INFRAWORLD: MOTION-AWARE WORLD MODELING FOR ROADSIDE 4D OCCUPANCY FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Roadside 4D occupancy forecasting predicts future semantic occupancy in a fixed infrastructure coordinate system. Roadside scenes are globally structured yet locally evolving, requiring comprehensive scene-level context while precisely modeling sparse localized motion. However, existing forecasting paradigms struggle to tackle both requirements. To address this limitation, we propose **InfraWorld**, a world model formulating roadside occupancy forecasting as selective state evolution. InfraWorld maintains a dense BEV state for the complete scene and persistent motion queries for moving regions, assigning complementary states to global context and local dynamics. Specifically, we propose Dynamic Query Initiation (DQI) to initialize queries in regions supported by motion in the observed history. During rollout, the dense state is updated through Long–Short Temporal Fusion (LST), which combines short-term adjacent-frame continuity with long-term motion trends to preserve stable context while capturing scene evolution. Motion-Aware Query Injection (MAQI) then refines the queries and injects confidence-weighted query information into corresponding BEV tokens, enabling targeted dynamic updates without perturbing global context. Together, InfraWorld concentrates predictive capacity on evolving traffic regions. Extensive experiments on the InfraOcc and Occ3D–nus datasets demonstrate that InfraWorld achieves state-of-the-art performance for roadside 4D occupancy forecasting while retaining its effectiveness under ego-centric settings. Code will be released.

1 INTRODUCTION

Semantic occupancy represents free space, static infrastructure, and traffic participants in a unified 3D field (Cao & de Charette, 2022; Li et al., 2023; Huang et al., 2023; Zhang et al., 2023; Li et al., 2025). Extending this representation over time, 4D occupancy forecasting predicts future semantic occupancy from observed sequences (Ma et al., 2024; Zheng et al., 2024; Chen et al., 2025; Liao et al., 2025b) for planning, safety assessment, and simulation. However, existing studies have primarily focused on ego-centric settings. In contrast, roadside perception continuously observes a fixed traffic space with infrastructure-mounted sensors, often beyond the view of individual vehicles, supporting traffic monitoring and coordinated safety services. (Ye et al., 2022; Yu et al., 2022; Xiang et al., 2024; Hao et al., 2024; Zimmer et al., 2024). Recently, InfraOcc (Yang et al., 2026) established fixed-coordinate 3D occupancy prediction for roadside scenes, yet roadside 4D occupancy forecasting remains unexplored.

The fixed viewpoint gives roadside occupancy forecasting a distinctive spatiotemporal structure. As shown in Fig. 1(a), static infrastructure and stationary elements dominate the scene and remain spatially aligned over time, while moving participants occupy only a few localized regions yet account for most future changes. Consequently, roadside occupancy forecasting is globally structured yet locally evolving. On the one hand, forecasting a complete occupancy field requires scene-level context to preserve the overall spatial layout and provide a coherent global scaffold for future prediction. On the other hand, the actual scene evolution is determined by sparse local regions, whose continuous motion requires fine-grained and flexible modeling of their spatial displacement and temporal dynamics. An effective roadside world model therefore needs to retain comprehensive scene-level context while precisely modeling the sparse local motions that drive future changes, which raises a key question: *how should a world model reason when scene state and scene evolution have fundamentally different spatial support?*

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.

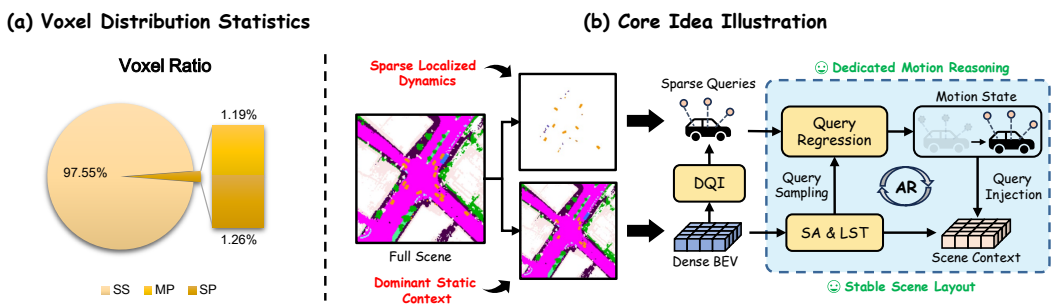


Figure 1: (a) Voxel distribution in InfraOcc: static scene (SS), moving participants (MP), and stationary participants (SP). (b) InfraWorld maintains the stable scene layout with a dense BEV state, while queries reason about localized dynamics and inject their updates into the scene context.

Existing ego-centric forecasting paradigms address this asymmetry only partially. Dense BEV-based methods (Zheng et al., 2024; Liao et al., 2025b; Yang et al., 2025) maintain a complete scene representation and effectively capture scene-level context, but their fixed-resolution grids evolve the scene uniformly, allowing dominant unchanged regions to dilute the sparse changes that drive future evolution. Flow-based methods (Zhang et al., 2024; Guo et al., 2026) explicitly model such changes by predicting displacement and warping occupancy or features, but primarily formulate scene evolution as spatial transport, which limits their ability to flexibly reason about the continuous evolution of localized traffic regions. Thus, existing formulations still struggle to preserve comprehensive scene context while devoting sufficient modeling capacity to the sparse local dynamics that determine future changes.

To reconcile comprehensive scene modeling with fine-grained reasoning over localized dynamics, we formulate roadside occupancy forecasting as selective state evolution, where persistent scene context and sparse scene changes are assigned complementary modeling roles. We instantiate this principle in **InfraWorld**, a roadside occupancy world model that maintains a dense BEV state for the spatially complete scene context while using a small set of persistent motion queries that flexibly reason about localized traffic evolution, as illustrated in Fig. 1(b). InfraWorld first employs Dynamic Query Initiation (DQI) to initialize queries where motion is likely to occur with motion evidence from the observed history. For temporal fusion, we develop Long–Short Temporal Fusion (LST) which combines recent continuity with longer-term change trends to model temporal evolution at complementary scales. Motion-Aware Query Injection (MAQI) then transfers reliable query-level dynamics to nearby BEV tokens, strengthening evolving regions without disrupting the stable scene context. With these designs, InfraWorld preserves complete roadside structure while focusing predictive capacity on localized dynamics. Our contributions are summarized below.

- We propose **InfraWorld**, the first framework for roadside 4D occupancy forecasting to the best of our knowledge, which formulates roadside forecasting as selective state evolution and assigns complementary modeling roles to spatially complete scene context and sparse localized dynamics.
- We introduce Dynamic Query Initiation (DQI) and Motion-Aware Query Injection (MAQI) to allocate dedicated motion reasoning to regions likely to evolve and selectively transfer reliable dynamic information back to the scene state.
- We develop Long–Short Temporal Fusion (LST), which combines recent-frame continuity with longer-term trends to strengthen temporal reasoning for scene evolution during rollout.
- Extensive experiments on InfraOcc and Occ3D–nus validate the effectiveness of InfraWorld for roadside 4D occupancy forecasting and its applicability to ego-centric settings.

2 Related Work

2.1 4D Occupancy World Models

4D occupancy world models forecast future semantic occupancy from observed scene history. Early methods predict LiDAR-derived occupancy through differentiable ray casting or point-cloud forecasting (Khurana et al., 2022; 2023), while Cam4DOcc (Ma et al., 2024) introduced a camera-only

benchmark and evaluation protocol. Subsequent approaches evolve compact latent scene representations. OccWorld (Zheng et al., 2024) combines VQ-VAE tokenization with autoregressive scene-token prediction, whereas RenderWorld (Yan et al., 2025) separately models empty and occupied voxels in latent space. Other studies incorporate textual scene knowledge (Wei et al., 2024; Xu et al., 2025), exploit 4D occupancy reconstruction for representation learning (Min et al., 2024; 2023), or condition generation on ego actions (Yang et al., 2025). For long-horizon forecasting, OccSora (Wang et al., 2026) and DOME (Gu et al., 2024) adopt diffusion-based generation, while I^2-World (Liao et al., 2025b) decouples spatial compression from temporal modeling. These methods primarily target ego-centric scenes, where ego motion and traffic dynamics jointly affect the global scene state. In contrast, InfraWorld operates in fixed roadside coordinates and assigns complementary roles to stable global context and localized dynamics.

2.2 Infrastructure-Side Occupancy and Motion-Aware Scene Modeling

Infrastructure-side occupancy forecasting operates in a fixed roadside coordinate system, where stationary sensors keep roads, buildings, and other infrastructure spatially aligned over time. As illustrated in Fig. 2, this stable background remains largely unchanged across consecutive occupancy frames, while temporal changes are concentrated in a small number of localized moving traffic participants. InfraOcc (Yang et al., 2026) formalizes this setting and highlights its static–dynamic asymmetry. Roadside forecasting must therefore preserve a spatially complete, temporally stable scene scaffold while focusing modeling capacity on the few regions that undergo change.

