KIO-WORLD: A KINEMATICS-INFORMED SPARSE OCCUPANCY WORLD MODEL

Anonymous authors
Paper under double-blind review

ABSTRACT

Occupancy world models reconstruct 3D scene geometry and semantics from past observations and predict their future evolution. Recent sparse models decode multiple future horizons from query anchors, but do not explicitly couple historical evidence alignment and future prediction through a shared motion state. Consequently, independently moving scene content can cause misaligned image feature sampling, while horizon-specific predictions lack a common motion reference. To address these issues, we present **KIO-World**, a kinematics-informed sparse occupancy world model that jointly models world states across horizons within a unified kinematic framework. Specifically, we equip each query anchor with a horizon-shared kinematic state that compensates for scene dynamics during historical feature sampling and provides world-state references for future query anchor evolution. To preserve spatial structure during scene evolution, we further introduce a hierarchical residual point refinement module, in which query points iteratively inherit parent geometry, and anchor-wise pose updates are decoupled from point-wise shape refinement. These mechanisms introduce complementary structural priors that support temporal reasoning and geometric refinement while retaining the flexibility of direct occupancy decoding. Extensive experiments demonstrate that KIO-World achieves state-of-the-art performance, outperforming existing camera-centric methods by 11.4% and 6.3% in average semantic mIoU and geometric IoU, respectively. Code will be released upon publication.

1 INTRODUCTION

Anticipating the evolution of 3D scenes is a crucial challenge in autonomous driving. Beyond reconstructing the current environment, an occupancy world model forecasts the future spatial layout and semantic composition of scene content (Zheng et al., 2024). Given historical camera observations and a future ego trajectory, this task requires reasoning jointly about semantic context, scene geometry, and environmental dynamics (Du et al., 2026). An effective world representation must therefore support both detailed geometric reconstruction and motion-informed forecasting.

Recent sparse occupancy models represent scenes using a compact set of learnable queries associated with explicit 3D points and latent features (Tang et al., 2024; Wang et al., 2024a; Shi et al., 2024). The locations of 3D points provide spatial support, while the query features encode semantic context and local geometric details, allowing occupancy to be recovered through point voxelization and feature decoding. Extending this representation to multiple future horizons facilitates scene evolution to be decoded in a single feed-forward pass (Dang et al., 2026; Du et al., 2026). However, this formulation still faces two closely related challenges: how to align historical image evidence with independently moving scene content, and how to keep the anchors predicted at different horizons consistent with a common motion.

A shared representation of scene motion can connect historical alignment with future prediction. Ego-pose transformations compensate for observer's movements, but do not by themselves account for independently moving scene content. During the historical feature sampling process, existing approaches (Tang et al., 2024; Wang et al., 2024a; 2026a; Du et al., 2026) project target-time query sampling locations into historical images using only ego poses, which can therefore retrieve evidence from the wrong source-time position on moving objects. Although deformable sampling and cross-frame attention enable adaptive spatiotemporal aggregation (Du et al., 2026), large dis-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

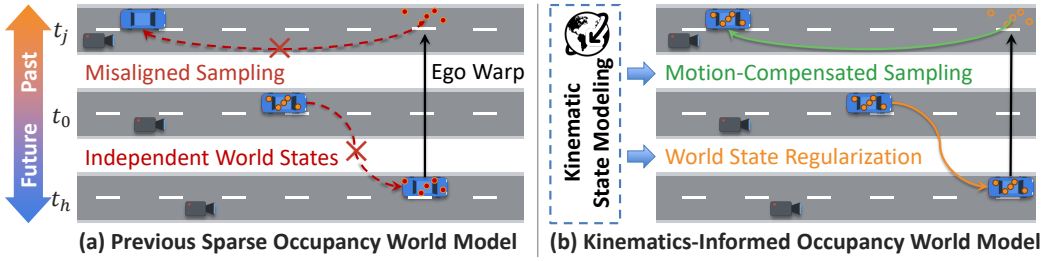

Figure 1: **A shared motion hypothesis for sparse occupancy world model.** (a) Previous methods may retrieve misaligned historical evidence for moving scene content, while horizon-specific anchors lack an explicit motion reference. (b) KIO-World uses a shared kinematic state to compensate for historical sampling locations and regularize future anchor centers.

placements over long temporal intervals can leave sampling locations far from their source-time correspondences, making reliable historical evidence hard to retrieve. Meanwhile, most sparse occupancy models (Dang et al., 2026; Du et al., 2026) directly estimate future horizons with a set of independent queries. Although temporal attention couples representations across forecasting horizons (Du et al., 2026), it does not explicitly regularize future query centers against a common motion reference. These limitations motivate a unified kinematic representation of world states that supports both motion-compensated historical alignment and explicit regularization of future anchor evolution.

Local geometric refinement poses a complementary challenge. Existing sparse occupancy models progressively expand their point sets by regressing offsets from a shared anchor centroid (Wang et al., 2024a; Du et al., 2026). During layer-wise refinement, this strategy discards the explicit local configuration of query points in the preceding layers, while newly predicted offsets must account for both anchor-level transformations and local shape variations. This entanglement makes it difficult to preserve and progressively refine the geometric structure established in earlier stages.

To address these issues, we present **KIO-World**, a kinematics-informed sparse occupancy world model that jointly estimates and refines world states under a unified kinematic framework. For temporal reasoning, we introduce a unified kinematic state, which estimates the velocity and yaw rate of each query anchor in the current ego frame and shares them across all prediction horizons. As shown in Fig. 1, this state is applied to backtrace sampling locations to their source-time positions, aligning the historical image evidence of moving objects that ego poses alone cannot compensate. It is also applied forward to advance current anchor centers to each future horizon, providing kinematic references that softly regularize the future anchor positions. For geometric refinement, we introduce a hierarchical residual point refinement module, which builds up explicit inheritance correspondences between successive refinement layers. This module preserves the preceding layer's geometry as a refinement reference while separately parameterizing anchor-level transformations and local shape variations. Extensive experiments on Occ3D-nuScenes (Tian et al., 2023) validate the effectiveness and superiority of the proposed occupancy world model. Notably, compared with the previous state-of-the-art method SparseWorld-TC (Du et al., 2026), KIO-World achieves a significant relative improvement of **11.4**% and **6.3**% on the average semantic mIoU and geometric IoU, respectively. Moreover, when compared with methods that take ground-truth occupancy as input, KIO-World achieves superior geometric forecasting performance than all of them and higher semantic accuracy than most of them, while relying only on camera observations. Our main contributions are threefold: (1) We introduce KIO-World, a sparse occupancy world model that incorporates kinematic states into world-state representation, achieving state-of-the-art performance in camera-centric 4D occupancy forecasting task. (2) We develop a unified kinematic state for bidirectional temporal reasoning, which can be utilized backward for motion-compensated historical evidence alignment, and forward for future query motion regularization. (3) We propose a hierarchical residual point refinement module, which preserves geometric priors from precedent refinement layers and decouples anchor-level transformations from point-specific shape refinement.

2 Related Work

Vision-Based Semantic Occupancy Prediction. Vision-based semantic occupancy prediction lifts images into a 3D representation of scene geometry and semantics. A common line of work orga-

nizes image evidence on voxel grids or predefined planes. Grid-based methods (Cao & de Charette, 2022; Tong et al., 2023; Zhang et al., 2023; Wei et al., 2023; Li et al., 2023b; Wang et al., 2024b) improve occupancy reasoning over structured voxel features, while other efforts (Li et al., 2022; Huang et al., 2023; Yu et al., 2023; 2024) factorize the 3D volume into bird's eye view or tri-perspective views for efficient scene representation. Nevertheless, their formulation remains coupled to spatial lattices, resulting in considerable computational redundancy. To address this, many works (Tang et al., 2024; Shi et al., 2024; Li et al., 2023a; Lu et al., 2024; Jiang et al., 2024) have proposed alternative sparse occupancy representations to achieve superior efficiency. In particular, the OPUS series (Wang et al., 2024a; 2026a) explores sparse set-based occupancy prediction, while the GaussianFormer series (Huang et al., 2024; 2025) further replaces voxels with sparse probabilistic Gaussian primitives. Building upon the sparse query-based scene representation, KIO-World propagates explicit parent geometry and decouples anchor-wise pose updates from point-wise shape refinement.

4D Occupancy Forecasting and World Models. 4D occupancy forecasting extends single-frame perception toward future scene evolution. Pioneering approaches predict future geometry through differentiable raycasting, point-cloud forecasting, or unsupervised occupancy fields (Khurana et al., 2022; 2023; Yang et al., 2024; Agro et al., 2024; Liu et al., 2024). Subsequent camera-only methods (Ma et al., 2024; Xu et al., 2025b; Chen et al., 2025) standardize vision-based occupancy forecasting, but remain coupled to task-specific representations and prediction heads. A parallel line formulates the problem as an occupancy world model. Token-based methods (Zheng et al., 2024; Min et al., 2024; Yang et al., 2025; Liao et al., 2025b; Yan et al., 2025; Li et al., 2025b) discretize scene states and forecast future occupancy autoregressively, while language-conditioned variants (Wei et al., 2024; Xu et al., 2025c) further unify occupancy with action or linguistic interfaces. Diffusion-based generators (Wang et al., 2026b; Gu et al., 2024; Shi et al., 2025; Li et al., 2025a; Jin et al., 2026; Bian et al., 2025) improve long-horizon synthesis, and other efforts (Xu et al., 2025a; Zhang et al., 2024) reduce temporal redundancy through compact scene updates. However, sequential rollout and iterative sampling still constrain efficiency and geometric flexibility. Sparse query-based world models (Dang et al., 2026; Du et al., 2026) instead decode multi-horizon occupancy in a single feed-forward pass. KIO-World follows this route and introduces an explicit kinematic state as a shared prior for historical sampling and future world state representation.

Dynamic Scene Modeling for Occupancy Forecasting. Accurate forecasting also requires historical evidence to be aligned with a changing scene. A common strategy compensates ego motion before temporal fusion (Li et al., 2022; Lin et al., 2022; Wang et al., 2023), but remains designed for perception rather than occupancy evolution. Occupancy forecasting methods therefore introduce explicit scene flow or motion transformations (Ma et al., 2024; Liao et al., 2025a). These methods capture scene dynamics, yet historical alignment and future evolution remain separately modeled. Gaussian-based approaches (Zuo et al., 2025; Chen et al., 2026) couple the two at the representation level through primitive dynamics, but still obtain occupancy by evolving prescribed Gaussians. KIO-World instead shares one kinematic anchor state across both directions, coupling historical evidence alignment and future center regularization while retaining direct occupancy decoding.

3 Methodology

3.1 Sparse Occupancy World Representation

Let $j \in \{-T, \dots, 0\}$ index the T historical frames and the current frame, and $h \in \{0, \dots, H\}$ index the current frame and H future horizons, with physical timestamps t_j and t_h in seconds. The camera parameters and ego poses are collectively denoted by $\mathcal{G}$. Given historical and current image observations $\mathcal{I}$ and the future ego trajectories $\mathcal{T} = \{\boldsymbol{T}_{h\to 0} \in \mathrm{SE}(3)\}_{h=0}^{H}$, the *Occupancy World Model* predicts the evolution of the scene occupancy as follows:

$$f(\mathcal{I}, \mathcal{G}, \mathcal{T}) \mapsto \{\mathcal{O}_h\}_{h=0}^{H}, \qquad \mathcal{O}_h = \{(\boldsymbol{u}_{h,m}, y_{h,m})\}_{m=1}^{M_h}, \tag{1}$$

where $\boldsymbol{u}_{h,m} \in \mathbb{Z}^3$ denotes one of the M_h occupied voxel indices, and $y_{h,m}$ denotes its semantic category. Inspired by recent sparse models (Dang et al., 2026; Du et al., 2026) that use a compact set of learnable queries to represent scene geometry and semantics, we formulate the world state $\mathcal{S}_h$ as a set of Q query anchors with learnable query features, point locations, and kinematic states:

$$\mathcal{S}_h = \{(\boldsymbol{k}_q, \boldsymbol{z}_{q,h}, \mathcal{P}_{q,h})\}_{q=1}^{Q}, \qquad \mathcal{P}_{q,h} = \{\boldsymbol{p}_{q,h,i} \in \mathbb{R}^3\}_{i=1}^{N}, \tag{2}$$

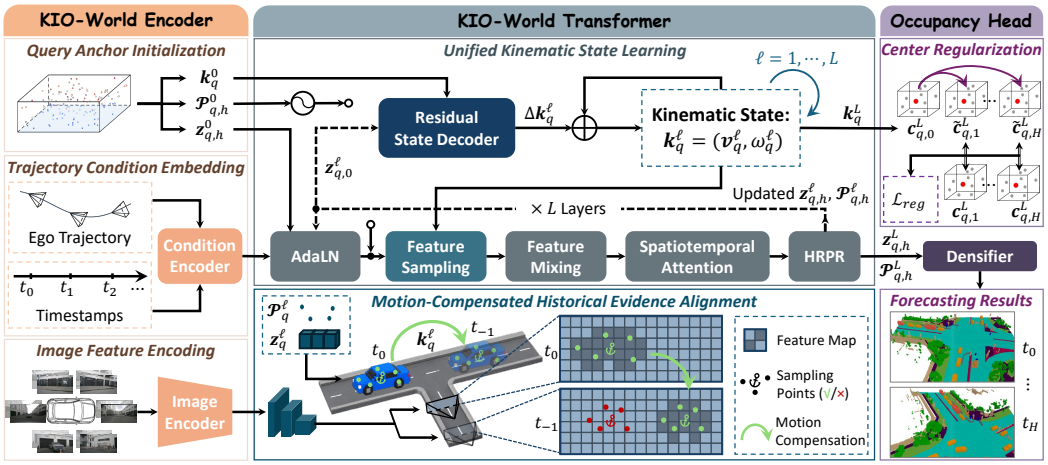

Figure 2: **Overview of KIO-World.** The encoder provides initialized query anchors, trajectory conditions, and image features. The transformer iteratively refines them with a shared kinematic state. The occupancy head finally regularizes future query centers and forecasts semantic occupancy.

where $\boldsymbol{k}_q$ denotes the kinematic state of anchor q shared by all horizons, $\boldsymbol{z}_{q,h}$ denotes the query feature at horizon h, and $\mathcal{P}_{q,h}$ denotes the query point set of N explicit 3D points. The query features and points play complementary roles in occupancy decoding: the query points provide explicit spatial evidence, while the latent features encode the associated semantic and geometric context. The unified kinematic state provides a common motion prior for bidirectional temporal reasoning, guiding the backtracing of sampling locations to align historical image evidence, and providing reference anchor locations for world state regularization.

