizon error at 3s to 0.85 m, demonstrating stronger temporal

TABLE III

COMPARISONS OF **FINE-GRAINED MMO AND MSO FORECASTING** ON NUSCENES-OCCUPANCY DATASET. PER-CLASS $\text{IO}\tilde{\text{U}}_f$ IS ALSO PROVIDED.

Method	mIoU_c	mIoU_f (2.5s)	Barrier	Bicycle	Bus	Car	Construction	Motorcycle	Pedestrian	Traffic cone	Trailer	Truck	Drivable surface	Other	Sidewalk	Terrain	Manmade	Vegetation	$\text{m}\tilde{\text{IoU}}_f$
Drive-OccWorld [6]	12.9	11.5	11.5	6.9	10.2	13.2	7.7	8.3	7.6	6.5	4.8	10.5	29.1	20.1	19.2	16.1	6.4	10.8	11.8
IR-WM (ours)	15.2	14.4	14.8	9.2	12.7	15.8	9.4	9.7	9.0	9.7	8.1	13.2	32.5	23.8	21.5	19.4	10.3	14.4	14.6

TABLE IV

END-TO-END PLANNING PERFORMANCE ON NUSCENES VALIDATION. THE EGO STATUS WAS NOT UTILIZED IN THE PLANNING MODULE.

Method	Input	Aux.Sup.	L2 (m) ↓				Collision (%) ↓			
			1s	2s	3s	Avg.	1s	2s	3s	Avg.
IL [50]	LiDAR	None	0.44	1.15	2.47	1.35	0.08	0.27	1.95	0.77
NMP [51]	LiDAR	Box & Motion	0.53	1.25	2.67	1.48	0.04	0.12	0.87	0.34
FF [52]	LiDAR	Freespace	0.55	1.20	2.54	1.43	0.06	0.17	1.07	0.43
EO [37]	LiDAR	Freespace	0.67	1.36	2.78	1.60	0.04	0.09	0.88	0.33
BevFormer [41]+OccWorld [17]	Camera	3D-Occ	0.43	0.87	1.31	0.87	-	-	-	-
BevFormer [41]+Occ-LLM [53]	Camera	3D-occ	0.26	0.67	0.98	0.64	-	-	-	-
ST-P3 [22]	Camera	Map & Box & Depth	1.33	2.11	2.90	2.11	0.23	0.62	1.27	0.71
UniAD [1]	Camera	Map & Box & Motion & Tracklets & Occ	0.48	0.96	1.65	1.03	0.05	0.17	0.71	0.31
VAD-Base [2]	Camera	Map & Box & Motion	0.54	1.15	1.98	1.22	0.04	0.39	1.17	0.53
OccNet [54]	Camera	3D-Occ & Map & Box	1.29	2.13	2.99	2.14	0.21	0.59	1.37	0.72
GenAD [55]	Camera	Map & Box & Motion	0.36	0.83	1.55	0.91	0.06	0.23	1.00	0.43
UAD [25]	Camera	Box	0.39	0.81	1.50	0.90	0.01	**0.12**	0.43	0.19
RenderWorld [56]	Camera	None	0.48	1.30	2.67	1.48	0.14	0.55	2.23	0.97
SSR [3]	Camera	None	0.24	0.65	1.36	0.75	**0.00**	0.10	0.36	**0.15**
PARA-Drive [26]	Camera	Map & Box & Motion & Tracklets & Occ	0.40	0.77	1.31	0.83	0.07	0.25	0.60	0.30
GaussianAD [57]	Camera	3D-Occ & Map & Box & Motion	0.40	0.64	0.88	0.64	0.09	0.38	0.81	0.42
Drive-OccWorld [6]	Camera	4D-Occ	0.32	0.75	1.49	0.85	0.05	0.17	0.64	0.29
IR-WM (ours)	Camera	4D-Occ	**0.23**	**0.51**	**0.85**	**0.53**	0.14	0.20	**0.16**	0.17

TABLE V

ABLATIONS ON **COMBINATIONS OF FORECASTING AND PLANNING**. WE PROVIDE **FINE-GRAINED MMO AND MSO FORECASTING** RESULTS ON NUSCENES-OCCUPANCY VALIDATION.