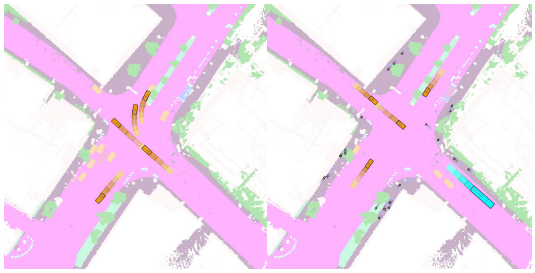


Figure 2: Sparse and localized dynamics in roadside occupancy. Across five consecutive frames, the static scene layout remains largely unchanged, while only a few moving traffic participants exhibit temporal changes.

Related motion-aware occupancy forecasting methods follow two main paradigms. Flow-based approaches explicitly model displacement: DFIT-OccWorld (Zhang et al., 2024) decouples static and dynamic voxels and warps dynamic occupancy with predicted flow, while FSF-Net (Guo et al., 2026) combines coarse BEV-flow forecasting with latent-state prediction. Query-based approaches (Dang et al., 2026) allocate computation selectively, representing scene evolution using adaptive sparse queries. Although developed for ego-centric scenes, these paradigms have complementary limitations for roadside forecasting. Flow-based methods primarily treat motion as spatial transport without maintaining a feature-rich persistent state for each moving region, whereas fully sparse query representations are less suited to preserving the spatially complete roadside background. InfraWorld addresses this trade-off by using dense BEV latents for global scene context and persistent motion-focused queries for localized dynamics.

3 Method

3.1 Overview

Given a sequence of H semantic occupancy observations $\mathcal{O}_{t-H+1:t} = \{O_{t-H+1}, \ldots, O_t\}$ in a fixed roadside coordinate system, InfraWorld predicts the subsequent F occupancy volumes:

$$\hat{\mathcal{O}}_{t+1:t+F} = \mathcal{F}_\theta(\mathcal{O}_{t-H+1:t}), \tag{1}$$

where θ denotes the model parameters. A pretrained and frozen VQ tokenizer $(\mathcal{E}, \mathcal{D})$ encodes O_s into a BEV latent $Z_s = \mathcal{E}(O_s) \in \mathbb{R}^{C \times H_z \times W_z}$ and decodes each forecast latent as $\hat{O}_s = \mathcal{D}(\hat{Z}_s)$.

InfraWorld autoregressively propagates a dense latent $\hat{Z}_\tau$ for the complete scene and a sparse query set Q_τ for predicted moving regions. Let $\mathcal{H}_\tau = [\mathbf{H}_\tau^{(1)}, \ldots, \mathbf{H}_\tau^{(L)}]$ denote a first-in–first-out queue of L dense latents ordered from oldest to newest. We initialize it from the L most recent observations as $\mathcal{H}_t = [Z_{t-L+1}, \ldots, Z_t]$, where $L \leq H$. After predicting $\hat{Z}_\tau$, we update the queue by

$$\mathcal{H}_\tau = [\mathbf{H}_{\tau-1}^{(2)}, \ldots, \mathbf{H}_{\tau-1}^{(L)}, \hat{Z}_\tau]. \tag{2}$$

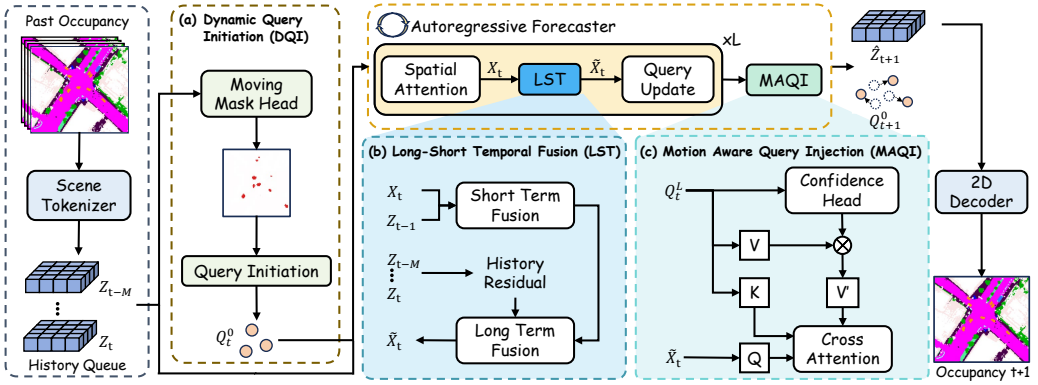


Figure 3: Architecture overview of InfraWorld. A frozen tokenizer encodes the observed occupancy into dense BEV latents, while DQI initializes persistent queries in predicted moving regions. At each autoregressive step, LST fuses spatial and temporal context to update the dense state and refine the queries, after which MAQI injects reliable query features into local BEV regions. The updated dense latent and queries are then propagated to the next forecasting step.

With $\hat{Z}_t = Z_t$ and Q_t initialized by Dynamic Query Initiation (DQI; Section 3.2), for $\tau = t+1, \ldots, t+F$, each forecasting step is

$$(\hat{Z}_\tau, Q_\tau) = \mathcal{W}_\theta(\hat{Z}_{\tau-1}, \mathcal{H}_{\tau-1}, Q_{\tau-1}). \tag{3}$$

As shown in Figure 3, $\mathcal{W}_\theta$ extracts multi-scale BEV features from $\hat{Z}_{\tau-1}$ and fuses them with $\mathcal{H}_{\tau-1}$ through Long–Short Temporal Fusion (LST; Section 3.3). The fused features recurrently refine the queries, after which Motion-Aware Query Injection (MAQI; Section 3.4) writes reliable query information into local BEV regions. A residual predictor then updates $\hat{Z}_\tau$ from $\hat{Z}_{\tau-1}$, and the new latent is appended to the history queue according to Eq. 2.

3.2 Dynamic Query Initiation

DQI initializes a limited set of persistent queries in motion-supported regions to provide dedicated representations for localized dynamics. Given the observed BEV latents $\{Z_s\}_{s=t-H+1}^{t}$, the query state propagated at rollout step τ is defined as

$$Q_\tau = \{(\mathbf{f}_i^\tau, \mathbf{P}_i^\tau, m_i)\}_{i=1}^{N_q}, \tag{4}$$

where N_q is the fixed query budget, $\mathbf{f}_i^\tau \in \mathbb{R}^C$ denotes the query feature, $\mathbf{P}_i^\tau \in [0,1]^{N_p \times 2}$ contains N_p normalized point coordinates, and $m_i \in \{0,1\}$ is the validity mask determined at initialization. The feature $\mathbf{f}_i^\tau$ encodes regional semantic and motion information, while $\mathbf{P}_i^\tau$ specifies where the query samples and updates local BEV features.

A lightweight moving-mask head fuses the current latent with the observed history to predict moving logits $\ell \in \mathbb{R}^{H_m \times W_m}$. As illustrated in Figure 4, we partition ℓ into non-overlapping tiles, rank them by their max-pooled logits, and select the N_q highest-scoring tiles.

For each selected tile $\mathcal{T}_i$, its anchor is obtained by soft localization:

$$\mathbf{a}_i = \sum_{u \in \mathcal{T}_i} \frac{\exp(\ell_u / T)}{\sum_{v \in \mathcal{T}_i} \exp(\ell_v / T)} \mathbf{x}_u, \tag{5}$$

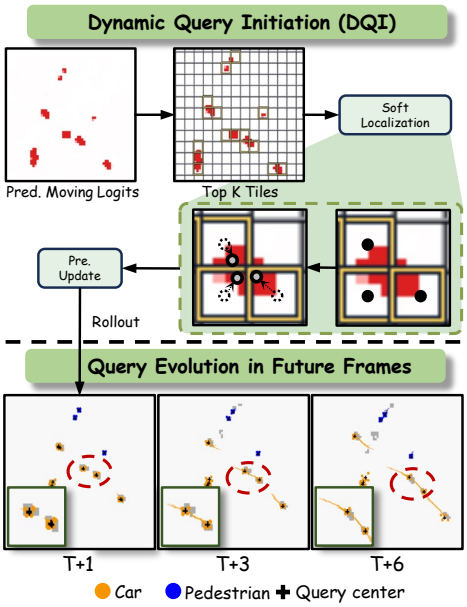


Figure 4: Dynamic Query Initiation (DQI). Top: DQI process for selecting and localizing motion-supported queries. Bottom: query evolution during rollout, where gray regions denote GT moving objects and highlighted trajectories show how queries capture localized motion evolution.

where $\mathbf{x}_u$ is the normalized center of cell u and $T > 0$ is the temperature. A query is marked invalid if the mean moving probability of its tile falls below a threshold. Tile-wise selection distributes the limited query budget across candidate motion regions, while soft localization places each anchor according to the intra-tile motion evidence.