3.2 KIO-World Overview

As illustrated in Fig. 2, KIO-World comprises three components: a KIO-World encoder, a KIO-World transformer, and an occupancy head. The encoder initializes sparse query anchors with point sets $\mathcal{P}^0_{q,h}$, query features $\boldsymbol{z}^0_{q,h}$, and zero-initialized kinematic states $\boldsymbol{k}^0_q$ shared across horizons. An image encoder extracts multi-scale image features from the multi-view observations, while a condition encoder embeds the ego trajectory and timestamps to provide horizon-specific conditioning. The transformer progressively refines the sparse query anchors through L decoder layers. Trajectory conditions modulate query features through adaptive layer normalization, and positional encodings incorporate the anchor geometry. A residual state decoder iteratively estimates the kinematic state $\boldsymbol{k}^\ell_q = (\boldsymbol{v}^\ell_q, \omega^\ell_q)$ from conditioned current-horizon query features, where $\boldsymbol{v}^\ell_q$ and ω^ℓ_q denote velocity and yaw rate in the current ego frame. Shared across target horizons, this state guides the backtracing of sampling locations to source timestamps for motion-compensated historical evidence alignment. The sampled image features are subsequently integrated through adaptive feature mixing and spatiotemporal attention, enabling evidence aggregation and interactions among anchors and horizons. The resulting query features drive Hierarchical Residual Point Refinement (HRPR), which progressively expands and refines the point sets by combining inherited parent geometry, anchor-level transformations, and point-specific shape residuals. The updated query features and geometry are passed to subsequent layers for further refinement. Finally, the occupancy head uses a densifier to convert the refined point sets $\mathcal{P}^L_{q,h}$ and features $\boldsymbol{z}^L_{q,h}$ into semantic occupancy predictions $\{\mathcal{O}_h\}^H_{h=0}$. During training, the final shared state $\boldsymbol{k}^L_q$ also advances current anchor centers to future timestamps, yielding kinematic references that softly constrain the directly predicted future anchor centers through $\mathcal{L}_{\text{reg}}$. This design connects historical evidence alignment and future query-center regularization through a shared motion prior, while retaining the flexibility of direct multi-horizon occupancy decoding.

3.3 Unified Kinematic State for Bidirectional Temporal Reasoning

Temporal perception models (Tang et al., 2024; Liu et al., 2023) commonly use ego poses to align historical observations, which relies on a static-world assumption without accounting for independently moving scene content. Meanwhile, existing occupancy forecasters (Dang et al., 2026; Du et al., 2026) directly estimate future occupancy at each horizon without an explicit shared motion ref-

erence for the decoded query centers. Motivated by these limitations, we propose a unified kinematic state model for bidirectional temporal reasoning. The model estimates a horizon-shared kinematic state for each anchor at the current time, and applies it backward for historical evidence alignment and forward for future world state regularization, providing motion-compensated feature sampling locations and cross-horizon motion references within the sparse query representation.

Horizon-Shared Kinematic State Modeling. Most movements in driving scenes are dominated by locally smooth translation and planar heading variation, which can be effectively approximated by a constant-velocity-with-yaw (CV-Yaw) model. However, estimating an independent motion state from each horizon can yield mutually incompatible kinematic hypotheses, which may project the same query slot to different locations in the historical images. We therefore estimate one unified kinematic state for each query anchor at the current timestamp, and share it with queries of the same index across all horizons. The query index identifies corresponding query slots, while their features and point support remain horizon-specific. The resulting correspondence provides a basis for center regularization without imposing object-instance identity or rigid shape preservation. For each query q at decoder layer ℓ, we define its kinematic state as $\boldsymbol{k}_q^\ell = (\boldsymbol{v}_q^\ell, \omega_q^\ell)$, where $\boldsymbol{v}_q^\ell \in \mathbb{R}^3$ is the anchor velocity and ω_q^ℓ is the yaw rate. Using the query feature $\boldsymbol{z}_{q,0}^\ell$ of the current horizon, the kinematic state is estimated and iteratively refined across layers as follows:

$$\widetilde{\boldsymbol{k}}_q^\ell = \widetilde{\boldsymbol{k}}_q^{\ell-1} + f_\theta^\ell(\boldsymbol{z}_{q,0}^\ell), \qquad \widetilde{\boldsymbol{k}}_q^0 = \mathbf{0}, \qquad \boldsymbol{k}_q^\ell = \psi(\widetilde{\boldsymbol{k}}_q^\ell), \qquad \ell = 1, \dots, L, \tag{3}$$

where f_θ^ℓ predicts layer-wise kinematics residuals, and ψ is a channel-wise bounded mapping that converts the accumulated state into metric CV-Yaw quantities. The kinematics logits $\widetilde{\boldsymbol{k}}_q^\ell$ are zero-initialized, so that motion-aware sampling initially reduces to rigid alignment. The estimated $\boldsymbol{k}_q^\ell$ is subsequently used to align the historical evidence during image feature sampling, and the final state $\boldsymbol{k}_q^L$ constructs future world state references for motion regularization across all future horizons.

Motion-Compensated Historical Evidence Alignment. Existing approaches (Dang et al., 2026; Du et al., 2026) use ego poses only to project anchor locations into historical images, which neglects the sampling misalignment induced by moving objects. We instead use the estimated kinematic state of each query anchor to backtrace every sampling point to its source-time location before feature sampling. Let $\boldsymbol{x}_{q,h}^\ell$ be a sampling point generated by query anchor q of the ℓ-th decoder layer in horizon h, $\boldsymbol{c}_{q,h}^\ell$ be the query anchor center, and $\Delta t_{j,h} = t_j - t_h$ be the signed time interval from target horizon h to source frame j. Using the future trajectory $\boldsymbol{T}_{h\to 0}$ and the kinematic state $\boldsymbol{k}_q^\ell = (\boldsymbol{v}_q^\ell, \omega_q^\ell)$, the motion-compensated sampling location $\widetilde{\boldsymbol{x}}_{q,h,j}^\ell$ at source time is derived as follows:

$$\widetilde{\boldsymbol{x}}_{q,h,j}^\ell = \boldsymbol{T}_{h\to 0}\boldsymbol{c}_{q,h}^\ell + \boldsymbol{v}_q^\ell \Delta t_{j,h} + \boldsymbol{R}_z(\omega_q^\ell \Delta t_{j,h})\boldsymbol{T}_{h\to 0}(\boldsymbol{x}_{q,h}^\ell - \boldsymbol{c}_{q,h}^\ell), \tag{4}$$

where $\boldsymbol{R}_z(\alpha)$ denotes the rotation matrix of angle α along the z-axis. The velocity term translates the target-time center to its source-time position, while the yaw term reverses the accumulated rotation of the local sampling layout. Subsequently, the compensated locations are projected into the source cameras for multi-scale image feature sampling.

Kinematics-Informed Future Motion Regularization. Direct multi-horizon decoding supervises occupancy predictions at each timestamp, but does not by itself provide an explicit motion prior for corresponding query centers. We therefore employ the same kinematic state of query anchors forward in time to regularize future query centers. Let $\boldsymbol{c}_{q,0}^L$ be the current-horizon anchor center after the last decoder layer, and $\boldsymbol{c}_{q,h}^L$ the freely predicted center at horizon h. Using the estimated query velocity $\boldsymbol{v}_q^L$ and the elapsed time t_h, the current query center is advanced to the future horizon as kinematic references and then regularized with predicted future centers as follows:

$$\mathcal{L}_{\text{reg}} = \frac{1}{QH}\sum_{q=1}^{Q}\sum_{h=1}^{H}\mathcal{H}\big(\boldsymbol{c}_{q,h}^L - \boldsymbol{T}_{h\to 0}^{-1}\big(\boldsymbol{c}_{q,0}^L + \boldsymbol{v}_q^L t_h\big)\big), \tag{5}$$

where $\mathcal{H}$ applies the Huber penalty (Huber, 1964) to each metric coordinate of the misalignment and sums the results. Since the yaw rate changes only the orientation of the local point layout, it is neglected during query center rollout. The regularizer provides a soft prior on corresponding anchor centers across horizons, while future occupancy remains directly decoded, so that deviations from non-rigid deformation and newly revealed regions can still be captured without accumulating rollout errors over long horizons.

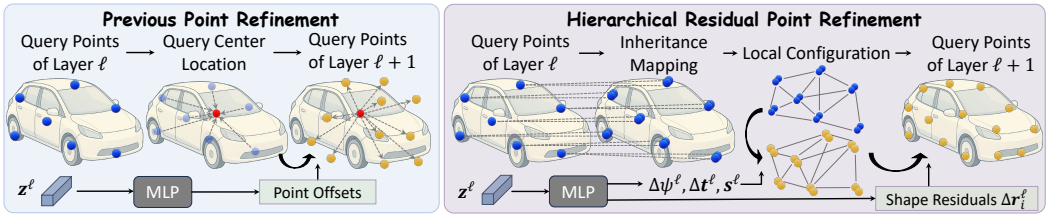

Figure 3: Illustration of the proposed HRPR module. Compared with previous centroid-based point refinement methods, HRPR inherits parent geometry and combines an anchor-level pose update with child-specific shape residuals.

3.4 Hierarchical Residual Point Refinement

As shown in Fig. 3, previous sparse occupancy models (Wang et al., 2024a; Du et al., 2026) progressively increase the number of query points by independently regressing point offsets with respect to a single anchor center inherited from the preceding refinement layer. Such centroid-based point refinement discards explicit parent geometry, and each predicted offset must simultaneously account for anchor-level pose changes and fine-grained shape deformation, thereby entangling global geometric transformations with local structural refinement. To address these limitations, we introduce a *Hierarchical Residual Point Refinement* (HRPR) module, which establishes explicit inheritance correspondences between consecutive refinement layers and decomposes point refinement into an anchor-level pose transformation and point-wise shape residuals.

Let $\mathcal{P}^\ell = \{\boldsymbol{p}_i^\ell \in \mathbb{R}^3\}_{i=1}^{N_\ell}$ be the set of N_ℓ query points at refinement layer ℓ, with their centroid denoted by $\boldsymbol{c}^\ell = \frac{1}{N_\ell}\sum_{i=1}^{N_\ell} \boldsymbol{p}_i^\ell$. To generate $N_{\ell+1}$ child points, we deterministically assign each child slot to a parent point through the inheritance mapping:

$$\kappa_\ell(i) = 1 + \lfloor \frac{(i-1)N_\ell}{N_{\ell+1}} \rfloor, \qquad i \in \{1, \dots, N_{\ell+1}\}, \tag{6}$$

where $\kappa_\ell(i)$ identifies the parent of the i-th point at layer $\ell+1$. Consequently, each refined point inherits an explicit geometric reference from the preceding layer instead of being generated independently from a shared anchor center. Given the query feature $\boldsymbol{z}^\ell$, HRPR predicts a query-level pose update comprising a translation $\Delta\boldsymbol{t}^\ell \in \mathbb{R}^3$, a yaw change $\Delta\psi^\ell$, and an anisotropic scale $\boldsymbol{s}^\ell \in \mathbb{R}^3_+$, together with a point-specific shape residual $\Delta\boldsymbol{r}_i^\ell \in \mathbb{R}^3$. Each child point is then refined relative to its designated parent as follows:

$$\boldsymbol{p}_i^{\ell+1} = \boldsymbol{c}^\ell + \Delta\boldsymbol{t}^\ell + \boldsymbol{R}_z(\Delta\psi^\ell)[\boldsymbol{s}^\ell \odot (\boldsymbol{p}_{\kappa_\ell(i)}^\ell - \boldsymbol{c}^\ell)] + \Delta\boldsymbol{r}_i^\ell, \tag{7}$$

where $\odot$ denotes element-wise multiplication. HRPR preserves the geometric structure inherited from parent points while retaining sufficient flexibility to capture fine-grained local shape variations, providing hierarchical spatial support for subsequent feature sampling and occupancy decoding.

