SCOPE-OCC: SCENE-CONDITIONED PROBE LIFECYCLES FOR TRAJECTORY-FREE CAMERA-ONLY OCCUPANCY WORLD MODELING

Anonymous authors
Paper under double-blind review

ABSTRACT

Camera-only occupancy world models forecast future 3D semantic scenes from historical observations. Existing approaches often condition on future ego trajectories, which require an upstream planner or predictor and may introduce oracle dependence or compounding errors. Without such inputs, models must infer how scenes evolve and when and where representational capacity is needed. We present **SCOPE-Occ**, a trajectory-free camera-only occupancy world model that treats sparse queries as scene-conditioned sensing probes. Its Scene-Conditioned Probe Lifecycle Field (SPLF) predicts temporal support functions governing each probe's emergence, persistence, and termination, supervised by probe-level support matching and scene-level capacity calibration. These signals route spatial evolution in Lifecycle-Routed Probe Dynamics (LPD), where persistent and emerging probes follow distinct update paths integrating observed ego-motion context, local visual evidence, learned mobility, and lifecycle-gated relational interactions. Existence scores further gate probe contributions to future occupancy decoding. On Occ3D-nuScenes, SCOPE-Occ outperforms prior trajectory-free camera-only methods, achieving 21.68 mIoU and 45.26 IoU averaged over the 1–3 s forecasting horizons, without future ego trajectories at inference.

1 INTRODUCTION

Semantic occupancy provides a structured representation of 3D driving scenes by jointly describing surrounding geometry and semantics. Compared with object-centric perception, it offers a more complete basis for free-space reasoning, occlusion handling, long-tail scene understanding, and safety-critical planning. Recent benchmarks and methods have substantially advanced camera-based 3D occupancy perception (Tian et al., 2023; Wang et al., 2023; Huang et al., 2023; Li et al., 2023; Wei et al., 2023; Zhang et al., 2023). Beyond current-scene perception, occupancy world models further forecast how semantic scenes evolve over the next few seconds, providing an important intermediate representation for autonomous systems that must reason before acting (Ma et al., 2024; Zheng et al., 2024; Gu et al., 2024; Shi et al., 2026; Chen et al., 2025).

Future ego trajectory is an effective condition for reducing the uncertainty of occupancy forecasting. It provides an explicit temporal reference for future horizons and a spatial reference for the evolution of the ego-centered scene. In sparse query-based models, future trajectory can condition horizon-specific queries and align their spatial interaction with historical visual observations (Du et al., 2026). In controllable or generative occupancy world models, trajectory is similarly encoded as an external condition that guides future scene generation or separates ego-induced changes from scene evolution (Gu et al., 2024; Shi et al., 2026). Therefore, as illustrated in Fig. 1, trajectory conditioning effectively provides *temporal anchoring* and *spatial alignment* for otherwise scene-independent future representations.

However, future ego trajectory is not an observed input under trajectory-free deployment. Once this external guidance is removed, a model must internally recover two representation-organizing capabilities. First, it must determine *when* each sparse representation should participate in future prediction, including whether it should persist, emerge only at a later time, or gradually decay. Second, it must determine *where* that representation should evolve while accounting for ego-conditioned

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.

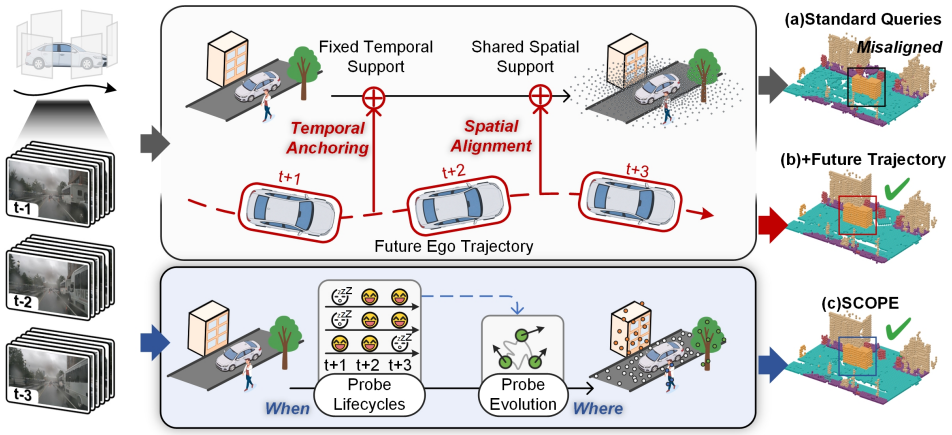

Figure 1: Future trajectories anchor conventional queries in time and space, while SCOPE learns probe lifecycles (*when*) and evolution (*where*) from historical observations.

scene changes, independent local motion, and newly supported regions. Existing query-based occupancy forecasters either reuse all queries across horizons or employ predefined temporal schedules, while their spatial support is commonly initialized from shared priors (Wang et al., 2024; Dang et al., 2026; Du et al., 2026). Although these queries can be refined during forecasting, their temporal and spatial organization remains insufficiently adaptive to the current scene.