For each valid query, the initial feature is formed by adding a learned query embedding to the current latent feature bilinearly sampled at $\mathbf{a}_i$, and is then refined using the observed history. Following SparseWorld (Dang et al., 2026), the coarse-to-fine refinement samples local history features around the query points, performs query-conditioned mixing across sampled points and channels, exchanges information among queries through self-attention, and predicts point semantics together with an expanded point set for the next stage. Invalid queries retain their states but are masked from query interactions. The resulting queries constitute the persistent motion states that are recurrently updated during forecasting.

3.3 Long–Short Temporal Fusion

LST is designed to jointly capture immediate temporal continuity and accumulated motion trends during autoregressive forecasting. At rollout step τ, the dense BEV backbone progressively applies spatial self-attention to $\hat{Z}_{\tau-1}$ at multiple BEV resolutions. The resulting multi-resolution features are then upsampled and fused to obtain the BEV feature $\mathbf{X}_\tau$.

For short-term temporal modeling, let $\bar{\mathbf{H}}_{\tau-1}^{(L-1)}$ denote the preceding history feature with its temporal embedding. Since the roadside coordinate system remains fixed over time, it can be directly fused with $\mathbf{X}_\tau$ through a ConvGRU-style update:

$$\mathbf{z}_\tau = \sigma(\mathrm{Conv}_z([\mathbf{X}_\tau, \bar{\mathbf{H}}_{\tau-1}^{(L-1)}])), \tag{6}$$

$$\mathbf{r}_\tau = \sigma(\mathrm{Conv}_r([\mathbf{X}_\tau, \bar{\mathbf{H}}_{\tau-1}^{(L-1)}])), \tag{7}$$

$$\mathbf{S}_\tau = \mathrm{Conv}_o\Big((1-\mathbf{z}_\tau)\odot\mathbf{X}_\tau + \mathbf{z}_\tau\odot\phi(\mathrm{Conv}_h([\mathbf{X}_\tau, \mathbf{r}_\tau\odot\bar{\mathbf{H}}_{\tau-1}^{(L-1)}]))\Big). \tag{8}$$

Here, $[\cdot,\cdot]$ denotes channel-wise concatenation, $\odot$ denotes element-wise multiplication, and ϕ is the candidate activation function. The update gate adaptively balances the current feature and history-conditioned information, while the reset gate suppresses irrelevant historical content, yielding the short-term state $\mathbf{S}_\tau$.

To capture longer-term motion trends, the long-term branch operates on the raw history queue $\mathcal{H}_{\tau-1}$. It first computes adjacent residuals $\Delta\mathbf{H}_{\tau-1}^{(j)} = \mathbf{H}_{\tau-1}^{(j+1)} - \mathbf{H}_{\tau-1}^{(j)}$ for $j = 1, \ldots, L-1$. A temporal Conv1D is then applied independently at each BEV location, followed by temporal averaging and a 3×3 spatial aggregation layer to produce the trend feature $\mathbf{R}_\tau$. This feature refines the short-term state as

$$\widetilde{\mathbf{X}}_\tau = \mathbf{S}_\tau + \mathcal{P}_l([\mathbf{S}_\tau, \mathbf{R}_\tau]), \tag{9}$$

where $\mathcal{P}_l$ denotes a lightweight projection whose final layer is zero-initialized. The refined feature $\widetilde{\mathbf{X}}_\tau$ is subsequently used to update the dynamic queries and predict the next dense latent state. In this way, LST combines local temporal continuity with accumulated motion trends to provide a temporally informed dense representation for both query refinement and global scene prediction.

3.4 Motion-Aware Query Injection

MAQI transfers query-level motion information to local BEV regions while suppressing unreliable updates during autoregressive rollout. At each forecasting step, the queries are recurrently refined from the current BEV features with a fixed number of query points.

Autoregressive propagation may introduce errors in query localization and semantics. We therefore estimate a confidence score $c_i^\tau \in [0,1]$ for each valid query before injection. As shown in Figure 5, the point-wise class log-probabilities are embedded and aggregated by average and max pooling into a semantic summary $\mathbf{s}_i^\tau$. The confidence head combines this summary with the normalized query feature:

$$c_i^\tau = \sigma(\mathcal{C}[\mathrm{LN}(\mathbf{f}_i^\tau), \mathbf{s}_i^\tau]), \tag{10}$$

where $\mathcal{C}$ is a lightweight MLP. The confidence target, defined in Section 3.5, jointly considers localization and semantic agreement.

At the final forecasting layer, the LST output $\widetilde{\mathbf{X}}_\tau$ is reshaped into BEV tokens $\mathbf{B}_\tau \in \mathbb{R}^{H_z W_z \times C}$, with $\mathbf{q}_\tau^{\text{bev}} = \mathbf{B}_\tau + \mathbf{E}_{\text{bev}}$. For each query, its feature, point-set centroid, relative point offsets, and dynamic class distribution are combined into a query token, from which $\mathbf{k}_i^\tau$ and $\mathbf{v}_i^\tau$ are generated. The BEV tokens then cross-attend to the query keys and confidence-weighted values:

$$\mathbf{B}'_\tau = \mathbf{B}_\tau + \mathcal{P}_o\Big(\text{MHA}\left(\mathbf{q}_\tau^{\text{bev}}, \{\mathbf{k}_i^\tau\}_{i=1}^{N_q}, \{c_i^\tau \mathbf{v}_i^\tau\}_{i=1}^{N_q}; \mathbf{b}_\tau\right)\Big), \tag{11}$$

where $\mathbf{b}_\tau$ combines a Gaussian centroid-distance bias with the invalid-query mask, and $\mathcal{P}_o$ is a zero-initialized output projection. The distance bias localizes each query's influence to nearby BEV tokens, while c_i^τ suppresses unreliable query updates. Finally, the dense latent is updated through a zero-initialized residual predictor:

$$\hat{Z}_\tau = \hat{Z}_{\tau-1} + \mathcal{P}_r(\text{reshape}(\mathbf{B}'_\tau)). \tag{12}$$

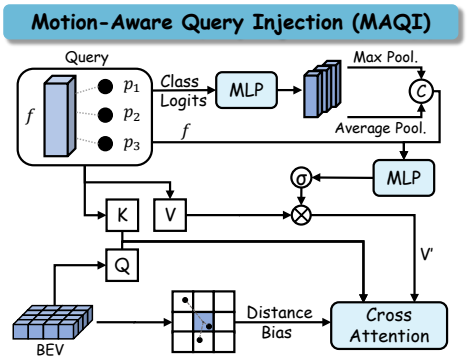


Figure 5: Motion-Aware Query Injection (MAQI). MAQI injects confidence-weighted query features into neighboring BEV tokens through distance-biased cross-attention.

3.5 Learning Objectives

We supervise the dense rollout by aligning each predicted BEV latent with the corresponding latent produced by the frozen tokenizer:

$$\mathcal{L}_{\text{latent}} = \sum_{k=1}^{F} \alpha_k \,\text{MSE}(\hat{Z}_{t+k}, Z_{t+k}), \tag{13}$$

where $\alpha_k \geq 0$ weights forecast horizon k. To guide DQI toward moving regions, the moving-mask head is supervised by a binary BEV target whose positive cells belong to dynamic semantic classes and have flow magnitude above a motion threshold:

$$\mathcal{L}_{\text{moving}} = \text{BCE}(\hat{\mathbf{M}}, \mathbf{M}) + \text{Dice}(\hat{\mathbf{M}}, \mathbf{M}), \tag{14}$$

where $\hat{\mathbf{M}}$ and $\mathbf{M}$ denote the predicted probability map and binary target, respectively. On InfraOcc, flow is generated offline from annotated 3D boxes and velocities as detailed in the appendix, and is used only after thresholding to identify moving occupancy cells. The model neither takes flow as input nor directly predicts the flow vectors.

For query supervision, moving BEV cells are converted into metric point sets with semantic labels. For valid queries, $\mathcal{L}_{\text{point}}$ uses a class-weighted bidirectional nearest-neighbor Smooth-L_1 loss for point alignment, while $\mathcal{L}_{\text{cls}}$ applies class-balanced focal loss to point semantics. Both losses supervise the pre-rollout refinement stages and future query updates. We further supervise the query confidence used by MAQI. Let d_{ip} denote the metric distance from point p of query i to its nearest ground-truth moving point, and $s_{ip} \in \{0, 1\}$ indicates semantic agreement. The confidence target is

$$c_i^* = \frac{1}{N_p} \sum_{p=1}^{N_p} \exp\left(-\frac{d_{ip}^2}{2\sigma_r^2}\right) s_{ip}, \tag{15}$$

where $\sigma_r > 0$ controls the localization tolerance, and the confidence loss is $\mathcal{L}_{\text{rel}} = \text{BCE}(c_i, c_i^*)$.