3.5 Training Objectives

Following prior works (Dang et al., 2026; Du et al., 2026), all decoder layers and $H+1$ target horizons receive a Chamfer distance loss $\mathcal{L}_{\text{cd}}$ to align the predicted point distribution with the target occupancy ground truth, and a class-weighted focal loss $\mathcal{L}_{\text{focal}}$ for voxel semantics. Besides the aforementioned regularization loss $\mathcal{L}_{\text{reg}}$, we further introduce a static-scene prior $\mathcal{L}_{\text{static}}$, which penalizes kinematic states on anchors whose current center lies near the occupied voxel of a static class, preventing background anchors from absorbing prediction noise as spurious motion. A feasibility loss $\mathcal{L}_{\text{feas}}$ is also proposed to penalize anchor velocity and yaw rate above prescribed bounds. A detailed description of the training objectives can be found in the appendix. The overall objective is formulated as follows:

$$\mathcal{L} = \mathcal{L}_{\text{cd}} + \mathcal{L}_{\text{focal}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}} + \lambda_{\text{static}}\mathcal{L}_{\text{static}} + \lambda_{\text{feas}}\mathcal{L}_{\text{feas}}. \tag{8}$$

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Dataset and Metrics. Experiments are conducted on the Occ3D-nuScenes benchmark (Caesar et al., 2020; Tian et al., 2023). The model receives RGB images of six surround-view cameras at eight historical timestamps, including the current frame, and forecasts semantic occupancy at 0, 1, 2, and 3 seconds. The evaluation volume is a $200 \times 200 \times 16$ grid with $0.4\,\text{m}$ voxel size, and only camera-visible voxels are evaluated. The future ego trajectory is provided as a conditioning input. Geometric IoU, semantic mIoU are adopted as evaluation metrics for semantic occupancy prediction and forecasting (Du et al., 2026).

Implementation Details. To ensure fair comparison, we use the same experimental setup as prior occupancy world models (Zheng et al., 2024; Dang et al., 2026; Du et al., 2026). Images are resized to 256×704 pixels and encoded with ResNet-50 (He et al., 2016) and an FPN (Lin et al., 2017a) neck. We release two versions of KIO-World: KIO-World-Small uses 600 query anchors with 128 final points per anchor, and KIO-World-Large uses $4,800$ query anchors with 16 final points per anchor. Both models are trained for 70 epochs on eight NVIDIA RTX 4090 GPUs with a total batch size of 8, using the AdamW optimizer (Loshchilov & Hutter, 2019) with an initial learning rate of 2×10^{-4} and a cosine annealing scheme (Loshchilov & Hutter, 2017).

4.2 MAIN RESULTS

4D Occupancy Forecasting. We compare KIO-World with existing occupancy forecasting methods and occupancy world models on the Occ3D-nuScenes benchmark (Tian et al., 2023). We distinguish methods that receive ground-truth occupancy from those that receive camera observations, and use the matched SparseWorld-TC (Du et al., 2026) variants as our principal comparisons. As shown in Table 1, KIO-World achieves state-of-the-art performance among the compared camera-based methods. Compared with SparseWorld-TC-Large, KIO-World-Large achieves a significant improvement of 3.0 points (11.4% relative) on the average semantic mIoU and 3.1 points (6.3% relative) on the average geometric IoU, and KIO-World-Small achieves an improvement of 2.6 points (10.2% relative) on the average semantic mIoU and 2.8 points (5.7% relative) on the average geometric IoU. Moreover, KIO-World achieves strong geometric forecasting performance compared with methods that use ground-truth occupancy as input, while using only surround-view images as scene observations. In terms of average geometric IoU, both variants surpass all Occ-GT-based methods, exceeding the strongest one, I^2-World (Liao et al., 2025b), by at least 2.0 points. KIO-World-Large also exhibits smaller decreases in both metrics over the forecasting horizon. Its geometric IoU degrades far more slowly from 1s to 3s, dropping only 2.6 points, compared with 8.6 for I^2-World, and the decrease in semantic mIoU is only 31.5% of that observed for I^2-World. These results highlight the geometric forecasting capability of our camera-based model. Fig. 4 presents a qualitative comparison of 4D occupancy forecasting. In the yellow boxes, the truck is missing from SparseWorld-TC's forecast, whereas KIO-World retains it near its ground-truth location. In the red boxes, KIO-World predicts more compact and better-separated object shapes than SparseWorld-TC.

Long-Term Occupancy Forecasting. We directly extend KIO-World from the 0–3 s training horizons to forecasting horizons of up to 8s at test time, without additional training. SparseWorld-TC-Large is evaluated using the same direct test-time extension procedure. As shown in Table 2, KIO-World-Large outperforms SparseWorld-TC-Large at every horizon in both metrics, improving average semantic mIoU by 2.5 points (11.2% relative) and average geometric IoU by 3.0 points (6.6% relative). Moreover, the advantage over methods that take occupancy ground truth as input becomes more pronounced as the horizon grows. Compared with COME (Shi et al., 2025), it improves average semantic mIoU by 5.7 points (29.8% relative) and average geometric IoU by 18.4 points (61.3% relative). Although COME achieves a higher semantic mIoU at 1s, KIO-World-Large retains 59.5% of its 1s semantic mIoU at 8s, compared with 35.5% for COME. These results show higher occupancy forecasting accuracy over longer prediction horizons, including horizons beyond those used to train KIO-World.

Table 1: **4D occupancy forecasting performance on Occ3D-nuScenes.** "-STC" and "-F" indicate models incorporating occupancy results from STCOcc and FB-Occ, respectively.

Method	Input	Semantic mIoU (%) ↑				Geometric IoU (%) ↑			
		1s	2s	3s	Avg.	1s	2s	3s	Avg.
OccWorld-O (Zheng et al., 2024)	Occ GT	25.8	15.1	10.5	17.1	34.6	25.1	20.2	26.6
OccLLaMA-O (Wei et al., 2024)	Occ GT	25.1	19.5	15.3	20.0	34.6	28.5	24.4	29.2
RenderWorld (Yan et al., 2025)	Occ GT	28.7	18.9	14.8	20.8	37.7	28.4	24.1	30.1
Occ-LLM (Xu et al., 2025c)	Occ GT	24.0	21.6	17.3	21.0	36.6	32.1	28.8	32.5
DFIT-OccWorld (Zhang et al., 2024)	Occ GT	31.7	21.3	15.2	22.7	40.3	31.2	25.3	32.3
DOME (Gu et al., 2024)	Occ GT	35.1	25.9	20.3	27.1	44.0	35.4	29.7	36.4
UniScene (Li et al., 2025a)	Occ GT	35.4	29.6	25.1	30.0	38.3	32.7	29.1	33.4
I^2-World (Liao et al., 2025b)	Occ GT	47.6	38.6	33.0	39.7	54.3	49.4	45.7	49.8
OccWorld-D (Zheng et al., 2024)	Camera	11.6	8.1	6.2	8.6	18.9	16.3	14.4	16.5
OccLLaMA (Wei et al., 2024)	Camera	10.3	8.7	7.0	8.7	25.8	23.2	20.0	23.0
PreWorld (Li et al., 2025b)	Camera	12.3	9.2	7.2	9.6	23.6	21.6	19.6	21.6
Occ-LLM (Xu et al., 2025c)	Camera	11.3	10.2	9.1	10.2	27.1	24.1	20.2	23.8
DFIT-OccWorld (Zhang et al., 2024)	Camera	13.4	10.2	8.0	10.5	19.2	16.9	15.0	17.0
OccTENS-F (Jin et al., 2026)	Camera	17.2	10.4	7.8	11.8	27.6	25.1	20.3	24.3
SparseWorld (Dang et al., 2026)	Camera	14.9	13.2	11.5	13.2	23.0	22.1	21.1	22.1
DOME-STC (Gu et al., 2024)	Camera	17.8	14.2	11.6	14.5	26.4	23.2	20.4	23.3
I^2-World-STC (Liao et al., 2025b)	Camera	21.7	18.8	16.5	19.0	30.6	28.8	27.0	28.8
DOME-F (Gu et al., 2024)	Camera	24.1	17.4	13.2	18.2	35.2	27.9	23.4	28.8
DTT (Xu et al., 2025a)	Camera	24.9	18.3	15.6	19.6	39.0	37.5	31.9	36.1
COME (Shi et al., 2025)	Camera	26.6	21.7	18.5	22.3	48.1	43.8	40.3	44.1
SparseWorld-TC-Small (Du et al., 2026)	Camera	27.9	25.5	23.4	25.6	50.7	49.1	47.2	49.0
SparseWorld-TC-Large (Du et al., 2026)	Camera	28.6	26.3	24.4	26.4	50.6	49.3	47.8	49.2
KIO-World-Small (Ours)	Camera	30.7	28.1	25.9	28.2	53.3	51.9	50.3	51.8
KIO-World-Large (Ours)	Camera	**31.6**	**29.6**	**27.0**	**29.4**	**53.5**	**52.3**	**50.9**	**52.3**

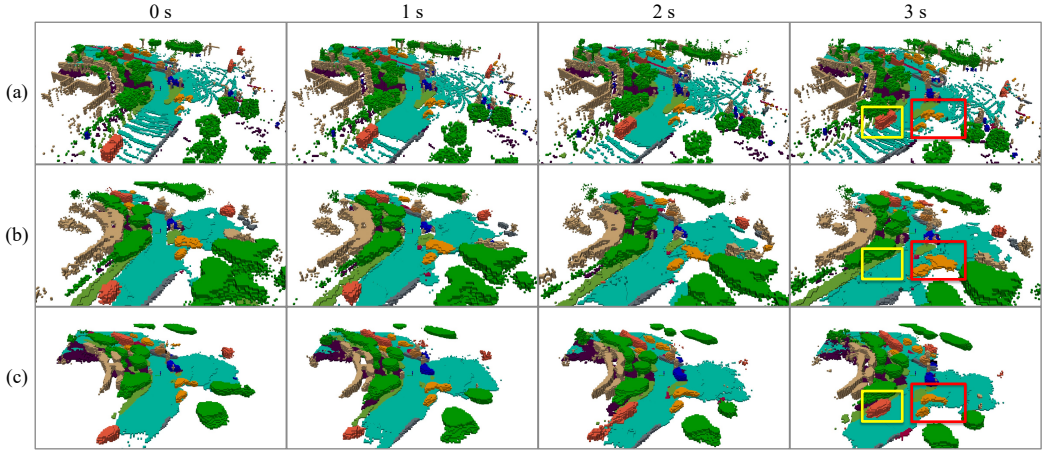

Figure 4: **Qualitative comparison of 4D occupancy forecasting on Occ3D-nuScenes.** Each row shows: (a) occupancy ground truth. (b) forecasting results of SparseWorld-TC, and (c) KIO-World.

4.3 Ablation Study

Cumulative Effect of Proposed Components. Table 3 adds the proposed components one at a time on top of our reproduced direct forecasting baseline, which adopts centroid-based point refinement and rigid historical alignment. HRPR raises semantic mIoU from 25.4 to 26.8 and geometric IoU from 48.7 to 49.6. Motion-compensated historical alignment then brings the two metrics to 27.9 and 50.4. Future motion regularization, which reuses the same kinematic state, adds a further 0.3 mIoU and 1.4 IoU and reaches 28.2 and 51.8. Relative to the baseline, the full model improves semantic mIoU by 2.8 points (11.0% relative) and geometric IoU by 3.1 points (6.4% relative). HRPR and historical alignment account for most of the semantic gain, while the geometric gain is largest when the shared motion state regularizes future anchor centers.

Table 2: **Long-term 4D occupancy forecasting performance on Occ3D-nuScenes.** KIO-World-Large and SparseWorld-TC-Large use the same direct test-time horizon-extension protocol.

Method	Input	Semantic mIoU (%) ↑								
		1s	2s	3s	4s	5s	6s	7s	8s	Avg.
DOME (Gu et al., 2024)	Occ GT	30.1	21.4	17.4	14.9	12.6	11.0	10.0	9.3	15.8
COME (Shi et al., 2025)	Occ GT	**33.8**	24.6	21.4	18.2	15.8	13.9	13.0	12.0	19.1
SparseWorld-TC (Du et al., 2026)	Camera	28.6	26.3	24.4	22.7	21.1	19.7	18.5	17.4	22.3
KIO-World (Ours)	Camera	31.6	**29.6**	**27.0**	**25.5**	**23.9**	**21.6**	**20.5**	**18.8**	**24.8**
Method	Input	Geometric IoU (%) ↑								
		1s	2s	3s	4s	5s	6s	7s	8s	Avg.
DOME (Gu et al., 2024)	Occ GT	39.0	31.2	27.1	24.7	22.3	20.3	19.1	18.0	25.2
COME (Shi et al., 2025)	Occ GT	44.2	36.2	32.9	30.0	26.9	24.7	23.3	21.4	30.0
SparseWorld-TC (Du et al., 2026)	Camera	50.6	49.3	47.8	46.2	44.6	43.0	41.4	40.0	45.4
KIO-World (Ours)	Camera	**53.5**	**52.3**	**50.9**	**49.0**	**47.8**	**46.1**	**44.6**	**43.2**	**48.4**

Table 3: Ablation on proposed modules. "Align." and "Reg." denote historical alignment and future regularization.