We address this limitation with **SCOPE-Occ**, referred to as **SCOPE** hereafter, a trajectory-free camera-only occupancy world model that recasts sparse queries as scene-conditioned sensing probes and learns their temporal and spatial organization from historical images and ego states. Its **Scene-Conditioned Probe Lifecycle Field** (SPLF) predicts scene-conditioned temporal support functions, determining when probes emerge, persist, or decay and producing adaptive existence signals. **Lifecycle-Routed Probe Dynamics** (LPD) then uses these signals, together with ego-conditioned context, local visual evidence, mobility, and lifecycle-gated relations, to route persistent evolution, mobility-aware correction, emergence relocation, and occupancy materialization. SCOPE thereby internalizes the temporal support and spatial guidance otherwise supplied by future trajectories.

Experiments on Occ3D-nuScenes demonstrate that SCOPE achieves strong camera-only trajectory-free forecasting performance, reaching 21.68 mIoU and 45.26 IoU. The results show that scene-conditioned probe organization can effectively improve long-horizon occupancy prediction without future ego trajectories or poses.

Our contributions are summarized as follows:

- We formulate trajectory-free sparse occupancy forecasting as a problem of scene-adaptive temporal and spatial representation organization, where the model must infer *when* probes should participate and *where* they should evolve without future trajectory guidance.
- We propose SPLF, which replaces scene-independent query participation with input-conditioned temporal support functions and produces existence, persistence, and emergence signals.
- We introduce LPD, which reuses lifecycle signals to guide mobility-aware interactions, spatial updates, emergence relocation, and occupancy materialization.
- SCOPE achieves strong trajectory-free camera-only occupancy forecasting results on Occ3D-nuScenes, outperforming prior camera-only methods and narrowing the gap to trajectory-conditioned references.

2 Related Work

2.1 3D Occupancy Perception

3D semantic occupancy jointly represents surrounding geometry and semantics beyond predefined object categories. Occ3D (Tian et al., 2023) and OpenOccupancy (Wang et al., 2023) provide large-scale annotations and evaluation protocols for camera-based occupancy perception. Existing methods employ structured representations such as tri-perspective planes (Huang et al., 2023), sparse voxel proposals (Li et al., 2023), and dense multi-scale voxel features (Wei et al., 2023; Zhang et al., 2023). More recent sparse or primitive-based approaches represent occupied scenes using semantic Gaussians (Huang et al., 2024; 2025) or sparse point sets (Wang et al., 2024), reducing the redundancy of dense voxel processing. While these methods mainly estimate the observed scene, our work focuses on forecasting its future evolution from historical images.

2.2 Occupancy Forecasting and World Models

Camera-only forecasting methods predict future occupancy directly from historical visual observations. Cam4DOcc (Ma et al., 2024) establishes a camera-only 4D forecasting benchmark, while OccProphet (Chen et al., 2025) improves efficiency through an observer–forecaster–refiner pipeline. Other world models operate primarily in occupancy or latent space: OccWorld (Zheng et al., 2024) autoregressively predicts future scene and ego tokens, DOME (Gu et al., 2024) introduces trajectory-controllable diffusion generation, and COME (Shi et al., 2026) separates scene evolution from ego-induced perspective changes through scene-centric control. GaussianWorld (Zuo et al., 2025) exploits explicit scene-evolution priors for streaming occupancy prediction, whereas GEM (Chen et al., 2026) models future scenes with continuous 4D Gaussian primitives that can be queried at arbitrary timestamps. SparseWorld-TC (Du et al., 2026) directly conditions sparse camera-based forecasting on future ego trajectories. In contrast, SCOPE follows a strict trajectory-free protocol, using neither future ego trajectories nor future poses as model inputs at inference.

2.3 Sparse and Continuous Occupancy Representations

Learnable queries have enabled efficient temporal perception in BEVFormer (Li et al., 2025b) and Sparse4D (Lin et al., 2023). For occupancy modeling, OPUS (Wang et al., 2024) formulates current-scene occupancy as sparse set prediction, while SparseWorld (Dang et al., 2026) extends dynamic queries to perception, forecasting, and planning. S2GO (Park et al., 2026) propagates compact 3D queries online and decodes them into semantic Gaussians, whereas GEM (Chen et al., 2026) and DIO (Diehl et al., 2025) model scene evolution through continuous space-time primitives or implicit occupancy-flow fields. These methods demonstrate the effectiveness of compact scene representations, but they either propagate a shared representation set, model temporal support separately from downstream routing, or organize future prediction using external conditions. SCOPE instead learns an input-conditioned lifecycle for each probe and reuses its existence, persistence, and emergence signals to jointly control relational interaction, spatial evolution, and occupancy materialization.

3 SCOPE

3.1 Overview and Problem Formulation

Given historical multi-view images, camera calibration, and historical ego states, SCOPE predicts semantic occupancy $\{\hat{\mathcal{O}}^h\}_{h=1}^H$ at future timestamps $\{\tau_h\}_{h=1}^H$, without using future ego trajectories or poses as model inputs. Let X denote the historical observations and $\mathcal{F}_X$ the corresponding multi-scale image features. All predicted probe coordinates and future occupancy targets are expressed in the current reference frame. Future ego poses, when required for offline annotation transformation, are never provided to the network.