The overall training objective is

$$\mathcal{L} = \lambda_l \mathcal{L}_{\text{latent}} + \lambda_m \mathcal{L}_{\text{moving}} + \lambda_p \mathcal{L}_{\text{point}} + \lambda_c \mathcal{L}_{\text{cls}} + \lambda_r \mathcal{L}_{\text{rel}}. \tag{16}$$

4 Experiments

4.1 Experimental Setup

Datasets and metrics. We evaluate our method on InfraOcc (Yang et al., 2026) and Occ3D–nus (Tian et al., 2023). InfraOcc consists of 215 training sequences and 75 evaluation sequences

Table 1: 4D occupancy forecasting performance on InfraOcc validation set.

Method	Input	mIoU_{dyn} (%) ↑				mIoU_{sta} (%) ↑			
		1 s	2 s	3 s	Avg.	1 s	2 s	3 s	Avg.
OccWorld-O (ECCV 2024)	3D-Occ	25.64	21.24	19.76	22.21	97.52	97.42	97.37	97.44
I^2-World-O (ICCV 2025)	3D-Occ	33.34	29.67	28.66	30.56	97.30	97.27	97.07	97.21
InfraWorld-O (Ours)	3D-Occ	**46.03**	**38.08**	**34.63**	**39.58**	**97.79**	**97.72**	**97.68**	**97.73**
OccWorld-STC (ECCV 2024)	Camera	16.20	14.29	13.69	14.73	78.29	78.27	78.24	78.27
I^2-World-STC (ICCV 2025)	Camera	19.47	16.95	15.73	17.38	78.71	78.70	**78.69**	78.70
SparseWorld (AAAI 2026)	Camera	8.52	8.28	8.16	8.32	73.67	73.80	73.12	73.53
InfraWorld-STC (Ours)	Camera	**22.52**	**20.73**	**20.43**	**21.23**	**78.75**	**78.71**	78.68	**78.71**

collected in roadside scenarios, with occupancy grids of $320 \times 320 \times 16$ voxels at a resolution of 0.4 m. We report mIoU_{dyn} and mIoU_{sta} over the dynamic and static categories, respectively. Occ3D–nus is a 4D occupancy forecasting benchmark built upon nuScenes, comprising 600 training scenes and 150 validation scenes. Its occupancy range spans $[-40, 40]$ m along both the x- and y-axes and $[-1, 5.4]$ m along the z-axis, with a voxel size of 0.4 m and a sampling rate of 2 Hz. On Occ3D–nus, we report mIoU at 1, 2, and 3 s, together with the average over the three horizons.

Implementation details. We use 2 seconds of historical occupancy observations to forecast the subsequent 3 seconds. The tokenizer is pretrained for 24 epochs and subsequently frozen. InfraWorld-O is trained with ground truth 3D occupancy and InfraWorld-STC is trained with predictions from STCOcc (Liao et al., 2025a). We train the model for 24 epochs on InfraOcc and 48 epochs on Occ3D–nus, using AdamW with an initial learning rate of 1×10^{-3}, a weight decay of 10^{-2}, and a batch size of 8 per GPU on four NVIDIA RTX 4090 GPUs. All inference results are measured on a single NVIDIA RTX 4090 GPU with 24GB of memory.

4.2 Main Results

Table 2: Forecasting performance on Occ3D–nus.

Method	Input	mIoU (%) ↑			
		1 s	2 s	3 s	Avg.
OccWorld-O	Occ	25.78	15.14	10.51	17.14
OccLLaMA-O	Occ & Text	25.05	19.49	15.26	19.93
DFIT-OccWorld	Occ & Camera	31.68	21.29	15.19	22.71
DOME	Occ	35.11	25.89	20.29	27.10
UniScene	Occ	35.37	29.59	25.08	31.76
I^2-World-O	Occ	47.62	38.58	32.98	39.73
InfraWorld-O	Occ	**50.34**	**41.16**	**35.25**	**42.25**
OccWorld-STC	Camera	15.16	10.97	10.19	12.11
SparseWorld	Camera	14.93	13.15	11.51	13.20
I^2-World-STC	Camera	19.60	16.71	14.48	16.93
InfraWorld-STC	Camera	**20.56**	**18.50**	**16.61**	**18.56**

Table 1 presents the main comparison on InfraOcc. With occupancy inputs, InfraWorld-O achieves an average mIoU_{dyn} of 39.58%, outperforming I^2-World-O by 9.02 points, and retains a 5.97-point advantage at the 3 s horizon. This improvement validates our key insight that persistent motion queries can complement the dense scene state by providing dedicated modeling capacity for the sparse localized regions that drive future evolution. Meanwhile, InfraWorld-O maintains a high average mIoU_{sta} of 97.73%, indicating that the dense pathway effectively preserves the stable scene layout while the query pathway focuses on evolving regions.

Table 2 further evaluates InfraWorld on Occ3D–nus, an ego-centric benchmark with scene dynamics distinct from the roadside setting. InfraWorld-O and InfraWorld-STC achieve average mIoUs of 42.25% and 18.56%, respectively, consistently outperforming their I^2-World counterparts across all forecasting horizons by 2.52 and 1.63 points on average. These results show that the proposed dense scene state and focused motion reasoning remain effective for ego-centric occupancy forecasting.

Figure 6 provides a qualitative comparison with I^2-World. Both methods largely preserve the static roadside layout, whereas the differences are more evident in localized regions containing moving participants. At the longer 3 s horizon, I^2-World tends to miss or distort the occupancy of moving objects, while InfraWorld better preserves their spatial locations and shapes, producing predictions that more closely match the ground truth in the highlighted regions. This behavior is consistent with our design, where LST integrates recent temporal continuity and longer-term motion trends to

provide more informative temporal cues, while persistent motion queries recurrently refine localized motion states throughout the rollout.

Together with the dense scene state that preserves the stable roadside context, these components enable more accurate prediction of localized traffic evolution over longer horizons.

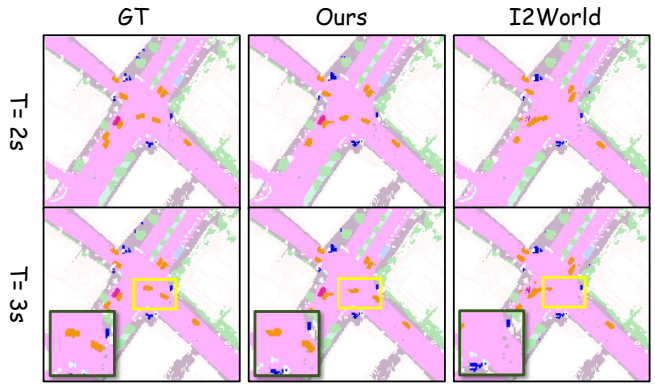


Figure 6: Qualitative comparison between InfraWorld and I^2-World on InfraOcc. Yellow boxes highlight dynamic regions where InfraWorld more closely matches the ground truth.

4.3 Ablation Studies

Overall ablation. Table 3 evaluates the contribution of the principal components of InfraWorld. For variants with MAQI but without DQI, we initialize one query at the center of every tile. Starting from the dense forecasting baseline, introducing LST improves the average mIoU$_{\text{dyn}}$ by 1.94 points. By combining recent-frame continuity with accumulated changes over the observed history, LST captures both the immediate state of local motion and its longer-term evolution trend, providing stronger temporal cues for forecasting where and how the scene will change. MAQI produces a larger gain of 3.42 points, supporting our central design that persistent motion queries provide complementary information for localized regions whose dynamics are difficult to capture through the dense state alone. When DQI is further introduced, mIoU$_{\text{dyn}}$ increases by another 2.40 points, indicating that assigning the limited query budget according to observed motion evidence makes this localized reasoning more effective. Combining all three components yields 39.58% mIoU$_{\text{dyn}}$, an 8.65-point improvement over the baseline. These results validate that preserving a complete dense scene state while allocating dedicated modeling capacity to sparse evolving regions is effective for roadside occupancy forecasting.

Table 3: Component-wise ablation of InfraWorld.

LST	MAQI	DQI	mIoU$_{\text{dyn}}$ (%) ↑			
			1 s	2 s	3 s	Avg.
–	–	–	33.68	32.54	26.58	30.93
✓	–	–	35.20	32.95	30.46	32.87
–	✓	–	36.81	34.33	31.91	34.35
–	✓	✓	42.76	34.87	32.62	36.75
✓	✓	✓	**46.03**	**38.08**	**34.63**	**39.58**

Ablation on LST. Table 4 examines the two temporal branches in LST. Removing either the short-term or long-term branch degrades performance, with drops of 1.33 and 1.48 points in average mIoU$_{\text{dyn}}$, respectively. The short-term branch exploits the strong temporal continuity between adjacent frames under the fixed roadside coordinate system, while the long-term branch summarizes residual changes across the history to capture accumulated motion trends. Their comparable contributions indicate that recent temporal continuity and longer-term motion trends capture complementary aspects of scene evolution, and jointly provide a richer temporal representation for forecasting dynamic regions. Therefore, their combination provides a more suitable temporal representation for both dense-state prediction and subsequent query refinement.