Variants	mIoU_c	mIoU_f (2.5 s)	$\text{m}\tilde{\text{IoU}}_f$	L2 (m) ↓				Latency (ms)
				1s	2s	3s	Avg.	
Tightly Coupled	15.1	14.3	14.6	0.23	0.50	0.84	0.52	764
Semi-Coupled	15.2	14.4	14.6	0.23	0.51	0.85	0.53	706
Decoupled	14.9	14.3	14.5	0.29	0.78	1.55	0.87	558

consistency and trajectory stability. For safety, our method obtains an average collision rate of 0.17%, markedly better than prior occupancy-based planners like Drive-OccWorld (0.29%). These results highlight that IR-WM generates more informative implicit representations for downstream decision making, yielding both accurate and safe trajectories.

C. Visualizations

In Fig. 5, we present qualitative results for fine-grained occupancy forecasting and planning across consecutive frames. The visualizations show that IR-WM produces reliable predictions under varying time-of-day conditions. It also accurately captures the motion trends of dynamic objects. In particular, predictions for static objects remain highly stable over time, highlighting the effectiveness of our implicit residual world models.

D. Analysis

We conduct ablation studies to analyze the influence of the key components.

Interaction between Forecasting and Planning. We present the 4D occupancy forecasting and planning results of different variants in Table V. The forecasting performance of the three variants is largely comparable, while their planning accuracy and efficiency differ significantly. We observe that (1) using occupancy to filter candidate trajectories appears unnecessary, as it has only a marginal effect on planning performance (tightly coupled vs. semi-coupled) while bringing extra latency; (2) future BEV features generated by world models substantially improves planning accuracy (Semi-Coupled vs. Decoupled), although it inevitably increases inference latency, which highlights the effectiveness of world models in improving planning accuracy; (3) world models also provide strong supervisory signals for trajectory learning. Even in the fully decoupled setting, it still achieves performance comparable to other methods in Table IV.

Impact of Condition Interface and Feature Alignment. We evaluate the effect of action conditioning and the alignment module. The detailed results are summarized in Table VII. Using addition for conditioning achieves better performance compared to cross-attention while reducing the parameters. Moreover, as shown in Table VIII, removing the alignment module leads to a drop in IoU_f for typical static objects, indicating that proper alignment of predicted BEV

TABLE VI

ABLATIONS ON THE **TEMPORAL SELF-SUPERVISION (TSS)** ON BEV FEATURES. WE PROVIDE **FINE-GRAINED GMO AND GSO FORECASTING** RESULTS ON NUSCENES-OCCUPANCY VALIDATION.

Method	IoU_c			IoU_f (2 s)			$\widetilde{\text{IoU}}_f$
	GMO	GSO	mean	GMO	GSO	mean	GMO
IR-WM	17.50	23.10	20.30	15.20	24.50	19.90	15.70
W/ TSS	17.50	23.20	20.30	15.10	24.50	19.80	15.60

TABLE VII

ABLATIONS ON THE **ACTION CONDITIONING**. WE PROVIDE **FINE-GRAINED MMO AND MSO FORECASTING** RESULTS ON NUSCENES-OCCUPANCY VALIDATION.

CA	Additon	mIoU_c	mIoU_f (2 s)	$\text{mI}\widetilde{\text{oU}}_f$
✓		14.2	13.8	13.9
	✓	14.3	13.9	14.1

features can slightly alleviate error accumulation over time.

Exploration on Temporal Self-Supervision. We further evaluate temporal self-supervision (TSS) on BEV features. During training, the scene encoder $\mathscr{E}$ extracts BEV features $\mathbf{S}^{obs}_{1:f-1}$ from future multiview images $\mathbf{I}_{1:f-1}$, and an L2 loss is applied to supervise the predicted BEV features $\mathbf{S}^{bev}_{1:f-1}$. As shown in Table VI, TSS provides only a marginal regularization effect on occupancy prediction. This indicates that our implicit residual world model, guided by 4D occupancy, already enforces strong temporal consistency, leaving limited benefit for additional temporal constraints.