SCOPE is built upon a sparse query-based occupancy forecaster with N shared probe slots $q_i = (\mathbf{f}_i, \mathbf{c}_i)$, $i = 1, \ldots, N$, where $\mathbf{f}_i$ and $\mathbf{c}_i$ denote the feature and spatial reference of probe slot i, respectively. The same probe slots are evaluated across future horizons, providing index-level correspondence for lifecycle modeling without assuming instance identity. Given $\mathcal{F}_X$ and a target

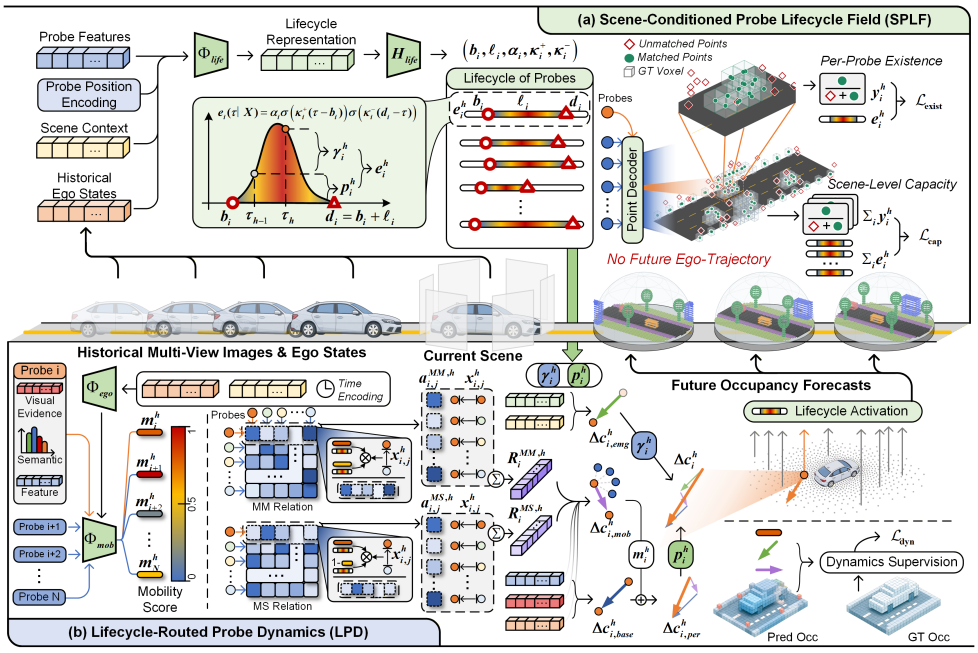

Figure 2: Overview of SCOPE. From historical multi-view images and ego states, (a) SPLF predicts scene-conditioned temporal support with per-probe existence and scene-level capacity supervision; (b) LPD uses the resulting lifecycle signals to model moving-to-moving and moving-to-stable relations and route persistent and emerging updates. Ground-truth assignments and occupancy are training-only; inference requires no future ego trajectory.

timestamp τ_h, the base sparse decoder produces

$$\bar{Z}_i^h = \left(\bar{\mathbf{f}}_i^h,\, \bar{P}_i^h,\, \bar{S}_i^h\right). \tag{1}$$

Here, $\bar{\mathbf{f}}_i^h$ is the candidate probe feature, $\bar{P}_i^h \in \mathbb{R}^{R\times 3}$ is its point bundle, and $\bar{S}_i^h \in \mathbb{R}^{R\times C}$ contains the point-level semantic logits. Its candidate center is $\bar{\mathbf{c}}_i^h = R^{-1}\sum_{n=1}^{R} \bar{\mathbf{p}}_{i,n}^h$.

As illustrated in Fig. 2, SPLF augments candidate probes with scene-conditioned temporal support and derives existence, persistence, and emergence signals. LPD combines these signals with ego-conditioned context, local visual evidence, mobility, and relational interactions to refine probe geometry and semantics before materializing future occupancy. All future horizons are processed in parallel without autoregressive rollout.

3.2 Scene-Conditioned Probe Lifecycle Field

SPLF assigns each probe slot a scene-conditioned temporal support function evaluated at the target forecast timestamps. The resulting lifecycle controls temporal participation and provides routing signals to LPD.

Lifecycle Construction. For each probe slot i, we construct

$$\mathbf{u}_i = \Phi_{\text{life}}\left[\mathbf{f}_i;\, \mathbf{c}_X;\, \psi_p(\mathbf{c}_i);\, \mathbf{h}_{\text{ego}}\right], \tag{2}$$

where $\mathbf{c}_X$ is the pooled scene context, while ψ_p and $\mathbf{h}_{\text{ego}}$ encode probe position and historical ego states, respectively. Instead of independent horizon-wise gates, the lifecycle head predicts one temporal support function per probe slot:

$$(b_i, \ell_i, \alpha_i, \kappa_i^+, \kappa_i^-) = H_{\text{life}}(\mathbf{u}_i), \qquad d_i = b_i + \ell_i. \tag{3}$$

Here, b_i and d_i are the emergence and termination centers, α_i is the scene-conditioned activation amplitude, and κ_i^+, κ_i^- control transition sharpness. Sigmoid or softplus transformations constrain these parameters to valid ranges.