Ablation on MAQI. We further compared different strategies for writing query information into the dense BEV state. Replacing MAQI with bilinear splatting reduces the average mIoU$_{\text{dyn}}$ from 39.58% to 37.78%. Unlike fixed spatial interpolation, the distance-biased cross-attention in MAQI allows each query to interact with nearby BEV features according to the local scene context, providing a more adaptive mechanism for localized updates. Removing confidence weighting further decreases performance to 38.18%. This confirms the importance of filtering query updates during autoregressive rollout, where localization and semantic errors can accumulate over time. By suppressing unreliable writeback, MAQI allows the query pathway to refine evolving regions without unnecessarily disturbing the dense scene state. The qualitative results in Figure 7 show that introducing dynamic queries produces more accurate forecasts of moving objects. This improvement is consistent with persistent queries maintaining and recurrently updating localized motion states throughout the rollout, providing dedicated motion cues for predicting future occupancy changes in these localized regions.

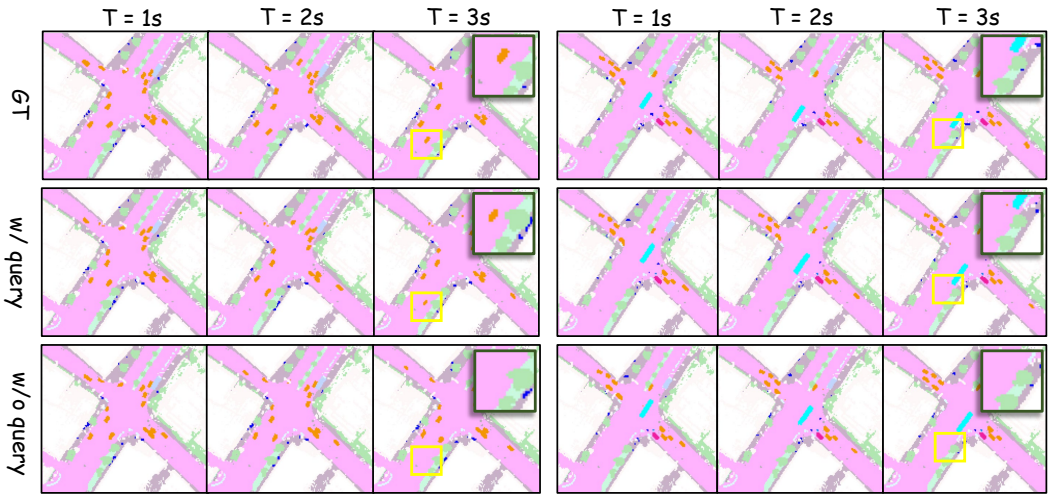


Figure 7: Qualitative comparison of occupancy forecasts with and without dynamic queries. Yellow boxes highlight moving regions where dynamic queries improve agreement with the GT.

Table 4: Joint ablation of LST and MAQI on InfraOcc.

Fusion method	Writeback method	mIoU_{dyn} (%) ↑			
		1 s	2 s	3 s	Avg.
w/o short-term fusion	MAQI	44.06	36.67	34.03	38.25
w/o long-term fusion	MAQI	45.02	36.18	33.09	38.10
LST	MAQI	**46.03**	**38.08**	**34.63**	**39.58**
LST	Bilinear Splat	42.83	36.18	34.34	37.78
LST	MAQI w/o conf.	45.11	36.32	33.11	38.18

4.4 Model Efficiency

Table 5 compares parameters, FLOPs, inference latency and forecasting accuracy. InfraWorld introduces a moderate increase in computation over I^2-World while improving mIoU_{dyn} by 9.02 points. To isolate the effect of model capacity, we further enlarge I^2-World to I^2-World-L with a parameter count comparable to InfraWorld. Despite similar model size and FLOPs, I^2-World-L still trails InfraWorld by 6.68 points, indicating that the improvement does not simply arise from increased model capacity but from the targeted designs for roadside scene evolution. DQI allocates persistent queries to motion-supported regions, LST integrates recent continuity with longer-term motion trends, and MAQI injects reliable query-level dynamics into the corresponding local BEV regions. InfraWorld also remains more efficient than OccWorld, achieving a favorable accuracy–efficiency trade-off.

Table 5: Efficiency–accuracy comparison. Lat: bs=1 inference latency.

Method	Params. (M)	FLOPs (G)	Lat. (ms)	mIoU_{dyn} ↑
OccWorld	71.1	1291	162	22.21
I^2-World	22.5	91	85	30.56
I^2-World-L	28.9	169	91	32.90
InfraWorld	29.1	166	119	39.58

5 Conclusion

In this work, we formulate roadside occupancy forecasting as selective state evolution and propose **InfraWorld**, which maintains a dense BEV state for complete scene context while using persistent motion queries to model sparse localized dynamics. DQI allocates queries to motion-supported regions, LST integrates recent continuity with longer-term motion trends, and MAQI selectively injects reliable query-level dynamics into local BEV regions. Together, these designs preserve the stable roadside layout while concentrating predictive capacity on the regions that drive future evolution. Experiments on InfraOcc and Occ3D–nus validate the effectiveness of InfraWorld for roadside forecasting and demonstrate its broader applicability to ego-centric settings.

AI USE STATEMENT

In this work, we used generative AI tools to assist with code development, including drafting and debugging auxiliary code, and to improve the presentation of the manuscript, including language polishing and table formatting. All AI-assisted content was carefully reviewed by the authors. AI-assisted code was inspected and tested for correctness before being used in our experiments. Language revisions were manually checked to ensure that they preserved the intended technical meaning and did not alter the scientific claims. AI-assisted table formatting was verified against the original experimental results to ensure numerical accuracy and consistency. We take responsibility for the final content of this work, including text, claims, code, and artifacts produced with the aid of generative AI.

REFERENCES

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 (CVPR)*, pp. 3991–4001, 2022.

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 *Proceedings of the International Conference on Learning Representations (ICLR)*, pp. 88425–88440, 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. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 3497–3505, 2026.

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:2410.10429*, 2024.

Erxin Guo, Pei An, You Yang, Qiong Liu, and An-An Liu. FSF-Net: Enhance 4D Occupancy Forecasting with Coarse BEV Scene Flow for Autonomous Driving. *Pattern Recognition*, pp. 112372, 2026.

Ruiyang Hao, Siqi Fan, Yingru Dai, Zhenlin Zhang, Chenxi Li, Yuntian Wang, Haibao Yu, Wenxian Yang, Jirui Yuan, and Zaiqing Nie. RCooper: A Real-World Large-Scale Dataset for Roadside Cooperative Perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22347–22357, 2024.

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 (CVPR)*, pp. 9223–9232, 2023.

Tarasha Khurana, Peiyun Hu, Achal Dave, Jason Ziglar, David Held, and Deva Ramanan. Differentiable Raycasting for Self-supervised Occupancy Forecasting. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 353–369, 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 (CVPR)*, pp. 1116–1124, 2023.

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 (CVPR)*, pp. 11971–11981, 2025.

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 (CVPR)*, pp. 9087–9098, 2023.

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 (CVPR)*, pp. 1516–1526, June 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 (ICCV)*, pp. 25810–25819, 2025b.

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 (CVPR)*, pp. 21486–21495, 2024.

Chen Min, Dawei Zhao, Liang Xiao, Yiming Nie, and Bin Dai. UniWorld: Autonomous Driving Pre-training via World Models. *arXiv:2308.07234*, 2023.

Chen Min, Dawei Zhao, Liang Xiao, Jian Zhao, Xinli Xu, Zheng Zhu, Lei Jin, Jianshu Li, Yulan Guo, Junliang Xing, et al. 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 (CVPR)*, pp. 15522–15533, 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 *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

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, 2026.

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:2409.03272*, 2024.

Hao Xiang, Zhaoliang Zheng, Xin Xia, Runsheng Xu, Letian Gao, Zewei Zhou, Xu Han, Xinkai Ji, Mingxi Li, Zonglin Meng, Li Jin, Mingyue Lei, Zhaoyang Ma, Zihang He, Haoxuan Ma, Yunshuang Yuan, Yingqian Zhao, and Jiaqi Ma. V2X-Real: A Large-Scale Dataset for Vehicle-to-Everything Cooperative Perception. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 455–470, 2024.

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 *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 8434–8441, 2025.