HRPR	Align.	Reg.	mIoU	IoU
			25.4	48.7
✓			26.8	49.6
✓	✓		27.9	50.4
✓	✓	✓	**28.2**	**51.8**

Table 4: HRPR as a plug-and-play module. OPUS-T is evaluated and reported on the current frame only.

Method	mIoU	IoU
OPUS-T	33.2	52.6
+ HRPR	**34.1**	**54.2**
SparseWorld-TC-Small	25.6	49.0
+ HRPR	**26.9**	**50.1**

Table 5: Ablation study on choices of kinematic models. mIoU and IoU are averaged over 1–3 s.

Kinematic Model	mIoU	IoU
CV	27.9	51.4
CTRV	28.1	51.8
CV-Yaw	**28.2**	**51.8**

Generalizability of HRPR. Table 4 evaluates HRPR as a plug-and-play module for centroid-based point refinement methods. We incorporate HRPR into OPUS-T (Wang et al., 2024a) for current-frame occupancy prediction and SparseWorld-TC-Small (Du et al., 2026) for 4D occupancy forecasting, keeping the remaining configurations unchanged. On OPUS-T, HRPR improves semantic mIoU and geometric IoU by 0.9 and 1.6 points, respectively. On SparseWorld-TC-Small, it achieves a relative improvement of 5.1% and 2.2% on average semantic mIoU and geometric IoU over 1–3 s, respectively. The consistent gains across both models support the applicability of HRPR to current-frame occupancy prediction and multi-horizon forecasting.

Choice of Kinematic Model. Table 5 compares motion parameterizations under the same horizon-sharing scheme and future motion regularization. CV (Li & Jilkov, 2003) models Cartesian translation without rotating local sampling offsets. CTRV (Schubert et al., 2008) couples translational motion with a constant yaw rate, whereas CV-Yaw models Cartesian translation and local yaw rotation separately. Compared with CV, CV-Yaw improves average semantic mIoU by 0.3 points and average geometric IoU by 0.4 points, reaching 28.2 and 51.8, respectively. CTRV achieves comparable performance, with 28.1 semantic mIoU and the same geometric IoU of 51.8. These results support the use of yaw-aware sampling, while the additional translation–yaw coupling in CTRV provides no further geometric IoU gain in this comparison.

5 Conclusion

We presented KIO-World, a sparse occupancy world model that connects historical evidence alignment and future prediction through a unified kinematic state. The state guides motion-compensated image sampling and supplies soft references for future anchor centers, while HRPR retains explicit geometric inheritance across decoder layers. Together, these components provide a cross-horizon motion prior for query centers and explicit geometric references for point refinement, improving occupancy forecasting accuracy on Occ3D-nuScenes. The reported results establish state-of-the-art performance among the compared camera-based methods. In future work, we plan to address the growing uncertainty of long-horizon forecasting in distant and newly revealed regions.

REFERENCES

Ben Agro, Quinlan Sykora, Sergio Casas, Thomas Gilles, and Raquel Urtasun. UnO: Unsupervised Occupancy Fields for Perception and Forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14487–14496, 2024.

Hengwei Bian, Lingdong Kong, Haozhe Xie, Liang Pan, Yu Qiao, and Ziwei Liu. DynamicCity: Large-Scale 4D Occupancy Generation from Dynamic Scenes. In *International Conference on Learning Representations*, 2025.

Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A Multimodal Dataset for Autonomous Driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11621–11631, 2020.

Zhaowei Cai and Nuno Vasconcelos. Cascade R-CNN: Delving into High Quality Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6154–6162, 2018.

Anh-Quan Cao and Raoul de Charette. MonoScene: Monocular 3D Semantic Scene Completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3991–4001, 2022.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-End Object Detection with Transformers. In *European Conference on Computer Vision*, pp. 213–229, 2020.

Cheng Chen, Hao Huang, and Saurabh Bagchi. GEM: Gaussian Evolution Model for Occupancy Forecasting and Motion Planning. *arXiv preprint arXiv:2605.17682*, 2026.

Junliang Chen, Huaiyuan Xu, Yi Wang, and Lap-Pui Chau. OccProphet: Pushing the Efficiency Frontier of Camera-Only 4D Occupancy Forecasting with an Observer-Forecaster-Refiner Framework. In *International Conference on Learning Representations*, 2025.

Chenxu Dang, Haiyan Liu, Jason Bao, Pei An, Xinyue Tang, An Pan, Jie Ma, Bingchuan Sun, and Yan Wang. SparseWorld: A Flexible, Adaptive, and Efficient 4D Occupancy World Model Powered by Sparse and Dynamic Queries. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(5):3497–3505, 2026.

Jiayuan Du, Yiming Zhao, Zhenglong Guo, Yong Pan, Wenbo Hou, Zhihui Hao, Kun Zhan, and Qijun Chen. SparseWorld-TC: Trajectory-Conditioned Sparse Occupancy World Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7425–7434, 2026.

Ziteng Gao, Limin Wang, Bing Han, and Sheng Guo. AdaMixer: A Fast-Converging Query-Based Object Detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5364–5373, 2022.

Benjamin Graham, Martin Engelcke, and Laurens van der Maaten. 3D Semantic Segmentation with Submanifold Sparse Convolutional Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9224–9232, 2018.

Songen Gu, Wei Yin, Bu Jin, Xiaoyang Guo, Junming Wang, Haodong Li, Qian Zhang, and Xiaoxiao Long. DOME: Taming Diffusion Model into High-Fidelity Controllable Occupancy World Model. *arXiv preprint arXiv:2410.10429*, 2024.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2961–2969, 2017.

Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Tri-Perspective View for Vision-Based 3D Semantic Occupancy Prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9223–9232, 2023.

Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. GaussianFormer: Scene as Gaussians for Vision-Based 3D Semantic Occupancy Prediction. In *European Conference on Computer Vision*, volume 15085 of *Lecture Notes in Computer Science*, pp. 376–393. Springer, 2024.

Yuanhui Huang, Amonnut Thammatadatrakoon, Wenzhao Zheng, Yunpeng Zhang, Dalong Du, and Jiwen Lu. GaussianFormer-2: Probabilistic Gaussian Superposition for Efficient 3D Occupancy Prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27477–27486, 2025.

Peter J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964.

Haoyi Jiang, Tianheng Cheng, Naiyu Gao, Haoyang Zhang, Tianwei Lin, Wenyu Liu, and Xinggang Wang. Symphonize 3D Semantic Scene Completion with Contextual Instance Queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20258–20267, 2024.

Bu Jin, Songen Gu, Xiaotao Hu, Yupeng Zheng, Xiaoyang Guo, Qian Zhang, Xiaoxiao Long, and Wei Yin. OccTENS: 3D Occupancy World Model via Temporal Next-Scale Prediction. *IEEE Robotics and Automation Letters*, 11(3):3566–3573, 2026.

Tarasha Khurana, Peiyun Hu, Achal Dave, Jason Ziglar, David Held, and Deva Ramanan. Differentiable Raycasting for Self-Supervised Occupancy Forecasting. In *European Conference on Computer Vision*, volume 13698 of *Lecture Notes in Computer Science*, pp. 353–369. Springer, 2022.

Tarasha Khurana, Peiyun Hu, David Held, and Deva Ramanan. Point Cloud Forecasting as a Proxy for 4D Occupancy Forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1116–1124, 2023.

Alex H. Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. PointPillars: Fast Encoders for Object Detection from Point Clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12697–12705, 2019.

Bohan Li, Jiazhe Guo, Hongsi Liu, Yingshuang Zou, Yikang Ding, Xiwu Chen, Hu Zhu, Feiyang Tan, Chi Zhang, Tiancai Wang, Shuchang Zhou, Li Zhang, Xiaojuan Qi, Hao Zhao, Mu Yang, Wenjun Zeng, and Xin Jin. UniScene: Unified Occupancy-centric Driving Scene Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11971–11981, 2025a.

X. Rong Li and Vesselin P. Jilkov. Survey of Maneuvering Target Tracking. Part I: Dynamic Models. *IEEE Transactions on Aerospace and Electronic Systems*, 39(4):1333–1364, 2003.

Xiang Li, Pengfei Li, Yupeng Zheng, Wei Sun, Yan Wang, and Yilun Chen. Semi-Supervised Vision-Centric 3D Occupancy World Model for Autonomous Driving. In *International Conference on Learning Representations*, 2025b.

Yiming Li, Zhiding Yu, Christopher Choy, Chaowei Xiao, Jose M. Alvarez, Sanja Fidler, Chen Feng, and Anima Anandkumar. VoxFormer: Sparse Voxel Transformer for Camera-Based 3D Semantic Scene Completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9087–9098, 2023a.

Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. BEVFormer: Learning Bird's-Eye-View Representation from Multi-Camera Images via Spatiotemporal Transformers. In *European Conference on Computer Vision*, volume 13669 of *Lecture Notes in Computer Science*, pp. 1–18. Springer, 2022.

Zhiqi Li, Zhiding Yu, David Austin, Mingsheng Fang, Shiyi Lan, Jan Kautz, and Jose M. Alvarez. FB-OCC: 3D Occupancy Prediction based on Forward-Backward View Transformation. *arXiv preprint arXiv:2307.01492*, 2023b.

Zhimin Liao, Ping Wei, Shuaijia Chen, Haoxuan Wang, and Ziyang Ren. STCOcc: Sparse Spatial-Temporal Cascade Renovation for 3D Occupancy and Scene Flow Prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1516–1526, 2025a.

Zhimin Liao, Ping Wei, Ruijie Zhang, Shuaijia Chen, Haoxuan Wang, and Ziyang Ren. I^2-World: Intra-Inter Tokenization for Efficient Dynamic 4D Scene Forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 25810–25819, 2025b.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision*, pp. 740–755, 2014.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature Pyramid Networks for Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2117–2125, 2017a.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal Loss for Dense Object Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2980–2988, 2017b.

Xuewu Lin, Tianwei Lin, Zixiang Pei, Lichao Huang, and Zhizhong Su. Sparse4D: Multi-view 3D Object Detection with Sparse Spatial-Temporal Fusion. *arXiv preprint arXiv:2211.10581*, 2022.

Haisong Liu, Yao Teng, Tao Lu, Haiguang Wang, and Limin Wang. SparseBEV: High-Performance Sparse 3D Object Detection from Multi-Camera Videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18580–18590, 2023.

Xinhao Liu, Moonjun Gong, Qi Fang, Haoyu Xie, Yiming Li, Hang Zhao, and Chen Feng. LiDAR-based 4D Occupancy Completion and Forecasting. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 11102–11109. IEEE, 2024.

Ilya Loshchilov and Frank Hutter. SGDR: Stochastic Gradient Descent with Warm Restarts. In *International Conference on Learning Representations*, 2017.

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*, 2019.

Yuhang Lu, Xinge Zhu, Tai Wang, and Yuexin Ma. OctreeOcc: Efficient and Multi-Granularity Occupancy Prediction Using Octree Queries. In *Advances in Neural Information Processing Systems*, volume 37, pp. 79618–79641, 2024.

Junyi Ma, Xieyuanli Chen, Jiawei Huang, Jingyi Xu, Zhen Luo, Jintao Xu, Weihao Gu, Rui Ai, and Hesheng Wang. Cam4DOcc: Benchmark for Camera-Only 4D Occupancy Forecasting in Autonomous Driving Applications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21486–21495, 2024.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *European Conference on Computer Vision*, pp. 405–421, 2020.

Chen Min, Dawei Zhao, Liang Xiao, Jian Zhao, Xinli Xu, Zheng Zhu, Lei Jin, Jianshu Li, Yulan Guo, Junliang Xing, Liping Jing, Yiming Nie, and Bin Dai. DriveWorld: 4D Pre-trained Scene Understanding via World Models for Autonomous Driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15522–15533, 2024.

William Peebles and Saining Xie. Scalable Diffusion Models with Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.

Robin Schubert, Eric Richter, and Gerd Wanielik. Comparison and Evaluation of Advanced Motion Models for Vehicle Tracking. In *2008 11th International Conference on Information Fusion*, pp. 1–6. IEEE, 2008.

Yiang Shi, Tianheng Cheng, Qian Zhang, Wenyu Liu, and Xinggang Wang. Occupancy as Set of Points. In *European Conference on Computer Vision*, volume 15119 of *Lecture Notes in Computer Science*, pp. 72–87. Springer, 2024.

Yining Shi, Kun Jiang, Qiang Meng, Ke Wang, Jiabao Wang, Wenchao Sun, Tuopu Wen, Mengmeng Yang, and Diange Yang. COME: Adding Scene-Centric Forecasting Control to Occupancy World Model. In *Advances in Neural Information Processing Systems*, volume 38, pp. 114–140, 2025.

Pin Tang, Zhongdao Wang, Guoqing Wang, Jilai Zheng, Xiangxuan Ren, Bailan Feng, and Chao Ma. SparseOcc: Rethinking Sparse Latent Representation for Vision-Based Semantic Occupancy Prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15035–15044, 2024.