Impact of Different BEV Feature Resolutions. Finally, we compare different BEV feature resolutions. Results in Table IX showcase that our method consistently outperforms the Drive-OccWorld baseline under the same resolution, demonstrating the effectiveness of our implicit residual world model. Moreover, higher spatial resolution (200×200) provides a significant advantage over coarser features (128×128), underlining the importance of fine-grained BEV representation for accurate occupancy prediction.

V. CONCLUSION

We presented IR-WM, an end-to-end framework that shifts from reconstructing full future scenes to explicitly modeling their changes. IR-WM builds a strong BEV representation of the current state from raw visual inputs, then uses the previous timestep's BEV as a spatiotemporal prior to predict only action-conditioned residuals. We deliver that forecasting BEV features with occupancy as supervision markedly boost planning accuracy despite added latency, while using occupancy to filter candidate trajectories offers only marginal gains. At the same time, the world model provides strong supervisory signals for trajectory learning, enabling a fully decoupled variant to achieve performance comparable to recent methods with low latency.

VI. ACKNOWLEDGMENT

This research was supported by Zhejiang Provincial Natural Science Foundation of China under Grant No.LQN25F030006.

TABLE VIII

ABLATIONS ON THE **FEATURE ALIGNMENT (FA)**. WE PROVIDE PER-CLASS IoU_f RESULTS ON NUSCENES-OCCUPANCY VALIDATION.

Variants	barrier	construction	sidewalk	driveable surface
w/o FA	14.0	8.5	20.7	32.0
w/ FA	14.4 (**+0.4**)	8.9 (**+0.4**)	21.1 (**+0.4**)	32.2 (**+0.2**)

TABLE IX

ABLATIONS ON THE **BEV FEATURE RESOLUTION**. WE PROVIDE **FINE-GRAINED GMO FORECASTING** RESULTS ON NUSCENES-OCCUPANCY VALIDATION.

Method	Resolution	IoU_c	IoU_f (2 s)	$\widetilde{\text{IoU}}_f$
Drive-OccWorld	200×200	13.60	12.00	12.40
IR-WM (ours)	200×200	16.20	14.50	15.00
IR-WM (ours)	128×128	11.80	10.80	11.10

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 *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2023, pp. 17 853–17 862.

[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 *IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 8340–8350.

[3] P. Li and D. Cui, "Navigation-guided sparse scene representation for end-to-end autonomous driving," *Int. Conf. Learn. Represent*, 2024.

[4] 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.

[5] C. Min, D. Zhao, L. Xiao, J. Zhao, X. Xu, Z. Zhu, L. Jin, J. Li, Y. Guo, J. Xing *et al.*, "Driveworld: 4d pre-trained scene understanding via world models for autonomous driving," in *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2024, pp. 15 522–15 533.

[6] Y. Yang, J. Mei, Y. Ma, S. Du, W. Chen, Y. Qian, Y. Feng, and Y. Liu, "Driving in the occupancy world: Vision-centric 4d occupancy forecasting and planning via world models for autonomous driving," in *AAAI Conf. Artifi. Intell.*, vol. 39, no. 9, 2025, pp. 9327–9335.

[7] 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 *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2020, pp. 11 621–11 631.

[8] X. Wang, Z. Zhu, W. Xu, Y. Zhang, Y. Wei, X. Chi, Y. Ye, D. Du, J. Lu, and X. Wang, "Openoccupancy: A large scale benchmark for surrounding semantic occupancy perception," *arXiv preprint arXiv:2303.03991*, 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] X. Wang, Z. Zhu, G. Huang, X. Chen, and J. Lu, "Drivedreamer: Towards real-world-driven world models for autonomous driving," *arXiv preprint arXiv:2309.09777*, 2023.

[11] X. Li, Y. Zhang, and X. Ye, "Drivingdiffusion: Layout-guided multi-view driving scene video generation with latent diffusion model," *arXiv preprint arXiv:2310.07771*, 2023.