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

Lei Yang, Xiaokai Bai, Boqi Li, Chunmian Lin, Li Wang, Ziying Song, Jiahuan Zhang, Enhui Ma, Haibao Yu, Jiaqi Ma, and Kaicheng Yu. InfraOcc: An Infrastructure Occupancy Benchmark with Static-to-Dynamic Reasoning. *arXiv:2608.30657*, 2026.

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

Xiaoqing Ye, Mao Shu, Hanyu Li, Yifeng Shi, Yingying Li, Guangjie Wang, Xiao Tan, and Errui Ding. Rope3D: The Roadside Perception Dataset for Autonomous Driving and Monocular 3D Object Detection Task. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21341–21350, 2022.

Haibao Yu, Yizhen Luo, Mao Shu, Yiyi Huo, Zebang Yang, Yifeng Shi, Zhenglong Guo, Hanyu Li, Xing Hu, Jirui Yuan, and Zaiqing Nie. DAIR-V2X: A Large-Scale Dataset for Vehicle-Infrastructure Cooperative 3D Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21361–21370, 2022.

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: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 (ICCV)*, 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 55–72, 2024.

Walter Zimmer, Gerhard Arya Wardana Wardana, Suren Sritharan, Xingcheng Zhou, Rui Song, and Alois C. Knoll. TUMTraf V2X Cooperative Perception Dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22668–22677, 2024.

A Appendix

A.1 Loss Calculation

We provide the implementation details of the training objectives introduced in Section 3.5. Losses are averaged over valid samples, while future frames without valid annotations are excluded. Throughout this section, $\mathcal{D}$ denotes the set of dynamic semantic classes.

Dense latent loss. For each forecast step k, we align the predicted latent $\hat{Z}_{t+k}$ with the target latent Z_{t+k} produced by the frozen tokenizer:

$$\mathcal{L}^{k}_{\text{latent}} = \frac{1}{CH_zW_z} \left\| \hat{Z}_{t+k} - Z_{t+k} \right\|^2_F . \tag{17}$$

In Eq. 17, C, H_z, and W_z denote the channel, height, and width of the BEV latent, respectively, and $\| \cdot \|_F$ is the Frobenius norm.

Moving-mask loss. Motion supervision is derived jointly from occupancy semantics and occupancy flow. Let y_{uvz} and $\mathbf{f}_{uvz} \in \mathbb{R}^2$ denote the semantic label and flow vector of voxel (u, v, z), respectively. A BEV cell is marked as moving if it contains at least one dynamic voxel whose flow magnitude exceeds τ_m:

$$M_{uv} = \mathbb{I}[\exists z : \; y_{uvz} \in \mathcal{D} \; \wedge \; \|\mathbf{f}_{uvz}\|_2 > \tau_m] . \tag{18}$$

We use max pooling to downsample the binary target to the resolution of the moving-mask head, preserving any motion evidence within each pooling region. Let A_{uv} be the predicted logit, $P_{uv} = \sigma(A_{uv})$ its corresponding moving probability, and Ω the set of BEV cells. The moving-mask loss is

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{|\Omega|} \sum_{(u,v)\in\Omega} [\eta M_{uv} \log P_{uv} + (1 - M_{uv}) \log(1 - P_{uv})] , \tag{19}$$

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2 \sum_\Omega P_{uv} M_{uv} + 1}{\sum_\Omega P_{uv} + \sum_\Omega M_{uv} + 1}, \tag{20}$$

$$\mathcal{L}_{\text{moving}} = \mathcal{L}_{\text{BCE}} + \mathcal{L}_{\text{Dice}}. \tag{21}$$

To account for the sparsity of moving cells, the positive-class factor is $\eta = \text{clip}(N_-/N_+, 1, 8)$, where N_+ and N_- are the numbers of positive and negative cells, respectively. The motion threshold is set to $\tau_m = 0.8$ in all experiments.

Query point loss. To supervise query locations, we convert each positive cell in Eq. 18 to its center in metric BEV coordinates. These centers form the ground-truth point set $\mathcal{G}$, while $\mathcal{P}$ contains the valid points predicted by the dynamic queries. We measure their alignment in both directions:

$$\mathcal{L}_{\text{point}} = \mathcal{D}(\mathcal{P}, \mathcal{G}) + \mathcal{D}(\mathcal{G}, \mathcal{P}). \tag{22}$$

The directed distance $\mathcal{D}(\mathcal{A}, \mathcal{B})$ averages the Smooth-L_1 distance from each point in $\mathcal{A}$ to its nearest neighbor in $\mathcal{B}$, weighted by the semantic class of the matched ground-truth point. We use a transition parameter of $\beta = 1$. The two directions penalize unsupported predictions and missed moving regions, respectively.

Query classification loss. The semantic target of each predicted point is obtained from a nearby ground-truth point. For this assignment, we divide the Euclidean distance by the square root of the normalized class weight, reducing the bias toward frequent dynamic classes. Let c_a denote the assigned class of predicted point a, and let π_{a,c_a} be its predicted probability for that class. We apply the class-balanced focal loss

$$\mathcal{L}_{\text{cls}} = -\frac{\sum_a w_{c_a} (1 - \pi_{a,c_a})^\gamma \log \pi_{a,c_a}}{\sum_a w_{c_a}}. \tag{23}$$

The probability π_{a,c_a} is obtained by applying softmax to the point-wise semantic logits, and w_{c_a} is the weight of class c_a. The focal focusing parameter is set to $\gamma = 2$.

Query confidence loss. The confidence target measures both localization quality and semantic agreement. For query i, let d_{ip} be the metric distance from its p-th point to the nearest ground-truth

Table 6: Per-class IoU (%) of InfraWorld on InfraOcc at different forecasting horizons.

	Dynamic classes						Static classes								Free
Horizon	Bicycle	Bus	Car	Motorcycle	Pedestrian	Truck	Other	Barrier	Cone	Driveable	Sidewalk	Terrain	Manmade	Vegetation	Free
1 s	41.41	67.01	58.49	18.12	35.68	55.47	96.65	90.73	98.44	98.03	99.40	99.48	99.85	99.73	99.65
2 s	39.53	51.60	47.89	12.28	27.53	49.62	96.32	90.63	98.37	97.92	99.38	99.55	99.85	99.75	99.58
3 s	40.06	46.61	42.68	9.58	23.11	45.76	96.09	90.53	98.35	97.90	99.38	99.58	99.86	99.77	99.55

Table 7: Per-class IoU (%) of InfraWorld on Occ3D–nus at different forecasting horizons.

	Dynamic classes							Static classes								Free
Horizon	Bicycle	Bus	Car	Motorcycle	Pedestrian	Trailer	Truck	Other	Barrier	Cone	Driveable	Sidewalk	Terrain	Manmade	Vegetation	Free
1 s	41.22	47.97	48.33	38.37	38.78	48.57	51.82	50.29	59.36	41.98	67.28	56.52	54.76	48.03	44.16	96.98
2 s	28.81	31.75	35.40	27.14	22.43	38.57	39.69	43.53	49.87	31.44	61.97	50.49	48.12	42.84	40.22	96.63
3 s	21.58	23.28	28.76	20.89	13.57	31.78	32.80	38.97	43.05	24.62	57.86	45.98	43.00	38.71	36.83	96.40

moving point. We set $s_{ip} = 1$ if their semantic classes agree and $s_{ip} = 0$ otherwise, yielding

$$c_i^* = \frac{1}{N_p} \sum_{p=1}^{N_p} \exp\left(-\frac{d_{ip}^2}{2\sigma_r^2}\right) s_{ip}, \tag{24}$$

where N_p is the number of points carried by query i, and σ_r controls the tolerance to localization error. Given the predicted confidence logit r_i, we optimize

$$\mathcal{L}_{\text{rel}} = \frac{1}{|\mathcal{Q}_v|} \sum_{i \in \mathcal{Q}_v} \text{BCEWithLogits}(r_i, c_i^*), \tag{25}$$

where $\mathcal{Q}_v$ is the set of valid queries. We detach c_i^* during optimization so that the reliability loss trains only the confidence head and does not introduce an additional gradient path to the query coordinates or semantic predictions. We set $\sigma_r = 2.0$ meters.

Loss weights. The losses are combined according to Eq. 16. The weights for $\mathcal{L}_{\text{latent}}$, $\mathcal{L}_{\text{moving}}$, $\mathcal{L}_{\text{point}}$, $\mathcal{L}_{\text{cls}}$, and $\mathcal{L}_{\text{rel}}$ are set empirically to keep the weighted loss terms at comparable numerical scales during training. We use 10, 0.1, 0.001, 0.02, and 0.02, respectively. The six future steps are weighted by $(1.00, 0.85, 0.70, 0.55, 0.40, 0.25)$, with the same schedule used for the query point, classification, and confidence losses during rollout. Pre-rollout query losses use unit weight. Point and classification supervision is applied to all pre-rollout refinement layers and to the final refinement layer of each rollout step, whereas confidence supervision is used only at the final rollout layer. The reported model does not use the query-flow or dense dynamic-map loss.