Xiaoyu Tian, Tao Jiang, Longfei Yun, Yucheng Mao, Huitong Yang, Yue Wang, Yilun Wang, and Hang Zhao. Occ3D: A Large-Scale 3D Occupancy Prediction Benchmark for Autonomous Driving. In *Advances in Neural Information Processing Systems*, volume 36, pp. 64318–64330, 2023.

Wenwen Tong, Chonghao Sima, Tai Wang, Li Chen, Silei Wu, Hanming Deng, Yi Gu, Lewei Lu, Ping Luo, Dahua Lin, and Hongyang Li. Scene as Occupancy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8406–8415, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In *Advances in Neural Information Processing Systems*, pp. 5998–6008, 2017.

Jiabao Wang, Zhaojiang Liu, Qiang Meng, Liujiang Yan, Ke Wang, Jie Yang, Wei Liu, Qibin Hou, and Ming-Ming Cheng. OPUS: Occupancy Prediction Using a Sparse Set. In *Advances in Neural Information Processing Systems*, volume 37, pp. 119861–119885, 2024a.

Jiabao Wang, Qiang Meng, Liujiang Yan, Ke Wang, Qibin Hou, and Ming-Ming Cheng. OPUS-V2: Bridging the Gap between Sparse Points and Dense Voxels. *arXiv preprint arXiv:2608.29187*, 2026a.

Lening Wang, Wenzhao Zheng, Yilong Ren, Han Jiang, Zhiyong Cui, Haiyang Yu, and Jiwen Lu. OccSora: 4D Occupancy Generation Models as World Simulators for Autonomous Driving. *IEEE Transactions on Image Processing*, 35:4947–4960, 2026b.

Shihao Wang, Yingfei Liu, Tiancai Wang, Ying Li, and Xiangyu Zhang. Exploring Object-Centric Temporal Modeling for Efficient Multi-View 3D Object Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3621–3631, 2023.

Yuqi Wang, Yuntao Chen, Xingyu Liao, Lue Fan, and Zhaoxiang Zhang. PanoOcc: Unified Occupancy Representation for Camera-based 3D Panoptic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17158–17168, 2024b.

Julong Wei, Shanshuai Yuan, Pengfei Li, Qingda Hu, Zhongxue Gan, and Wenchao Ding. OccLLaMA: An Occupancy-Language-Action Generative World Model for Autonomous Driving. *arXiv preprint arXiv:2409.03272*, 2024.

Yi Wei, Linqing Zhao, Wenzhao Zheng, Zheng Zhu, Jie Zhou, and Jiwen Lu. SurroundOcc: Multi-Camera 3D Occupancy Prediction for Autonomous Driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21729–21740, 2023.

Haoran Xu, Peixi Peng, Guang Tan, Yiqian Chang, Yisen Zhao, and Yonghong Tian. Delta-Triplane Transformers as Occupancy World Models. *arXiv preprint arXiv:2503.07338*, 2025a.

Jingyi Xu, Xieyuanli Chen, Junyi Ma, Jiawei Huang, Jintao Xu, Yue Wang, and Ling Pei. Spatiotemporal Decoupling for Efficient Vision-Based Occupancy Forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22338–22347, 2025b.

Tianshuo Xu, Hao Lu, Xu Yan, Yingjie Cai, Bingbing Liu, and Yingcong Chen. Occ-LLM: Enhancing Autonomous Driving with Occupancy-Based Large Language Models. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pp. 8434–8441, 2025c.

Ziyang Yan, Wenzhen Dong, Yihua Shao, Yuhang Lu, Haiyang Liu, Jingwen Liu, Haozhe Wang, Zhe Wang, Yan Wang, Fabio Remondino, and Yuexin Ma. RenderWorld: World Model with Self-Supervised 3D Label. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pp. 6063–6070, 2025.

Yu Yang, Jianbiao Mei, Yukai Ma, Siliang Du, Wenqing Chen, Yijie Qian, Yuxiang Feng, and Yong Liu. Driving in the Occupancy World: Vision-Centric 4D Occupancy Forecasting and Planning via World Models for Autonomous Driving. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(9):9327–9335, 2025.

Zetong Yang, Li Chen, Yanan Sun, and Hongyang Li. Visual Point Cloud Forecasting enables Scalable Autonomous Driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14673–14684, 2024.

Zichen Yu, Changyong Shu, Jiajun Deng, Kangjie Lu, Zongdai Liu, Jiangyong Yu, Dawei Yang, Hui Li, and Yan Chen. FlashOcc: Fast and Memory-Efficient Occupancy Prediction via Channel-to-Height Plugin. *arXiv preprint arXiv:2311.12058*, 2023.

Zichen Yu, Changyong Shu, Qianpu Sun, Yifan Bian, Xiaobao Wei, Jiangyong Yu, Zongdai Liu, Dawei Yang, Hui Li, and Yan Chen. Panoptic-FlashOcc: An Efficient Baseline to Marry Semantic Occupancy with Panoptic via Instance Center. *arXiv preprint arXiv:2406.10527*, 2024.

Haiming Zhang, Ying Xue, Xu Yan, Jiacheng Zhang, Weichao Qiu, Dongfeng Bai, Bingbing Liu, Shuguang Cui, and Zhen Li. An Efficient Occupancy World Model via Decoupled Dynamic Flow and Image-assisted Training. *arXiv preprint arXiv:2412.13772*, 2024.

Yunpeng Zhang, Zheng Zhu, and Dalong Du. OccFormer: Dual-path Transformer for Vision-based 3D Semantic Occupancy Prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9433–9443, 2023.

Wenzhao Zheng, Weiliang Chen, Yuanhui Huang, Borui Zhang, Yueqi Duan, and Jiwen Lu. OccWorld: Learning a 3D Occupancy World Model for Autonomous Driving. In *European Conference on Computer Vision*, volume 15071 of *Lecture Notes in Computer Science*, pp. 55–72. Springer, 2024.

Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the Continuity of Rotation Representations in Neural Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5745–5753, 2019.

Sicheng Zuo, Wenzhao Zheng, Yuanhui Huang, Jie Zhou, and Jiwen Lu. GaussianWorld: Gaussian World Model for Streaming 3D Occupancy Prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6772–6781, 2025.

Table 6: Summary of notation.

Symbol	Description
$q \in \{1, \dots, Q\}$	Query anchor index; Q denotes the number of anchors.
$h \in \{0, \dots, H\}$	Target horizon index; $h = 0$ is the current frame, and H denotes the number of future horizons.
$j \in \{-T, \dots, 0\}$	Source frame index; $j = 0$ is the current frame, and T denotes the number of historical frames.
$\ell \in \{0, \dots, L\}$	Decoder layer index; $\ell = 0$ denotes initialization, and L denotes the number of decoder layers.
$i \in \{1, \dots, N_\ell\}$	Point index within a query anchor; N_ℓ denotes the number of points per anchor at layer ℓ.
D	Query feature dimension.
$\mathcal{P}^\ell_{q,h} = \{\boldsymbol{p}^\ell_{q,h,i}\}_{i=1}^{N_\ell}$	Explicit 3D point set of anchor q at horizon h.
$\boldsymbol{c}^\ell_{q,h}$	Anchor center, the mean of $\mathcal{P}^\ell_{q,h}$.
$\boldsymbol{z}^\ell_{q,h} \in \mathbb{R}^D$	Query feature of anchor q at horizon h.
$\boldsymbol{k}^\ell_q = (\boldsymbol{v}^\ell_q, \omega^\ell_q)$	Horizon-shared kinematic state of anchor q, with velocity $\boldsymbol{v}^\ell_q \in \mathbb{R}^3$ in m/s and yaw rate ω^ℓ_q in rad/s.
t_h, t_j	Physical timestamps of target horizon h and source frame j in seconds, with $t_0 = 0$.
$\Delta t_{j,h} = t_j - t_h$	Signed time interval from target horizon h to source frame j; $\Delta t_{j,h} \leq 0$ for historical sources.
$\boldsymbol{T}_{h\to 0} \in \mathrm{SE}(3)$	Rigid transformation from the ego frame at horizon h to the current ego frame.
$\boldsymbol{g}_h \in \mathbb{R}^D$	Trajectory condition embedding of horizon h.
$\mathcal{O}_h$	Semantic occupancy prediction at horizon h.

A AI Assistance

Large language models were used as writing and verification assistants during the preparation of this manuscript. Their use covered language polishing, latex formatting, literature search, and consistency checks between the manuscript and the implementation. All AI-assisted material was critically reviewed and revised by the authors. The authors take full responsibility for the content of this work, including all text produced with AI assistance.

B Notation

Table 6 summarizes the notation used in the main paper and this appendix. Superscripts index the decoder layer, and the leading subscripts index the query anchor and the target horizon. Unless stated otherwise, metric quantities are expressed in the current ego frame.

C Additional Method Details

C.1 KIO-World Encoder

The KIO-World encoder prepares three inputs for the decoder: the initial sparse query anchors, a trajectory condition embedding for each target horizon, and multi-scale image features of the historical and current observations.

Query Anchor Initialization. Following previous sparse perception models (Carion et al., 2020; Wang et al., 2024a; Dang et al., 2026; Du et al., 2026), each query anchor q is initialized with a single

learnable 3D point, which is expressed in coordinates normalized by the occupancy perception range and initialized from a uniform distribution over the unit cube $[0,1]^3$. This point is shared by all $H+1$ horizons, so that anchors with the same index start from the same location and remain in correspondence across horizons. The query features $\boldsymbol{z}_{q,h}^0$ are initialized as zero vectors, and the raw kinematic state $\widetilde{\boldsymbol{k}}_q^0$ is initialized to zero, as in Eq. (3).

Trajectory Condition Embedding. For each target horizon h, the condition encoder embeds the future-to-current ego pose $\boldsymbol{T}_{h\to 0} \in \mathrm{SE}(3)$ together with the temporal offset of the horizon. We decompose each ego pose into a translation vector $\boldsymbol{t}_h$ and a rotation matrix $\boldsymbol{R}_h$, with the former normalized by the half-extent of the occupancy volume, and the latter represented by its first two columns, yielding $\widetilde{\boldsymbol{t}}_h$ and $\rho(\boldsymbol{R}_h)$, respectively. Both components thus have comparable magnitudes, and the resulting 6D rotation is a continuous parameterization of $\mathrm{SO}(3)$ (Zhou et al., 2019). For the temporal embedding, the horizon timestamp t_h is encoded by a sinusoidal encoding (Vaswani et al., 2017) as follows:

$$\boldsymbol{e}(t_h) = \big[\sin(t_h\boldsymbol{\nu}); \cos(t_h\boldsymbol{\nu})\big], \qquad \nu_k = 10000^{-k/(D/2-1)}, \quad k = 0, \ldots, D/2-1. \tag{9}$$

The pose coefficients and the temporal encoding are subsequently concatenated and mapped to the condition embedding by a two-layer MLP as follows:

$$\boldsymbol{g}_h = \mathrm{MLP}\Big(\big[\widetilde{\boldsymbol{t}}_h,\ \rho(\boldsymbol{R}_h),\ \boldsymbol{e}(t_h)\big]\Big), \tag{10}$$

where $\boldsymbol{g}_h \in \mathbb{R}^D$ is the condition embedding shared by all anchors of that horizon and is injected into every decoder layer through AdaLN-Zero (Appendix C.2).

Image Feature Encoding. The image encoder consists of a ResNet-50 (He et al., 2016) backbone and an FPN (Lin et al., 2017a) neck. The backbone is initialized from the Cascade Mask R-CNN (Cai & Vasconcelos, 2018; He et al., 2017) pretrained on COCO (Lin et al., 2014) and nuImages (Caesar et al., 2020). During training, the stem and the first residual stage are frozen, and all batch normalization layers use frozen statistics. The encoder processes the 6 surround-view images at each of the 8 input timestamps, and produces image features at four scales with a unified channel dimension D, which forms a shared feature space for subsequent fusion with the query features.

C.2 KIO-World Transformer

Each of the L decoder layers executes the following operations in order: trajectory-conditioned AdaLN-Zero, positional encoding, kinematic state estimation, image feature sampling, adaptive feature mixing, temporal attention across horizons, spatial attention within each horizon, HRPR, and point feature decoding. We detail each operation below following this execution order.

Trajectory-Conditioned AdaLN-Zero. Following the adaptive layer normalization with zero initialization (AdaLN-Zero) of DiT (Peebles & Xie, 2023), the condition $\boldsymbol{g}_h$ from Eq. (10) is injected at the entrance of every decoder layer. A zero-initialized linear projection predicts a shift $\boldsymbol{\beta}_h^\ell$, a scale $\boldsymbol{\gamma}_h^\ell$, and a gate $\boldsymbol{\alpha}_h^\ell$, which are broadcast to every anchor of horizon h:

$$\begin{aligned} (\boldsymbol{\beta}_h^\ell, \boldsymbol{\gamma}_h^\ell, \boldsymbol{\alpha}_h^\ell) &= \boldsymbol{W}_\ell\, \mathrm{SiLU}(\boldsymbol{g}_h) + \boldsymbol{b}_\ell, \\ \widetilde{\boldsymbol{z}}_{q,h}^\ell &= \boldsymbol{z}_{q,h}^{\ell-1} + \boldsymbol{\alpha}_h^\ell \odot \Big((1+\boldsymbol{\gamma}_h^\ell) \odot \mathrm{LN}(\boldsymbol{z}_{q,h}^{\ell-1}) + \boldsymbol{\beta}_h^\ell\Big), \end{aligned} \tag{11}$$

where LN denotes layer normalization without learnable affine parameters. Since $\boldsymbol{W}_\ell$ and $\boldsymbol{b}_\ell$ are initialized to zero, Eq. (11) reduces to the identity at initialization, and the trajectory condition is learned on top of this mapping. Unlike the original DiT block, which modulates both its attention and feed-forward branches, KIO-World applies a single gated modulation to the query stream at the entrance of each decoder layer.