The temporal support of probe i is

$$e_i(\tau \mid X) = \alpha_i \sigma\big(\kappa_i^+(\tau - b_i)\big)\, \sigma\big(\kappa_i^-(d_i - \tau)\big)\,. \tag{4}$$

We evaluate the continuous support only at the benchmark timestamps, $e_i^h = e_i(\tau_h \mid X)$, with $h = 0$ denoting the current timestamp; $\sum_i e_i^h$ gives the probe capacity at horizon h.

Lifecycle Signals and Supervision. We split support at horizon h into inherited and newly activated parts:

$$p_i^h = \min(e_i^{h-1}, e_i^h), \qquad \gamma_i^h = [e_i^h - e_i^{h-1}]_+. \tag{5}$$

Here, $[x]_+ = \max(x, 0)$. Since $e_i^h = p_i^h + \gamma_i^h$, p_i^h represents inherited support and γ_i^h newly activated support. Decreasing e_i^h progressively suppresses probes approaching termination.

Since probe existence has no direct annotation, we reuse the point-to-occupancy assignment from the base occupancy supervision. Let $\mathcal{M}_h$ contain candidate points assigned to occupied ground-truth locations at horizon h. The soft matched-support target is

$$y_i^h = \frac{1}{R} \sum_{n=1}^{R} \mathbf{1}[(i, n) \in \mathcal{M}_h]. \tag{6}$$

We supervise individual probes using

$$\mathcal{L}_{\text{exist}} = \frac{1}{NH} \sum_{h=1}^{H} \sum_{i=1}^{N} \text{FocalBCE}(e_i^h, y_i^h). \tag{7}$$

Let $\bar{e}^h = N^{-1} \sum_i e_i^h$ and $\bar{y}^h = N^{-1} \sum_i y_i^h$. The scene-level capacity loss is

$$\mathcal{L}_{\text{cap}} = \frac{1}{H} \sum_{h=1}^{H} \text{SmoothL1}\left(\bar{e}^h, \bar{y}^h\right). \tag{8}$$

The lifecycle objective is $\mathcal{L}_{\text{life}} = \mathcal{L}_{\text{exist}} + \mathcal{L}_{\text{cap}}$. SPLF provides e_i^h for probe materialization, p_i^h for persistent evolution, and γ_i^h for emergence relocation in LPD.

3.3 Lifecycle-Routed Probe Dynamics

Given e_i^h, p_i^h, and γ_i^h, LPD determines where temporally supported probes should evolve. It estimates probe mobility, aggregates context from mobile and stable-support neighbors, and routes persistent and emerging spatial updates.

Ego-Conditioned Mobility. Without future trajectories, we infer $\mathbf{z}_{\text{ego}}^h = \Phi_{\text{ego}}[\mathbf{h}_{\text{ego}}; \mathbf{c}_X; \psi_t(\tau_h)]$ from historical ego motion, scene context, and time. For each candidate probe, we sample local evidence $\mathbf{v}_i^h = \text{MVSample}(\mathcal{F}_X, \bar{\mathbf{c}}_i^h)$ by projecting its center into historical views with known camera geometry, and pool its semantics as $\bar{\boldsymbol{\pi}}_i^h = R^{-1} \sum_{n=1}^{R} \sigma(\bar{\mathbf{S}}_{i,n}^h)$. The mobility score is

$$m_i^h = \sigma\big(\Phi_{\text{mob}}\left[\bar{\mathbf{f}}_i^h; \bar{\boldsymbol{\pi}}_i^h; \mathbf{v}_i^h; \mathbf{z}_{\text{ego}}^h\right]\big)\,, \tag{9}$$

where $m_i^h \in [0, 1]$ estimates the motion tendency of the probe-supported region.

Lifecycle-Gated Relational Context. For each probe i, we construct a neighborhood $\mathcal{N}_i^h$ from candidate centers. A pair descriptor $\mathbf{x}_{ij}^h$ combines feature compatibility, relative geometry, semantic similarity, existence, and mobility. We consider moving-to-moving (MM) relations, which aggregate mobile neighbors, and moving-to-stable-support (MS) relations, which aggregate low-mobility neighbors. Their neighbor gates are $g_j^{\text{MM},h} = e_j^h m_j^h$ and $g_j^{\text{MS},h} = e_j^h (1 - m_j^h)$.

For $q \in \{\text{MM}, \text{MS}\}$, the normalized relation coefficient is

$$a_{ij}^{q,h} = \frac{g_j^{q,h} \exp\big(\chi_q(\mathbf{x}_{ij}^h)\big)}{\sum_{k \in \mathcal{N}_i^h} g_k^{q,h} \exp\big(\chi_q(\mathbf{x}_{ik}^h)\big) + \epsilon}, \tag{10}$$

and the corresponding context is

$$\mathbf{R}_i^{q,h} = \sum_{j \in \mathcal{N}_i^h} a_{ij}^{q,h} \Psi_q(\mathbf{x}_{ij}^h), \tag{11}$$

where χ_q predicts pairwise compatibility and Ψ_q produces relation messages. Neighbor gates determine which probes contribute, while source-probe persistence and mobility are applied during displacement routing.