A.2 Detailed Evaluation

Tables 6 and 7 report the class-wise IoU of InfraWorld at different forecasting horizons on InfraOcc and Occ3D–nus, respectively. Table 8 further groups the InfraOcc evaluation by the actual motion state of the occupied regions.

Table 8 shows that InfraWorld achieves a substantial improvement for moving participants at every forecasting horizon, increasing their average mIoU by 9.69 points over I^2-World. It also improves the prediction of stationary participants while retaining the high accuracy of the static background, indicating that the stable roadside context is well preserved during forecasting. On top of this stable

Table 8: Motion-state-aware forecasting result on InfraOcc.

Method	Moving participants				Stationary participants				Static background			
	mIoU (%) ↑				mIoU (%) ↑				mIoU (%) ↑			
	1 s	2 s	3 s	Avg.	1 s	2 s	3 s	Avg.	1 s	2 s	3 s	Avg.
I^2-World	25.24	12.13	6.59	14.65	60.55	53.42	50.61	54.86	97.30	97.27	97.07	97.21
InfraWorld	**38.36**	**22.12**	**12.54**	**24.34**	**64.61**	**61.39**	**58.58**	**61.53**	**97.79**	**97.72**	**97.68**	**97.73**

scene representation, persistent motion queries provide dedicated states for localized moving regions and recurrently refine their motion information throughout the rollout, enabling more accurate prediction of the regions that drive scene evolution.

The InfraOcc results support the selective state evolution underlying InfraWorld. Across the three-second rollout, all static categories remain above 90% IoU and most exceed 97%, showing that the dense BEV state preserves a complete and stable representation of the roadside scene. At the same time, the model maintains meaningful long-horizon predictions for diverse traffic participants; at 3 s, bicycle, bus, car, and truck still achieve IoUs above 40%. This class-level behavior complements the 9.02-point average dynamic mIoU improvement reported in the main paper. It demonstrates that persistent motion queries provide effective predictive capacity for localized traffic evolution, while confidence-weighted MAQI integrates these updates without compromising the scene context maintained by the dense pathway.

Figure 8 further visualizes the dynamic-class IoU gains over I^2-World on InfraOcc. The improvement is shared by all evaluated classes and remains evident at longer horizons, rather than being driven by a single category or short-term effect. This behavior is consistent with the complementary design of InfraWorld: the dense state retains the surrounding scene context, while persistent queries maintain dedicated motion information for localized traffic regions throughout rollout. Their selective integration therefore improves the prediction of diverse moving participants without requiring the dense pathway to model every local motion pattern on its own.

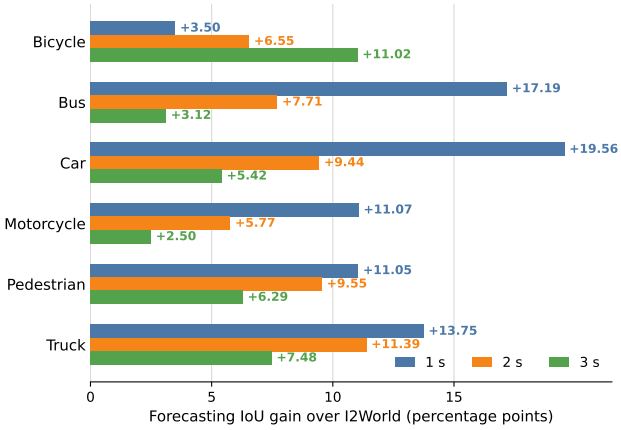


Figure 8: Per-class forecasting IoU gains of InfraWorld over I^2-World on InfraOcc. Gains are reported for dynamic classes at 1, 2, and 3 s.

The Occ3D–nus results provide a complementary evaluation under the ego-centric setting. InfraWorld maintains consistent performance across dynamic participants and surrounding scene categories. At the 3 s horizon, it obtains 24.67% dynamic mIoU and 34.20% semantic mIoU, while free-space IoU remains above 96% throughout the rollout. Together with the aggregate comparison in the main paper, these class-wise results suggest that the complementary dense and query representations can also provide useful predictive capability in ego-centric occupancy forecasting.

A.3 Implementation Details

Model configuration. On InfraOcc, the forecasting volume spans $[-64, 64]$ m along both horizontal axes and $[-4.8, 1.6]$ m vertically. On Occ3D–nus, it covers $[-40, 40]$ m along both horizontal axes and $[-1, 5.4]$ m vertically. The tokenizer uses a codebook of 512 entries with 128-dimensional embeddings. The world model takes the current latent state and four preceding states as input and autoregressively predicts the next six frames. The dense latent grids have spatial resolutions of 40×40 on InfraOcc and 50×50 on Occ3D–nus. The sparse branch maintains 20 persistent queries.

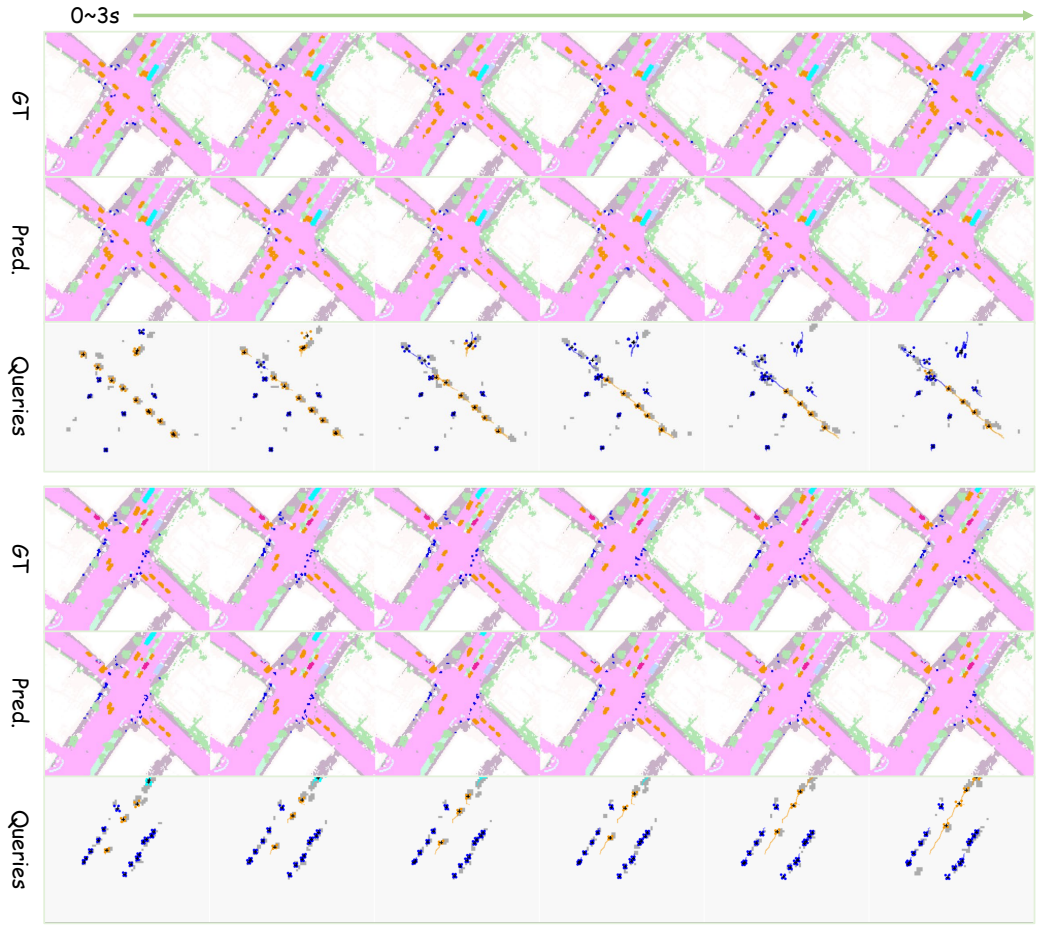


Figure 9: Additional qualitative results on InfraOcc. Each group shows a 0–3 s forecast sequence from left to right. The first two rows show ground-truth (GT) and predicted (Pred.) occupancy, respectively, while the third row visualizes the evolution of persistent motion queries during rollout.

The initial motion maps have resolutions of 80×80 and 100×100 on InfraOcc and Occ3D–nus, respectively, with a tile size of 4. Query refinement consists of three stages with $\{1, 2, 4\}$ sampled supports and eight attention heads. The validity threshold in DQI is set to 0.1.