Positional Encoding. The current anchor geometry is then injected by a Fourier positional encoding (Mildenhall et al., 2020) of the normalized coordinates of all points in $\mathcal{P}_{q,h}^{\ell-1}$, which is projected to the feature dimension by an MLP and added to $\widetilde{\boldsymbol{z}}_{q,h}^\ell$. The resulting pre-sampling feature $\bar{\boldsymbol{z}}_{q,h}^\ell$ encodes both the feature content and the current anchor geometry.

Kinematic State Estimation. A motion head reads only the current-horizon pre-sampling feature $\bar{\boldsymbol{z}}^{\ell}_{q,0}$, which is the input of f^{ℓ}_{θ} in Eq. (3), and predicts a residual that is accumulated into the raw state $\widetilde{\boldsymbol{k}}^{\ell}_{q}$. The raw state is mapped to the metric CV-Yaw state by a component-wise bounded function:

$$\boldsymbol{k}^{\ell}_{q} = \boldsymbol{b} \odot \tanh\left(\widetilde{\boldsymbol{k}}^{\ell}_{q} \oslash \boldsymbol{b}\right), \tag{12}$$

where $\odot$ and $\oslash$ denote element-wise multiplication and division, respectively, and $\boldsymbol{b}$ collects the upper bounds of the velocity components and the yaw rate. The motion head is zero-initialized, and the first decoder layer has no motion head, so that sampling starts from rigid ego-motion alignment and motion compensation is learned progressively. The bounded state is shared by all horizons of anchor q for image feature sampling.

Image Feature Sampling. Each anchor generates several groups of sampling points around its input geometry, by scaling a learnable sampling prototype with the per-axis spread of $\mathcal{P}^{\ell-1}_{q,h}$ and adding offsets predicted from $\bar{\boldsymbol{z}}^{\ell}_{q,h}$. The same local pattern is shared by all source frames. Half of the groups are backtraced to each source frame by Eq. (4) using the recorded source timestamps, whereas the other half follow rigid ego-motion alignment. Every anchor therefore aggregates both rigidly aligned and motion-compensated evidence, and the rigid groups remain informative when the estimated state is inaccurate. The resulting locations are projected into the surround-view cameras of the corresponding source frames, and the features are bilinearly sampled from the multi-scale feature maps and fused with scale weights predicted from the query.

Adaptive Feature Mixing. The sampled features are fused into the query by adaptive mixing (Gao et al., 2022; Wang et al., 2024a), in which mixing weights generated from $\bar{\boldsymbol{z}}^{\ell}_{q,h}$ first transform the channels of each sample and then aggregate information across sampling points and source frames. The mixed features are added to the query through a residual connection.

Spatiotemporal Attention. The query features then interact across horizons and anchors through factorized temporal and spatial attention. Temporal attention is applied independently to the $H+1$ horizon tokens of each anchor index, so that information is exchanged only among corresponding anchors. Following SparseBEV (Liu et al., 2023), spatial attention is applied among the Q anchors of each horizon, where a query-dependent penalty on the distance between anchor centers is subtracted from the attention logits as a learnable locality prior.

Hierarchical Residual Point Refinement. The output feature $\boldsymbol{z}^{\ell}_{q,h}$ drives HRPR, which refines $\mathcal{P}^{\ell-1}_{q,h}$ into $\mathcal{P}^{\ell}_{q,h}$ through Eqs. (6) and (7). Separate heads predict the anchor-level pose update and the point-wise shape residuals, and all updates are applied in metric coordinates. The yaw change is bounded by a $\tanh$ function, and the anisotropic scale is predicted in log space with clipping. The pose head is zero-initialized, so that each child initially coincides with its parent up to the shape residual. In the first decoder layer, each anchor contains a single point, so the rotation and scale terms vanish and the children are placed by the translation and shape residuals only.

Point Feature Decoding. An MLP maps $\boldsymbol{z}^{\ell}_{q,h}$ to a descriptor for each refined point in $\mathcal{P}^{\ell}_{q,h}$, and these point descriptors are passed to the occupancy head. During training, the descriptors are decoded at every decoder layer for deep supervision, while only descriptors of the last decoder layer is decoded at inference.

C.3 KIO-World Occupancy Head

Sparse Voxel Densification. The densifier in the KIO-World's occupancy head converts the refined points and their descriptors into sparse semantic voxels. The refined points of all anchors are first voxelized, and each point descriptor is augmented with the offsets of the point to its voxel center and to its anchor center. PointPillars-style point feature networks (Lang et al., 2019) then aggregate the points within each voxel. A sparse convolution expands the occupied set to neighboring voxels, and a submanifold sparse convolution (Graham et al., 2018) predicts the semantic logits at every active voxel. This expansion allows the head to recover occupied voxels adjacent to the refined points and provides free-space negatives for semantic supervision. At inference, only the last-layer

densifier is used. The active voxels whose maximum class score exceeds a threshold are assigned the corresponding class, and the remaining voxels are treated as free space.

C.4 Training Objectives

Occupancy Supervision. For each decoder layer and horizon, the point loss $\mathcal{L}_{\text{cd}}$ matches every refined point to its nearest occupied ground-truth voxel center and vice versa, and penalizes the metric offsets with a Smooth-L_1 loss. Ground-truth points of rare classes and camera-visible ground-truth points far from all predictions receive larger weights, and the initial anchor points are supervised in the same way. The semantic loss $\mathcal{L}_{\text{focal}}$ supervises every voxel from the densifier with the ground-truth label at its coordinate, treating free voxels as background, through a class-weighted sigmoid focal loss (Lin et al., 2017b). Both losses are summed over all decoder layers.

Static-Scene Prior. To construct the anchor set $\mathcal{S}$ of $\mathcal{L}_{\text{static}}$, each current-horizon center $\boldsymbol{c}_{q,0}^{L}$ is matched to its nearest camera-visible occupied ground-truth voxel. An anchor is included if its match lies within a threshold and belongs to a static class. Anchors matched to movable classes, including parked vehicles, as well as unmatched anchors, receive no prior. The prior penalizes the full kinematic state of the selected anchors:

$$\mathcal{L}_{\text{static}} = \frac{1}{\max(1, |\mathcal{S}|)} \sum_{q \in \mathcal{S}} \left(\left\| \boldsymbol{v}_q^L \right\|_2^2 + \left(\omega_q^L \right)^2 \right). \tag{13}$$

The yaw-rate term suppresses spurious rotation of the sampling layout, which $\mathcal{L}_{\text{reg}}$ cannot constrain. The matching is used only during training.

Motion Feasibility. With $[x]_+ = \max(0, x)$, the feasibility objective averages over all anchors

$$\mathcal{L}_{\text{feas}} = \frac{1}{Q} \sum_{q=1}^{Q} \left(\left[\| \boldsymbol{v}_{q,xy}^L \|_2 - v_{\max} \right]_+^2 + \left[|\omega_q^L| - \omega_{\max} \right]_+^2 \right), \tag{14}$$

where $v_{\max}$ and $\omega_{\max}$ are prescribed bounds on the planar speed and the yaw rate, respectively. Since Eq. (12) bounds the planar velocity components individually, the first term additionally constrains the planar speed.

Loss Weights. In Eq. (8), the per-layer point and semantic losses are weighted by 0.5 and 2.0, respectively, and we set $\lambda_{\text{reg}} = 0.05$, $\lambda_{\text{static}} = 0.1$, and $\lambda_{\text{feas}} = 0.05$.

D Additional Experimental Results

D.1 Ray-Level Forecasting Results

Table 7 reports RayIoU (Tang et al., 2024) at the current frame and each future horizon under distance thresholds of 1, 2, and 4 m. RayIoU evaluates the semantic class and depth of the first occupied voxel along each query ray, and therefore does not reward overly thick surface predictions. KIO-World outperforms SparseWorld-TC(Du et al., 2026) at every horizon and distance threshold, and its RayIoU advantage grows from 1.8 points at the current frame to 2.9 points at 3 s, reaching 31.1 versus 28.5 on average over 1–3 s. The improvement is similar across distance thresholds, ranging from 2.3 to 2.8 points on average over 1–3 s, which indicates that the gains do not rely on a loose depth tolerance.

D.2 Per-Class Forecasting Results

Tables 8 and 9 report the per-class IoU and RayIoU of SparseWorld-TC-Small and KIO-World-Small at the current frame and each future horizon, where the per-class RayIoU averages the three distance thresholds. In terms of RayIoU, the largest gains appear on small movable classes: at 3 s, KIO-World improves pedestrian from 5.9 to 11.3, bicycle from 7.6 to 13.5, and motorcycle from 6.7 to 9.8 points. For pedestrian and bicycle, the gap widens steadily as the horizon grows, which is consistent with the motion-compensated alignment of moving content. Static classes also benefit

Table 7: RayIoU (%) of 4D occupancy forecasting on Occ3D-nuScenes. S and K denote SparseWorld-TC-Small and KIO-World-Small, respectively, and Avg. is the mean over 1–3 s.

	0s		1s		2s		3s		Avg.	
	S	K	S	K	S	K	S	K	S	K
$\text{RayIoU}_{1\text{m}}$	30.2	31.6	25.9	27.8	22.8	25.1	20.4	23.0	23.0	**25.3**
$\text{RayIoU}_{2\text{m}}$	37.2	39.3	32.3	34.8	28.6	31.4	25.7	28.7	28.8	**31.6**
$\text{RayIoU}_{4\text{m}}$	42.3	44.3	37.2	39.8	33.3	36.1	30.2	33.1	33.6	**36.3**
RayIoU	36.6	38.4	31.8	34.1	28.2	30.9	25.4	28.3	28.5	**31.1**

Table 8: Per-class IoU (%) of 4D occupancy forecasting on Occ3D-nuScenes (Tian et al., 2023). S and K denote SparseWorld-TC-Small and KIO-World-Small, respectively, and Avg. is the mean over 1–3 s. SparseWorld-TC-Small is re-evaluated with its released checkpoint.

	0s		1s		2s		3s		Avg.	
	S	K	S	K	S	K	S	K	S	K
Others	10.4	11.5	9.7	11.2	9.0	10.8	8.3	10.2	9.0	**10.7**
Barrier	37.4	44.7	35.8	43.1	33.6	41.0	30.7	38.3	33.4	**40.8**
Bicycle	16.6	21.5	11.1	15.8	8.3	12.8	5.6	10.2	8.3	**12.9**
Bus	34.7	39.9	25.3	29.1	18.3	24.7	13.8	19.4	19.1	**24.4**
Car	42.2	46.5	30.6	37.9	24.4	27.6	20.5	23.3	25.2	**29.6**
Construction Vehicle	20.9	20.8	19.8	18.5	17.9	16.9	16.0	15.5	**17.9**	17.0
Motorcycle	16.4	22.3	9.7	15.4	7.6	13.3	6.1	11.7	7.8	**13.5**
Pedestrian	12.8	18.9	6.4	11.5	3.6	7.7	2.1	5.6	4.0	**8.3**
Traffic Cone	17.2	23.3	14.9	20.7	12.2	18.0	9.4	15.3	12.2	**18.0**
Trailer	22.5	23.8	20.3	20.9	17.9	19.9	15.6	17.5	17.9	**19.4**
Truck	29.8	31.4	23.3	28.5	19.1	22.4	16.3	19.3	19.6	**23.4**
Driveable Surface	73.3	73.4	72.5	72.6	71.5	71.7	70.0	70.5	71.3	**71.6**
Other Flat	40.9	41.2	40.5	40.9	39.4	40.2	37.6	38.8	39.2	**40.0**
Sidewalk	45.7	48.4	45.2	47.7	44.1	46.5	42.4	44.9	43.9	**46.4**
Terrain	43.7	45.9	42.9	45.1	41.7	44.1	40.1	42.6	41.6	**43.9**
Manmade	32.6	32.2	31.5	31.3	30.0	30.0	28.2	28.5	**29.9**	**29.9**
Vegetation	32.4	32.1	31.3	31.2	29.8	29.9	28.0	28.4	29.7	**29.8**
Semantic mIoU	31.1	34.0	27.7	30.7	25.2	28.1	23.0	25.9	25.3	**28.2**
Geometric IoU	51.1	54.5	49.9	53.3	48.4	51.9	46.6	50.3	48.3	**51.8**

at longer horizons: their gains are small at the current frame but grow steadily with the horizon, *e.g.*, from 1.6 to 4.9 points for driveable surface and from 1.7 to 4.2 points for other flat between 0 and 3 s. This trend is consistent with the future center regularization and the static-scene prior, which keep static anchors aligned with the ego motion. Overall, the RayIoU gap between the two methods widens with the forecasting horizon for most classes, both movable and static, indicating that beyond a better current-frame reconstruction, KIO-World maintains the scene more consistently over time.