Routed Evolution and Materialization. LPD predicts a persistent base update $\Delta\mathbf{c}_{i,\text{base}}^h$, a relation-aware mobility residual $\Delta\mathbf{c}_{i,\text{mob}}^h$, and an emergence relocation $\Delta\mathbf{c}_{i,\text{emg}}^h$. They are routed as

$$\Delta\mathbf{c}_{i,\text{per}}^h = \Delta\mathbf{c}_{i,\text{base}}^h + m_i^h \Delta\mathbf{c}_{i,\text{mob}}^h, \qquad \Delta\mathbf{c}_i^h = p_i^h \Delta\mathbf{c}_{i,\text{per}}^h + \gamma_i^h \Delta\mathbf{c}_{i,\text{emg}}^h. \tag{12}$$

Thus, inherited support follows base evolution augmented by a mobility-scaled residual, whereas newly activated support follows the emergence relocation path.

Fusing candidate features with visual, ego, and relational context into $\mathbf{h}_i^h$, we refine and materialize each probe as

$$\hat{P}_i^h = \bar{P}_i^h + \mathbf{1}_R \otimes \Delta\mathbf{c}_i^h, \qquad \hat{S}_i^h = \bar{S}_i^h + \Phi_{\text{sem}}(\mathbf{h}_i^h), \qquad \hat{\mathcal{O}}^h = \text{Vox}\Big(\{(\hat{P}_i^h, e_i^h \sigma(\hat{S}_i^h))\}_{i=1}^N\Big). \tag{13}$$

Here, p_i^h, γ_i^h, and m_i^h govern spatial refinement, while e_i^h gates occupancy contribution.

Dynamics Supervision. The refined points and semantic logits are supervised by the standard occupancy loss $\mathcal{L}_{\text{occ}}$. For slots with reliable geometric support at both adjacent horizons, the displacement between their matched-support centroids defines the soft mobility target

$$\tilde{m}_i^h = \text{clip}\left(\frac{\|\mathbf{g}_i^h - \mathbf{g}_i^{h-1}\|_2}{s_m}, 0, 1\right), \tag{14}$$

where s_m is the motion normalization scale and displacement is measured in the current reference frame. Let Ω_m denote these support-reliable slot–horizon pairs. For emerging probes, we define $\tilde{\gamma}_i^h = [y_i^h - y_i^{h-1}]_+$, and let Ω_e denote the reliable slot–horizon pairs with positive $\tilde{\gamma}_i^h$. The dynamics objective is

$$\begin{aligned}\mathcal{L}_{\text{dyn}} = \mathcal{L}_{\text{occ}} &+ \sum_{(i,h)\in\Omega_m} \text{SmoothL1}(m_i^h, \tilde{m}_i^h) \\ &+ \sum_{(i,h)\in\Omega_m} p_i^h (1 - \tilde{m}_i^h) \|\Delta\mathbf{c}_{i,\text{mob}}^h\|_1 + \sum_{(i,h)\in\Omega_e} \tilde{\gamma}_i^h \|\hat{\mathbf{c}}_i^h - \mathbf{g}_i^h\|_1,\end{aligned} \tag{15}$$

where $\hat{\mathbf{c}}_i^h$ is the centroid of $\hat{P}_i^h$. The auxiliary terms supervise mobility, suppress residual motion for persistent low-mobility probes, and relocate emerging probes toward new support. The total loss is $\mathcal{L} = \mathcal{L}_{\text{dyn}} + \lambda_{\text{life}} \mathcal{L}_{\text{life}}$.

At inference, the base decoder produces candidates for all horizons, which SPLF and LPD process in parallel to infer lifecycle support and mobility-aware relations, route spatial updates, and materialize semantic occupancy without autoregressive rollout.

4 Experiments

4.1 Experimental Settings

Dataset and Evaluation Protocol. We conduct experiments on the official nuScenes v1.0-trainval split (Caesar et al., 2020) using semantic occupancy annotations from Occ3D-nuScenes (Tian et al., 2023). Each input sequence contains five synchronized multi-view image sets, comprising four historical timestamps and the current timestamp. The model forecasts future occupancy at 1, 2, and 3 seconds. For semantic occupancy forecasting, we report occupancy IoU and semantic mIoU. For Table 2 only, we train a lightweight trajectory head with a multi-horizon L_2 loss. This head and loss are disabled for all occupancy results in Table 1.

Table 1: Main results on future semantic occupancy forecasting. Methods are grouped by input protocol. * denotes DINOv3 and † marks our reproductions; for SparseWorld-TC, only the Small configuration is publicly available.