[12] J. Yang, S. Gao, Y. Qiu, L. Chen, T. Li, B. Dai, K. Chitta, P. Wu, J. Zeng, P. Luo *et al.*, "Generalized predictive model for autonomous driving," in *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2024, pp. 14 662–14 672.

[13] L. Kong, W. Yang, J. Mei, Y. Liu, A. Liang, D. Zhu, D. Lu, W. Yin, X. Hu, M. Jia *et al.*, "3d and 4d world modeling: A survey," *arXiv preprint arXiv:2509.07996*, 2025.

[14] J. Mei, T. Hu, X. Yang, L. Wen, Y. Yang, T. Wei, Y. Ma, M. Dou, B. Shi, and Y. Liu, "Dreamforge: Motion-aware autoregressive video generation for multi-view driving scenes," *arXiv preprint arXiv:2409.04003*, 2024.

[15] X. Yang, L. Wen, T. Wei, Y. Ma, J. Mei, X. Li, W. Lei, D. Fu, P. Cai, M. Dou *et al.*, "Drivearena: A closed-loop generative simulation platform for autonomous driving," in *IEEE/CVF Int. Conf. Comput. Vis.*, 2025, pp. 26 933–26 943.

[16] 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.

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

[18] 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.

[19] H. Xu *et al.*, "Temporal triplane transformers as occupancy world models," *arXiv preprint arXiv:2503.07338*, 2025.

[20] Y. Yang, A. Liang, J. Mei, Y. Ma, Y. Liu, and G. H. Lee, "X-scene: Large-scale driving scene generation with high fidelity and flexible controllability," *Adv. Neural Inf. Process. Syst.*, 2025.

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

[22] 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 *Eur. Conf. Comput. Vis.* Springer, 2022, pp. 533–549.

[23] Y. Zhang, D. Qian, D. Li, Y. Pan, Y. Chen, Z. Liang, Z. Zhang, S. Zhang, H. Li, M. Fu *et al.*, "Graphad: Interaction scene graph for end-to-end autonomous driving," *arXiv preprint arXiv:2403.19098*, 2024.

[24] S. Doll, N. Hanselmann, L. Schneider, R. Schulz, M. Cordts, M. Enzweiler, and H. Lensch, "Dualad: Disentangling the dynamic and static world for end-to-end driving," in *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2024, pp. 14 728–14 737.

[25] M. Guo, Z. Zhang, Y. He, K. Wang, and L. Jing, "End-to-end autonomous driving without costly modularization and 3d manual annotation," *arXiv preprint arXiv:2406.17680*, 2024.

[26] X. Weng, B. Ivanovic, Y. Wang, Y. Wang, and M. Pavone, "Paradrive: Parallelized architecture for real-time autonomous driving," in *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2024, pp. 15 449–15 458.

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

[28] S. Song, F. Yu, A. Zeng, A. X. Chang, M. Savva, and T. Funkhouser, "Semantic scene completion from a single depth image," in *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 1746–1754.

[29] X. Yan, J. Gao, J. Li, R. Zhang, Z. Li, R. Huang, and S. Cui, "Sparse single sweep lidar point cloud segmentation via learning contextual shape priors from scene completion," in *AAAI Conf. Artifi. Intell.*, vol. 35, no. 4, 2021, pp. 3101–3109.

[30] X. Yang, H. Zou, X. Kong, T. Huang, Y. Liu, W. Li, F. Wen, and H. Zhang, "Semantic segmentation-assisted scene completion for lidar point clouds," in *IEEE/RSJ Int. Conf. Intell. Robots Syst.* IEEE, 2021, pp. 3555–3562.

[31] J. Mei, Y. Yang, M. Wang, T. Huang, X. Yang, and Y. Liu, "Ssc-rs: Elevate lidar semantic scene completion with representation separation and bev fusion," in *IEEE/RSJ Int. Conf. Intell. Robots Syst.* IEEE, 2023, pp. 1–8.

[32] Y. Li, Z. Yu, C. Choy, C. Xiao, J. M. Alvarez, S. Fidler, C. Feng, and A. Anandkumar, "Voxformer: Sparse voxel transformer for camera-based 3d semantic scene completion," in *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2023, pp. 9087–9098.