Training details. We train the model in two stages. The tokenizer is first pretrained on four-frame sequences for 24 epochs and then kept frozen during world-model training. For the baselines in Table 1, we use their official implementations and train the occupancy tokenizer or world model on InfraOcc as required, under the same forecasting protocol. As InfraOcc uses a fixed roadside coordinate system, the ego-state input is replaced with an identity transformation matrix. On Occ3D–nus, ego-motion compensation is applied when aggregating multi-frame features in both the tokenizer and LST temporal fusion. The world model is optimized with AdamW using a learning rate of 10^{-3}, a weight decay of 10^{-2}, and gradient clipping with a maximum norm of 5. The learning rate is linearly warmed up from 10^{-3} times its base value over the first 200 iterations and then decayed according to the training schedule. We train the world model for 24 epochs on InfraOcc and 48 epochs on Occ3D–nus, using four NVIDIA GeForce RTX 4090 GPUs with a batch size of eight per GPU. Scheduled sampling is applied during training, with the teacher-forcing probability following $\exp(-2.5t/T)$, where t denotes the current iteration and T is the total number of training iterations.

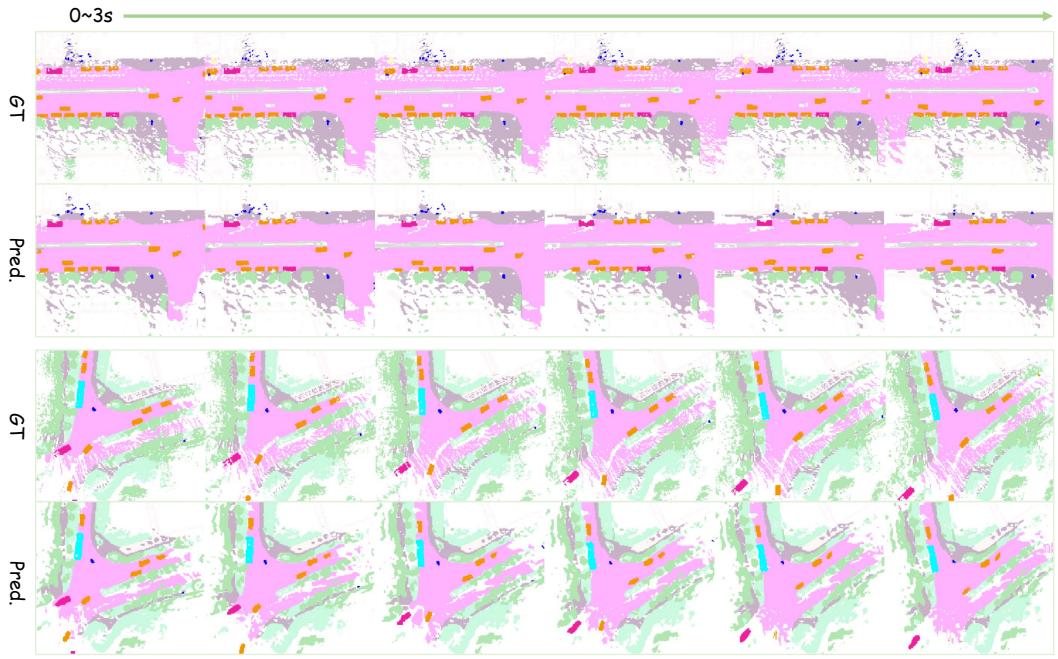


Figure 10: Additional qualitative results on Occ3D–nus. Each group shows a 0–3 s ego-centric forecast sequence from left to right, with ground-truth (GT) and predicted (Pred.) occupancy shown in paired rows.

A.4 Additional Visualization

Figures 9 and 10 provide additional qualitative results on InfraOcc and Occ3D–nus, respectively. Each sequence spans 0–3 s from left to right. Figure 9 additionally visualizes the evolution of persistent motion queries throughout the rollout.

As shown in Figure 9, InfraWorld preserves the dominant roadside scene layout while forecasting the evolution of localized traffic participants over the 3 s horizon. The query visualizations provide a closer view of the motion-focused reasoning during rollout. The persistent queries remain concentrated around a small number of dynamic regions and their sampling locations evolve with the predicted local motion patterns. This behavior is consistent with the intended role of the query pathway in maintaining dedicated states for the sparse regions that drive scene changes. DQI provides these queries with motion-supported initialization from the observed history, allowing the limited query budget to start from regions that are more likely to evolve. As forecasting proceeds, the queries are recurrently refined to preserve and update localized motion information over time. MAQI then selectively transfers reliable query information to the corresponding local BEV regions, allowing these motion-focused states to contribute to the dense scene prediction without broadly perturbing the surrounding context. The visualization therefore provides an intuitive view of how persistent queries remain associated with evolving traffic regions throughout autoregressive forecasting.

Figure 10 presents additional results on Occ3D–nus. Despite the shift from the fixed roadside setting to an ego-centric setting, InfraWorld produces predictions that remain broadly consistent with the ground truth over the forecast horizon and capture the major evolution of surrounding traffic.

A.5 Ablation on Hyperparameters

Tables 9 and 10 study two hyperparameters of the sparse query pathway on InfraOcc: the query budget N_q and the number of points N_p associated with each query. All other settings are kept unchanged.

Query budget. Increasing the query budget from 10 to 20 improves the average mIoU_{dyn} from 38.72% to 39.58%. With only 10 queries, the model may not provide sufficient spatial coverage of potential moving regions in traffic-dense scenes. A budget of 20 allows DQI to allocate dedicated query states to the major motion-supported regions in most cases. Further increasing N_q to 30 or 40 brings only limited improvements, as additional queries are more likely to cover weak or redundant motion candidates once the primary dynamic regions have been represented. Considering the marginal accuracy gains and the additional cost of query update and interaction, we set $N_q = 20$.

Points per query. Increasing N_p from 2 to 4 improves the average mIoU_{dyn} from 39.10% to 39.58%, indicating that four points provide sufficient local support for representing an evolving region. Using eight points achieves a slightly higher score of 39.66%, whereas further increasing N_p to 16 reduces the performance to 39.42%. Since the sampled point features are aggregated into a fixed-dimensional query state, densely sampling around the same query center introduces increasingly redundant features and may also incorporate less relevant context near semantic boundaries. This can dilute the localized motion information represented by the query while increasing the cost of point-wise localization and classification. We therefore use $N_p = 4$, which achieves nearly the best performance with lower sampling and prediction cost.

Table 9: Ablation of query budget N_q.

N_q	mIoU_{dyn} (%) ↑			
	1 s	2 s	3 s	Avg.
10	45.90	36.82	33.45	38.72
20	46.03	38.08	34.63	39.58
30	45.80	38.23	34.97	39.67
40	45.83	38.41	35.66	39.97

Table 10: Ablation of the number of points N_p per query.

N_p	mIoU_{dyn} (%) ↑			
	1 s	2 s	3 s	Avg.
2	44.74	37.58	34.99	39.10
4	46.03	38.08	34.63	39.58
8	45.99	38.24	34.74	39.66
16	45.59	37.79	34.88	39.42

A.6 Flow Ground-Truth Generation

We derive the InfraOcc flow labels by associating each occupancy grid with its annotated 3D boxes and velocities. The boxes and planar velocities are transformed from the LiDAR frame to the ego frame, and each oriented box of a dynamic category is rasterized onto the voxel grid. Voxels whose centers lie inside a box receive its class and planar velocity; overlapping boxes are resolved by the normalized distance to their centers. Thus, $\mathbf{f}_{uvz} = (v_x, v_y)$ for voxels assigned to moving dynamic boxes, and is zero for static, empty, or stationary regions. We enlarge the vertical box extent by a factor of 1.5 to better cover the occupied voxels.

To improve velocity reliability, we match dynamic boxes to temporal annotations by class and box geometry. Adjacent-frame annotations provide a temporal velocity estimate, and boxes with a temporal speed below 0.8 m/s are treated as stationary. For low-speed candidates, we further compare the future semantic consistency of the candidate velocity and zero flow at the next sample, mapping support voxels through the global coordinate frame; zero flow is selected only when it receives stronger future-occupancy support.

Box assignment may not cover every voxel of a dynamic object because of the difference between boxes and voxelized occupancy. We therefore propagate flow within face-connected components of each dynamic semantic class from assigned voxels to the remaining voxels, using the nearest assigned flow. The resulting two-dimensional field is stored with the semantic labels and generated at $2\times$, $4\times$, and $8\times$ downsampled resolutions to construct the moving-mask target in Eq. 18.

A.7 Limitations and Future Work

InfraWorld currently follows a two-stage training pipeline, where the occupancy tokenizer is first pretrained and subsequently frozen while the forecasting model is optimized separately. Although this design provides stable latent representations, it prevents the scene representation from being jointly adapted to the forecasting objective. An end-to-end framework that jointly optimizes occupancy representation and future prediction is therefore a natural direction for future work. In addition, while we evaluate InfraWorld in both roadside and ego-centric settings, the two perspectives are currently modeled independently. A further direction is to jointly exploit infrastructure-side and vehicle-side observations for cooperative occupancy forecasting, leveraging their complementary viewpoints and spatial coverage to build a more comprehensive predictive representation of the traffic scene.