Methods	mIoU				IoU				FPS
	1s	2s	3s	Avg	1s	2s	3s	Avg	
Multi-camera images + ground-truth future trajectories									
I^2-World-STC (Liao et al., 2025)	21.67	18.78	16.47	18.97	30.55	28.76	26.99	28.77	4.21
DOME-F (Gu et al., 2024)	24.12	17.41	13.24	18.25	35.18	27.90	23.44	28.84	–
COME (Shi et al., 2026)	26.56	21.73	18.49	22.26	48.08	43.84	40.28	44.07	–
SparseWorld-TC-Small (Du et al., 2026)	27.95	25.51	23.35	25.60	50.69	49.15	47.23	49.02	9.35
SparseWorld-TC-Large (Du et al., 2026)	28.64	26.28	24.36	26.42	50.57	49.26	47.80	49.21	3.58
SparseWorld-TC-Small* (Du et al., 2026)	30.09	27.61	25.43	27.71	52.66	51.20	49.31	51.05	5.63
SparseWorld-TC-Large* (Du et al., 2026)	32.76	29.62	27.28	29.89	55.28	53.56	51.71	53.52	2.08
Multi-camera images + predicted future trajectories									
COME (Shi et al., 2026)	25.57	18.35	13.41	19.11	45.36	37.06	30.46	37.63	–
SparseWorld-TC-Small† (Du et al., 2026)	25.28	21.46	18.06	21.60	47.89	45.01	41.83	44.91	3.92
Multi-camera images only (trajectory-free)									
RenderWorld (Yan et al., 2025)	2.83	2.55	2.37	2.58	14.61	13.61	12.98	13.73	–
PreWorld (Li et al., 2025a)	12.27	9.24	7.15	9.55	23.62	21.62	19.63	21.62	–
Occ-LLM (Xu et al., 2025)	11.28	10.21	9.13	10.21	27.11	24.07	20.19	23.79	–
I^2-World-STC† (Liao et al., 2025)	15.52	11.60	8.66	11.93	22.48	18.07	15.27	18.61	9.40
COME† (Shi et al., 2026)	14.53	10.39	8.39	11.10	22.34	18.09	15.48	18.64	2.37
OccLLaMA (Wei et al., 2024)	10.34	8.66	6.98	8.66	25.81	23.19	19.97	22.99	–
SparseWorld (Dang et al., 2026)	14.93	13.15	11.51	13.20	22.96	22.10	21.05	22.03	8.00
OccTENS-F (Jin et al., 2026)	17.17	10.38	7.82	11.79	27.60	25.14	20.33	24.35	–
SparseWorld-TC-Small† (Du et al., 2026)	20.61	15.22	12.16	15.99	41.19	36.59	32.97	36.91	<u>9.44</u>
SCOPE (This Paper)	<u>24.24</u>	<u>20.95</u>	<u>17.87</u>	<u>21.02</u>	<u>45.58</u>	<u>43.41</u>	<u>39.98</u>	<u>42.99</u>	**10.79**
SCOPE* (This Paper)	**25.31**	**21.56**	**18.17**	**21.68**	**48.18**	**45.30**	**42.29**	**45.26**	7.45

Implementation Details. We train SCOPE for 40 epochs on four NVIDIA A100 GPUs with a batch size of four per GPU. We instantiate two variants using ResNet-50 (He et al., 2016) and DINOv3-pretrained ConvNeXt-Base (Siméoni et al., 2025; Liu et al., 2022) as image backbones, each equipped with an FPN neck (Lin et al., 2017), and report both variants in the main comparison. We use AdamW (Loshchilov & Hutter, 2019) with an initial learning rate of 2×10^{-4}, a weight decay of 0.01, linear warmup, and cosine learning-rate decay. The model maintains $1{,}340$ sparse probes, each decoded into 64 points. During voxelization, class scores of points falling into the same voxel are aggregated by max pooling before applying the standard semantic occupancy evaluation. For trajectory prediction, we attach a lightweight trajectory head trained with a multi-horizon L_2 loss; its predictions are not used to condition occupancy forecasting.

Trajectory-Free Protocol. During inference, SCOPE receives only historical and current multi-view images, camera calibration, and historical/current ego states. Future ego trajectories, future ego poses, future occupancy observations, LiDAR, and radar are never used as model inputs. Future ego poses, when required for constructing supervision or evaluation labels, are used only by the offline data-processing pipeline. The trajectory results reported below evaluate predicted trajectories as model outputs rather than conditioning signals for occupancy forecasting.

4.2 Main Results

Semantic Occupancy Forecasting. Table 1 compares SCOPE with existing occupancy forecasting methods under different input protocols. Under the strict camera-only setting, SCOPE with ResNet-50 achieves 21.02 average mIoU and 42.99 average IoU, outperforming the strongest camera-only approach, SparseWorld-TC-Small†, by 5.03 and 6.08 points, respectively. The gains become more pronounced at the 3-second horizon, reaching 5.71 mIoU and 7.01 IoU. This trend indicates that scene-conditioned probe lifecycles and lifecycle-routed evolution are particularly effective for preserving and reallocating representation support during long-horizon forecasting. With