D.3 Ablations on the Kinematic State

The following ablations are conducted with KIO-World-Small under the same training setting, and semantic mIoU and geometric IoU are averaged over 1–3 s.

State Sharing across Directions and Horizons. Table 10 examines whether a single kinematic state should be shared by both temporal directions and by all horizons. The direction-specific variant estimates two states from the same current-horizon features, one for historical alignment and one for future center regularization. The horizon-specific variant estimates a separate state from the features of each horizon and uses it for both operations at that horizon. Without a kinematic state, the model uses rigid historical alignment and no future center regularization, which corresponds to the second row of Table 3. Sharing a single state across both directions and all horizons yields

Table 9: Per-class RayIoU (%) of 4D occupancy forecasting on Occ3D-nuScenes (Tian et al., 2023). Each entry averages the RayIoU over the distance thresholds of 1, 2, and 4 m. S and K denote SparseWorld-TC-Small and KIO-World-Small, respectively, and Avg. is the mean over 1–3 s.

	0s		1s		2s		3s		Avg.	
	S	K	S	K	S	K	S	K	S	K
Others	9.1	10.4	8.4	10.0	7.8	9.7	7.5	9.4	7.9	**9.7**
Barrier	41.5	44.6	40.2	42.9	38.0	41.0	35.6	38.7	38.0	**40.8**
Bicycle	22.1	26.2	14.4	19.2	10.8	16.2	7.6	13.5	11.0	**16.3**
Bus	57.2	61.2	42.1	48.4	31.5	36.7	25.1	31.5	32.9	**38.8**
Car	54.3	57.3	43.0	45.7	34.7	37.4	29.6	32.3	35.8	**38.5**
Construction Vehicle	27.7	26.4	26.2	25.3	24.9	24.1	23.6	22.6	**24.9**	24.0
Motorcycle	21.1	25.3	10.2	13.5	7.9	11.5	6.7	9.8	8.3	**11.6**
Pedestrian	26.9	29.8	16.8	20.5	9.5	14.8	5.9	11.3	10.7	**15.5**
Traffic Cone	27.0	29.0	24.5	27.1	21.5	24.7	18.0	22.1	21.3	**24.6**
Trailer	26.6	29.2	22.7	25.7	18.2	20.6	14.0	15.5	18.3	**20.6**
Truck	49.0	50.4	42.4	43.4	37.3	37.9	33.9	33.9	37.9	**38.4**
Driveable Surface	67.9	69.5	65.4	68.0	62.4	66.0	59.3	64.2	62.4	**66.1**
Other Flat	36.7	38.4	35.2	37.9	33.1	37.0	31.2	35.4	33.2	**36.8**
Sidewalk	33.2	33.8	31.9	33.0	30.1	31.8	28.3	30.4	30.1	**31.7**
Terrain	35.5	36.0	34.3	35.6	32.6	34.6	30.8	33.2	32.6	**34.5**
Manmade	46.1	46.6	44.6	45.7	42.7	44.2	40.5	42.5	42.6	**44.2**
Vegetation	39.6	38.8	38.3	37.9	36.4	36.4	34.3	34.7	**36.3**	**36.3**
RayIoU	36.6	38.4	31.8	34.1	28.2	30.9	25.4	28.3	28.5	**31.1**

Table 10: Ablation on kinematic state sharing. mIoU and IoU are averaged over 1–3 s.

Variant	Alignment State	Regularization State	mIoU	IoU
No Kinematic State	–	–	26.8	49.6
Direction-Specific	$\boldsymbol{k}_q^{\text{past}}$	$\boldsymbol{k}_q^{\text{fut}}$	27.8	50.7
Horizon-Specific	$\boldsymbol{k}_{q,h}$	$\boldsymbol{k}_{q,h}$	27.2	49.9
Shared (ours)	$\boldsymbol{k}_q$	$\boldsymbol{k}_q$	28.2	51.8

the best results. Separating the states of the two directions reduces mIoU and IoU to 27.8 and 50.7, respectively. Compared with the shared-state model without future center regularization in Table 11, this variant retains only a marginal geometric gain of 0.3 IoU and no semantic gain, *i.e.*, most of the benefit of the regularization vanishes once the states are separated. In this variant, the future center regularization only shapes the state used for the future references, whereas the state used for historical alignment is learned solely through alignment, so the two operations no longer benefit from each other's supervision. Estimating a separate state for each horizon further degrades the two metrics to 27.2 and 49.9, which is even below the shared-state model without future center regularization. With horizon-specific velocities, the fitting term can be satisfied at every horizon by adjusting the velocity of that horizon alone, so the future centers of non-static anchors are no longer tied to a common motion reference. Moreover, the same anchor is backtraced with inconsistent motion hypotheses at different horizons, which also impairs historical alignment. These results indicate that the benefit of the kinematic state comes from sharing it across both temporal directions and all horizons, rather than from introducing additional motion parameters.

Design of Future Center Regularization. Table 11 ablates the two training terms that constrain the kinematic state. Without both terms, the kinematic state is learned only through motion-compensated historical alignment, which corresponds to the third row of Table 3. The fitting term $\mathcal{L}_{\text{reg}}$ regularizes the future centers towards the kinematic references, and the static-scene prior $\mathcal{L}_{\text{static}}$ further suppresses spurious motion on static structure. Using the fitting term alone brings almost no improvement, reaching 27.9 mIoU and 50.6 IoU. We attribute this to the fact that spurious velocities of static anchors can absorb the discrepancies of their future centers, which weakens the regularization on static structure. Adding the static-scene prior on top of the fitting term raises the two metrics to 28.2 and 51.8, respectively. Once the velocities of static anchors are suppressed, the fitting term constrains their future centers to follow the ego motion, which mainly benefits the geometric accuracy dominated by static structure.

Table 11: Ablation on future center regularization. mIoU and IoU are averaged over 1–3 s.

$\mathcal{L}_{\text{reg}}$	$\mathcal{L}_{\text{static}}$	mIoU	IoU
		27.9	50.4
✓		27.9	50.6
✓	✓	28.2	51.8

Table 12: Soft kinematic prior versus hard rollout. mIoU and IoU are averaged over 1–3 s.

Variant	Future Geometry	Kinematic State	mIoU	IoU
Direct Decoding	Decoded	–	26.8	49.6
Static Ego-Warp	Rolled Out	–	24.9	41.7
Kinematic Rollout	Rolled Out	Hard Constraint	21.2	28.9
KIO-World	Decoded	Soft Prior	28.2	51.8

Soft Kinematic Prior versus Hard Rollout. Table 12 compares the soft kinematic prior with variants that generate future geometry by rolling out the current anchors instead of decoding each horizon. Static ego-warp transports the current anchors to each future horizon using only the ego motion, whereas kinematic rollout additionally displaces and rotates all anchors with the estimated kinematic state. Direct decoding without any kinematic prior corresponds to the second row of Table 3. Both rollout variants perform substantially worse than direct decoding. Static ego-warp reduces mIoU and IoU to 24.9 and 41.7, since rigidly transported anchors can represent neither the motion of dynamic objects nor the regions that become visible at future horizons. Kinematic rollout further degrades the two metrics to 21.2 and 28.9. When the estimated state serves as a hard constraint, its errors directly displace the future geometry, and the resulting displacement errors grow with the forecasting horizon. In contrast, using the same kinematic state as a soft prior improves direct decoding to 28.2 and 51.8, which supports our design of retaining direct multi-horizon decoding and using kinematics only for historical alignment and future regularization.

E Additional Qualitative Results

E.1 Long-Term Occupancy Forecasting

Fig. 5 demonstrates long-term forecasting results of KIO-World from 1 to 8 s, obtained by extending the forecasting horizon at test time without additional training. In the straight-driving case, the drivable areas, curbs, and buildings remain stable over all horizons, and the vehicles along the street keep coherent shapes. In the left-turn case, the predicted layout rotates consistently with the conditioned ego trajectory, and the junction and vegetated traffic islands preserve their shapes while moving across the view. Beyond 6 s, the forecasts gradually become less complete, particularly in regions far from the ego vehicle or newly revealed by the turn, reflecting the growing uncertainty of long-term forecasting.

E.2 Forecasting under Different Trajectory Conditions

Fig. 6 shows forecasts of KIO-World in two scenes, where the same observations are conditioned on two different future ego trajectories: turning left or going straight in the first scene, and going straight or turning right in the second scene. The current-frame forecasts are identical across conditions, since the trajectory condition only affects future horizons. As the horizon grows, the forecasts diverge according to the conditioned trajectory: for turning trajectories, the road layout rotates in the ego frame towards the corresponding direction, whereas for going straight the scene mainly shifts along the heading. Static structures, such as buildings and vegetation, keep consistent shapes across conditions, indicating that the forecasts respond to the trajectory condition while remaining consistent with the observed scene.

E.3 Qualitative Comparison with SparseWorld-TC

As shown in Figs. 7–9, we provide additional qualitative comparison between KIO-World and SparseWorld-TC (Du et al., 2026). Images at future timestamps are shown only for reference and are not used as inputs.

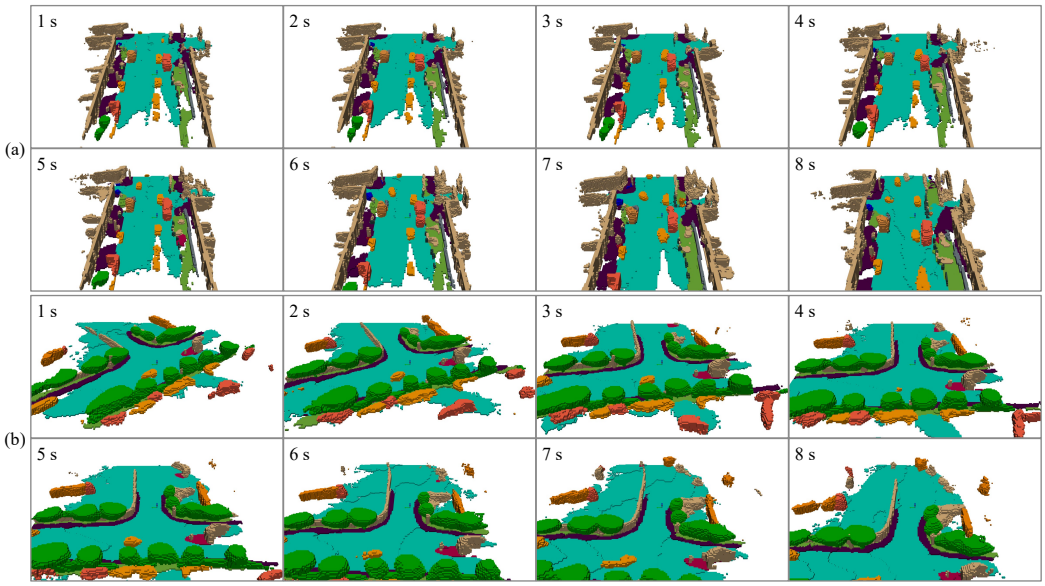

Figure 5: **Long-term occupancy forecasting results of KIO-World on Occ3D-nuScenes.** Each panel is the forecast at the labeled horizon from 1 to 8 s. Two different cases are selected: (a) straight driving and (b) left turn.

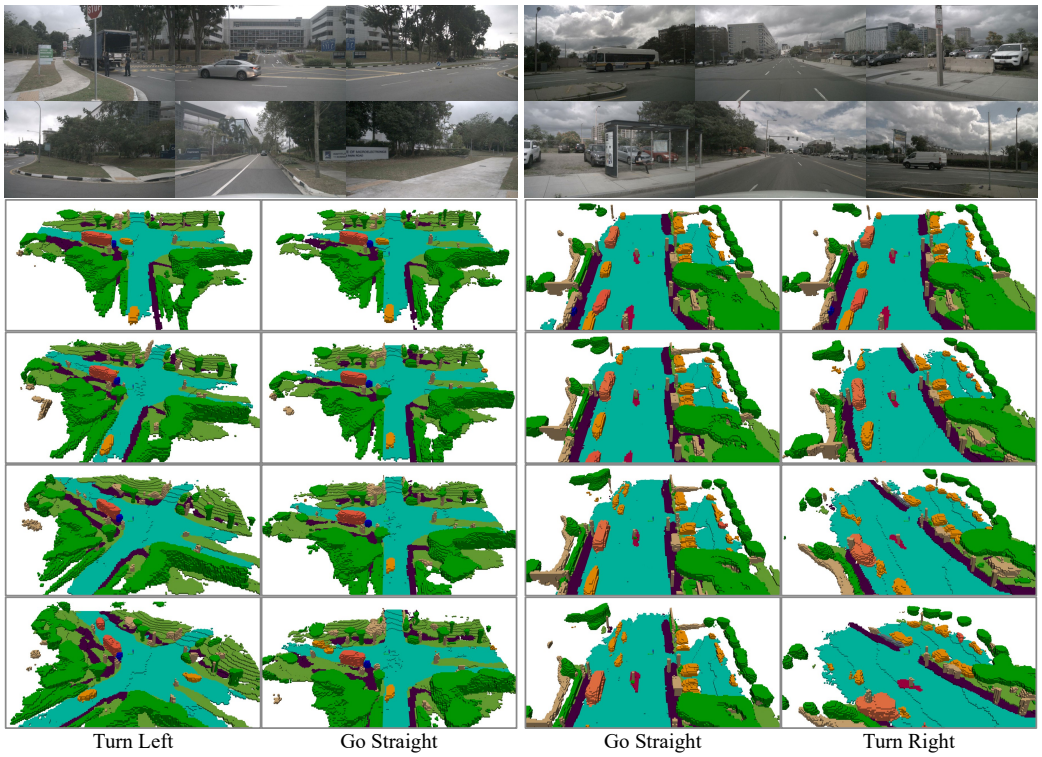

Figure 6: **Occupancy forecasting results of KIO-World under different trajectory conditions on Occ3D-nuScenes.** For each of the two scenes, the top panel shows the current surround-view images, and the rows below show the forecasts from 0 to 3 s under the trajectory condition labeled at the bottom.