[33] Y. Wei, L. Zhao, W. Zheng, Z. Zhu, J. Zhou, and J. Lu, "Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving," *arXiv preprint arXiv:2303.09551*, 2023.

[34] J. Mei, Y. Yang, M. Wang, J. Zhu, J. Ra, Y. Ma, L. Li, and Y. Liu, "Camera-based 3d semantic scene completion with sparse guidance network," *IEEE Trans. Image Process.*, 2024.

[35] Y. Zhang, Z. Zhu, and D. Du, "Occformer: Dual-path transformer for vision-based 3d semantic occupancy prediction," *arXiv preprint arXiv:2304.05316*, 2023.

[36] X. Tian, T. Jiang, L. Yun, Y. Wang, Y. Wang, and H. Zhao, "Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving," *arXiv preprint arXiv:2304.14365*, 2023.

[37] T. Khurana, P. Hu, A. Dave, J. Ziglar, D. Held, and D. Ramanan, "Differentiable raycasting for self-supervised occupancy forecasting," in *Eur. Conf. Comput. Vis.* Springer, 2022, pp. 353–369.

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

[39] J. Ma, X. Chen, J. Huang, J. Xu, Z. Luo, J. Xu, W. Gu, R. Ai, and H. Wang, "Cam4docc: Benchmark for camera-only 4d occupancy forecasting in autonomous driving applications," in *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2024, pp. 21 486–21 495.

[40] Z. Yang, L. Chen, Y. Sun, and H. Li, "Visual point cloud forecasting enables scalable autonomous driving," in *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2024, pp. 14 673–14 684.

[41] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Y. Qiao, and J. Dai, "Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers," in *Eur. Conf. Comput. Vis.* Springer, 2022, pp. 1–18.

[42] J. Xu, X. Sun, Z. Zhang, G. Zhao, and J. Lin, "Understanding and improving layer normalization," *Adv. Neural Inf. Process. Syst.*, vol. 32, 2019.

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

[44] M. Berman, A. Rannen 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 *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 4413–4421.

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

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

[47] J. Chen, H. Xu, Y. Wang, and L.-P. Chau, "Occprophet: Pushing efficiency frontier of camera-only 4d occupancy forecasting with observer-forecaster-refiner framework," *arXiv preprint arXiv:2502.15180*, 2025.

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

[49] X. Zhu, H. Zhou, T. Wang, F. Hong, Y. Ma, W. Li, H. Li, and D. Lin, "Cylindrical and asymmetrical 3d convolution networks for lidar segmentation," in *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2021, pp. 9939–9948.

[50] N. D. Ratliff, J. A. Bagnell, and M. A. Zinkevich, "Maximum margin planning," in *Int. Conf. Mach. Learn.*, 2006, pp. 729–736.

[51] W. Zeng, W. Luo, S. Suo, A. Sadat, B. Yang, S. Casas, and R. Urtasun, "End-to-end interpretable neural motion planner," in *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 8660–8669.

[52] P. Hu, A. Huang, J. Dolan, D. Held, and D. Ramanan, "Safe local motion planning with self-supervised freespace forecasting," in *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2021, pp. 12 732–12 741.

[53] 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.

[54] 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 *IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 8406–8415.

[55] W. Zheng, R. Song, X. Guo, C. Zhang, and L. Chen, "Genad: Generative end-to-end autonomous driving," in *Eur. Conf. Comput. Vis.* Springer, 2024, pp. 87–104.

[56] Z. Yan, W. Dong, Y. Shao, Y. Lu, H. Liu, J. Liu, H. Wang, Z. Wang, Y. Wang, F. Remondino *et al.*, "Renderworld: World model with self-supervised 3d label," in *IEEE Int. Conf. Robot. Autom.* IEEE, 2025, pp. 6063–6070.

[57] W. Zheng, J. Wu, Y. Zheng, S. Zuo, Z. Xie, L. Yang, Y. Pan, Z. Hao, P. Jia, X. Lang *et al.*, "Gaussianad: Gaussian-centric end-to-end autonomous driving," *arXiv preprint arXiv:2412.10371*, 2024.