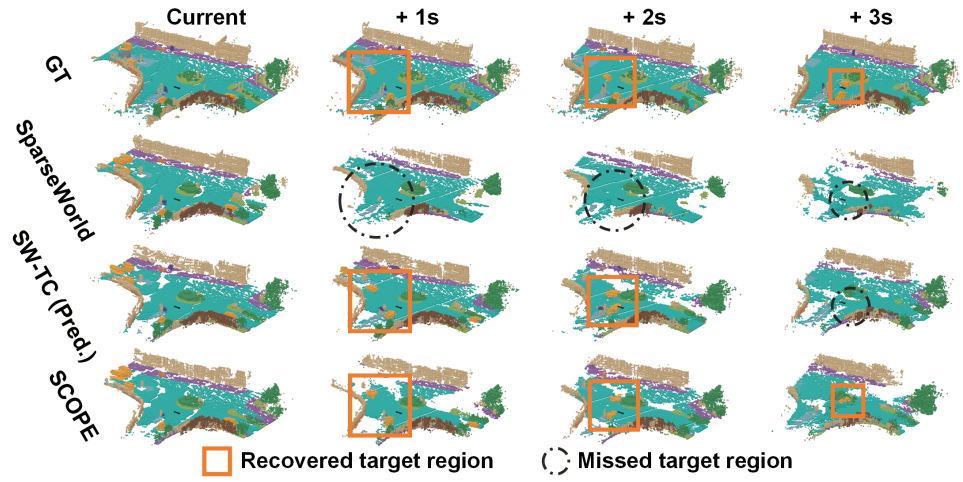

Figure 3: Qualitative comparison of current reconstruction and future occupancy forecasting. SW-TC (Pred.) denotes SparseWorld-TC-Small with predicted trajectories; boxes and circles indicate recovered and missed regions, respectively.

the DINOv3 ConvNeXt-Base backbone, SCOPE further improves the average results to 21.68 mIoU and 45.26 IoU, establishing new state-of-the-art performance among camera-only methods.

Without future trajectories, the ResNet-50 SCOPE narrows the gap to predicted-trajectory SparseWorld-TC-Small† to 0.58 average mIoU and 1.92 average IoU. With DINOv3, SCOPE* further slightly exceeds this reference by 0.08 mIoU and 0.35 IoU, although this comparison also reflects backbone differences. These results show that SCOPE substantially narrows the trajectory-conditioning gap. The ResNet-50 variant also achieves 10.79 FPS, offering a favorable accuracy–efficiency trade-off.

Table 2: Trajectory prediction results on nuScenes.

Method	L_2 (m) ↓				Collision Rate (%) ↓			
	1s	2s	3s	avg	1s	2s	3s	avg
UniAD	0.48	0.96	1.65	1.03	**0.05**	**0.17**	0.71	0.31
OccNet	1.29	2.13	2.99	2.14	0.21	0.59	1.37	0.72
OccWorld	0.52	1.27	2.41	1.40	0.12	0.40	2.08	0.87
OccLLaMA	0.38	1.07	2.15	1.20	0.06	0.39	1.65	0.70
PreWorld	0.22	0.30	0.40	0.31	0.21	0.66	0.71	0.53
SparseWorld	0.19	0.25	0.36	0.27	0.11	0.29	0.46	0.29
SCOPE	**0.16**	**0.20**	**0.24**	**0.20**	0.24	0.22	0.22	0.23
SCOPE*	**0.16**	**0.20**	**0.24**	**0.20**	0.17	**0.17**	**0.19**	**0.18**

Qualitative Results. In the example shown in Fig. 3, the methods produce visually similar current-scene reconstructions, while their future forecasts differ more substantially. SparseWorld loses the highlighted vehicle region at all future horizons. With predicted trajectory conditioning, SW-TC (Pred.) recovers the target at $+1$ and $+2$ seconds but misses it at $+3$ seconds. In contrast, SCOPE consistently preserves the target region through all forecasting horizons without using any future trajectory input. This result illustrates that scene-conditioned probe lifecycles and lifecycle-routed evolution can compensate for part of the temporal anchoring and spatial guidance otherwise supplied by predicted trajectories.

Trajectory Prediction. Table 2 compares SCOPE with the planning-oriented UniAD (Hu et al., 2023), the occupancy-based OccNet (Tong et al., 2023), and the occupancy world model OccWorld (Zheng et al., 2024), together with recent forecasting methods. Both SCOPE variants achieve the lowest L_2 errors at all horizons, reducing the average error of SparseWorld from 0.27 m to 0.20 m. The advantage increases at longer horizons, with the 3-second error decreasing from 0.36 m to 0.24 m. SCOPE* further obtains the lowest average collision rate of 0.18%, compared with 0.29% for SparseWorld, demonstrating accurate and safe long-horizon trajectory prediction.

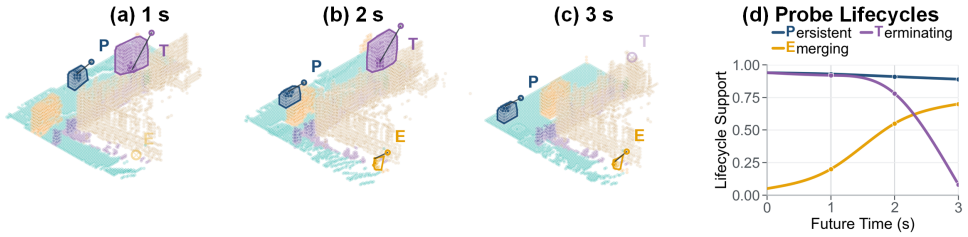

Figure 4: Probe lifecycle analysis. Persistent (P), emerging (E), and terminating (T) probes are shown over 1–3 s with their lifecycle-support curves.

4.3 Ablation and Analysis

Component Ablation. Table 3 evaluates the main components of SCOPE. SPLF improves the base model by 3.05 average mIoU and 4.79 average IoU, while mobility-aware evolution provides further gains. MM and MS are complementary: their combination clearly outperforms either branch alone. Emergence relocation further improves the 3-second mIoU and IoU by 0.68 and 1.10 points. Overall, full SCOPE surpasses the base model by 7.15 average mIoU and 10.67 average IoU.