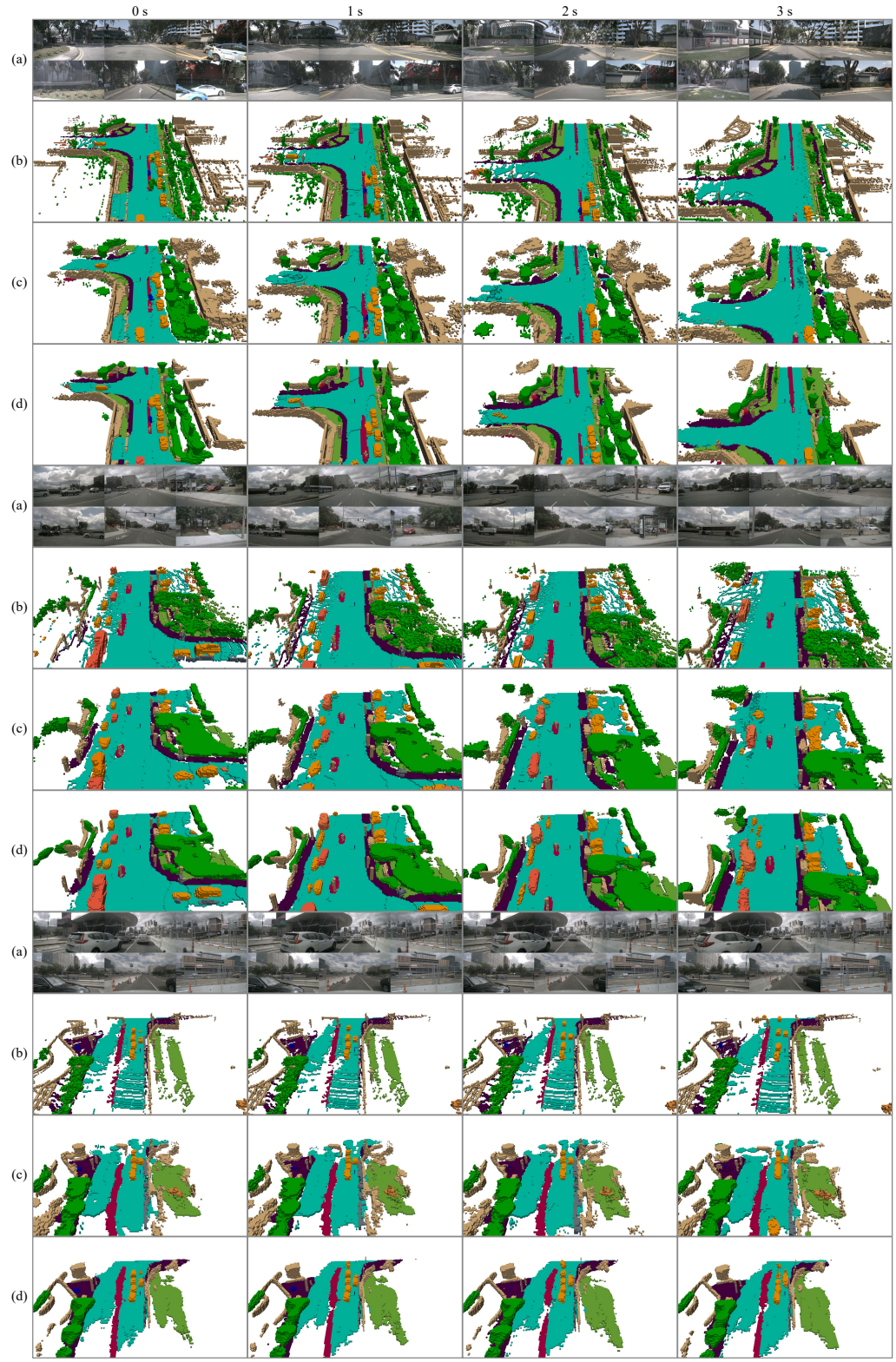

Figure 7: **Additional qualitative comparison of 4D occupancy forecasting on Occ3D-nuScenes.** Columns correspond to the horizons from 0 to 3 s. For each scene, the rows show (a) the surround-view images at the corresponding timestamps for reference, (b) the occupancy ground truth, and the forecasts of (c) SparseWorld-TC and (d) KIO-World.

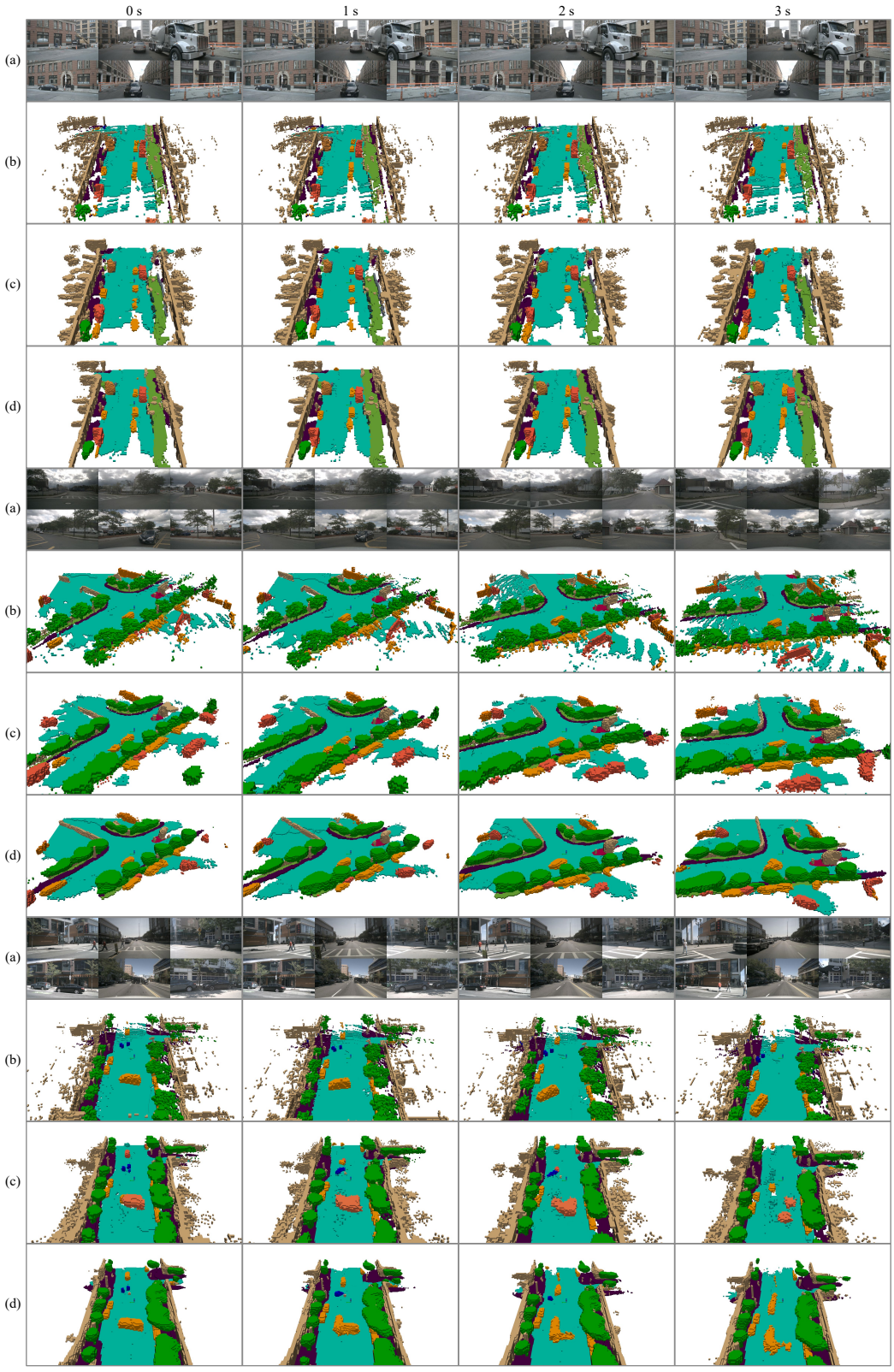

Figure 8: **Additional qualitative comparison of 4D occupancy forecasting on Occ3D-nuScenes.** The layout follows Fig. 7.

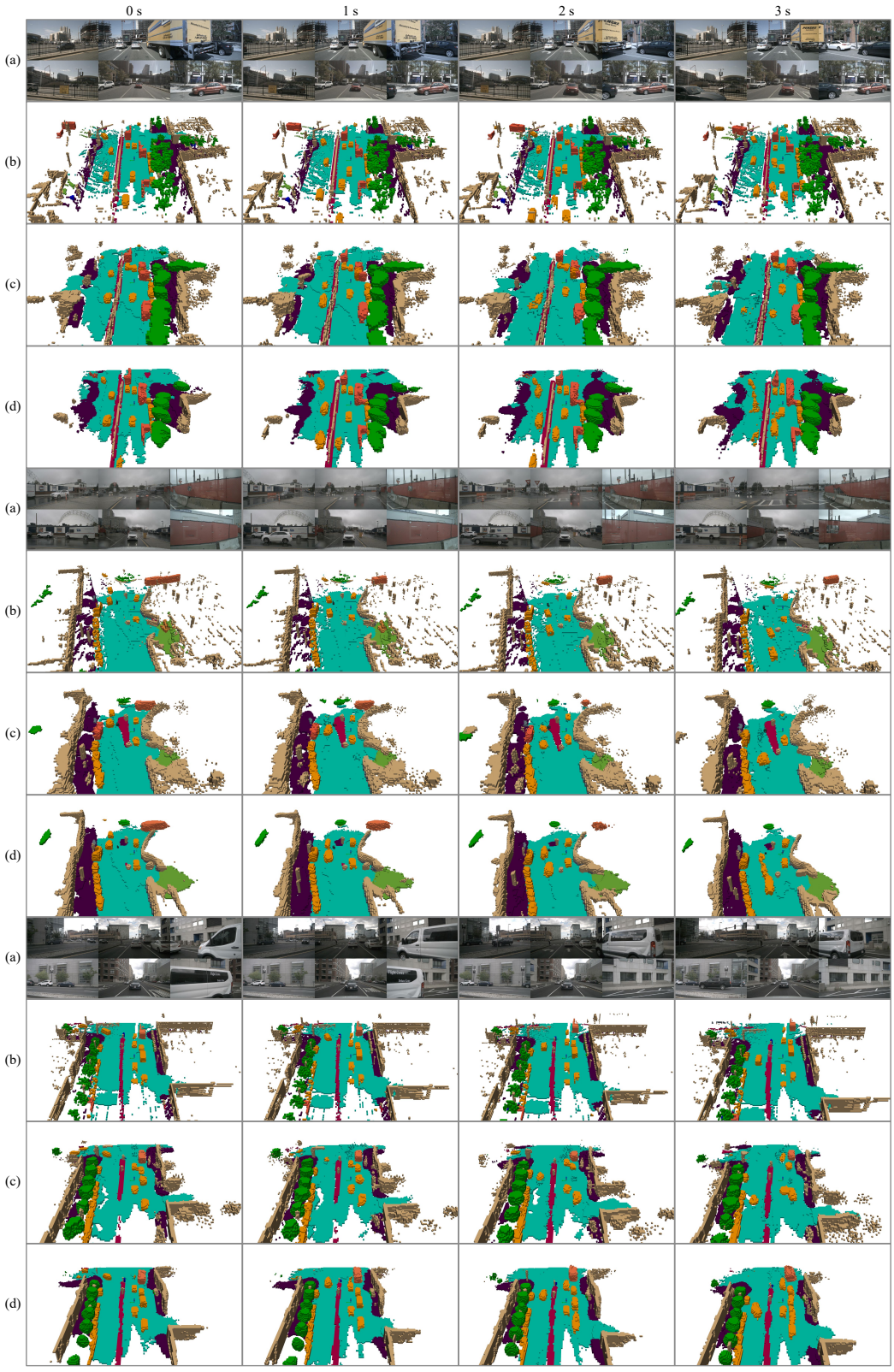

Figure 9: **Additional qualitative comparison of 4D occupancy forecasting on Occ3D-nuScenes.** The layout follows Fig. 7.