Table 3: Component ablation on Occ3D-nuScenes. *Mob.* denotes mobility-aware persistent evolution, and *Emg.* denotes emergence relocation.

Components					mIoU		IoU	
SPLF	Mob.	MM	MS	Emg.	3s	Avg	3s	Avg
					11.66	13.87	27.40	32.32
✓					14.21	16.92	34.96	37.11
✓	✓				15.23	18.36	36.03	39.34
✓	✓	✓			16.01	19.43	37.72	41.20
✓	✓		✓		15.44	18.79	37.17	40.91
✓	✓	✓	✓		17.19	20.54	38.88	42.01
✓	✓	✓	✓	✓	**17.87**	**21.02**	**39.98**	**42.99**

Table 4 examines the temporal-support design. Learned support already outperforms fixed all-on participation, while scene-conditioned continuous lifecycles further improve over static lifecycles and horizon-wise gates. Capacity calibration raises average mIoU and IoU by 1.13 and 1.38 points and reduces Cap. Err. from 0.19 to 0.13. Overall, full SPLF improves the fixed design by 4.96 mIoU and 5.53 IoU, while reducing Cap. Err. from 0.42 to 0.13.

Table 4: Ablation of SPLF temporal-support designs. mIoU and IoU are averaged over 1–3 s; *Cap. Err.* measures the active-capacity calibration error.

Support	Scene	Cont.	$\mathcal{L}_{\text{cap}}$	mIoU	IoU	Cap. Err.
Fixed all-on				16.06	37.46	0.42
Static lifecycle		✓		17.92	39.23	0.31
Horizon gate	✓			18.22	40.22	0.28
Continuous life.	✓	✓		19.89	41.61	0.19
Full SPLF	✓	✓	✓	**21.02**	**42.99**	**0.13**

Probe Lifecycle Analysis. Fig. 4 tracks three probe slots: persistent support remains high, emerging support grows, and terminating support decays toward the final horizon. These lifecycles reflect scene-adaptive participation rather than a shared horizon-wise schedule.

5 Conclusion

We presented **SCOPE-Occ**, a trajectory-free camera-only occupancy world model combining SPLF's scene-conditioned probe lifecycles with LPD's mobility-aware interactions, spatial updates, and occupancy materialization. Without future trajectories or poses, SCOPE achieves 21.68 average mIoU and 45.26 average IoU on Occ3D-nuScenes, establishing state-of-the-art trajectory-free camera-only performance and remaining competitive with predicted-trajectory methods. Ablations validate adaptive temporal support, relational dynamics, and emergence relocation. Weak support matching and a fixed probe budget may limit crowded or ambiguous scenes; future work will explore uncertainty-aware adaptive allocation and cross-dataset validation.

AI USE STATEMENT

Generative AI tools were used for language editing, including checking grammar, improving readability, and polishing the manuscript. OpenAI Codex was additionally used to assist with LaTeX format checking and ensure compliance with submission requirements.

ETHICS STATEMENT

This work studies camera-only occupancy forecasting for autonomous systems. Incorrect predictions under occlusion, rare classes, adverse weather, or domain shift may affect downstream planning. The model is a predictive perception component, not a safety guarantee; uncertainty estimation, redundant sensing, runtime monitoring, and conservative planning remain necessary. Dataset bias may lead to uneven performance across regions and traffic conditions. Further discussion of limitations and broader impacts appears in Appendices F.3 and F.4.

REPRODUCIBILITY STATEMENT

The model formulation and training objectives are described in Section 3, and the dataset, evaluation protocol, and principal training settings in Section 4.1. Appendix A details probe correspondence, coordinate conventions, weak supervision, and the algorithmic training and inference flow. Appendix B provides network configuration, optimization, and data processing details. Additional ablations and efficiency measurements are reported in Appendices C and D.

REFERENCES

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*, 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 efficiency frontier of camera-only 4d occupancy forecasting with 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. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 3497–3505, 2026. doi: 10.1609/aaai.v40i5.37347. Issue 5.

Christopher Diehl, Quinlan Sykora, Ben Agro, Thomas Gilles, Sergio Casas, and Raquel Urtasun. Dio: Decomposable implicit 4d occupancy-flow world model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

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*, 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 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 Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, Lewei Lu, Xiaosong Jia, Qiang Liu, Jifeng Dai, Yu Qiao, and Hongyang Li. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17853–17862, 2023.

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*, 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*, 2024.

Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, 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*, 2025.

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*, 2026. doi: 10.1109/LRA.2026.3655202.

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*, 2025a.

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*, 2023.

Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Qiao Yu, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from lidar-camera via spatiotemporal transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(3):2020–2036, 2025b. doi: 10.1109/TPAMI.2024.3515454.

Zhimin Liao, Ping Wei, Ruijie Zhang, Shuaijia Chen, Haoxuan Wang, and Ziyang Ren. I2-world: Intra-inter tokenization for efficient dynamic 4d scene forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 25810–25819, 2025.

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 Conference on Computer Vision and Pattern Recognition*, pp. 936–944, 2017.

Xuewu Lin, Zixiang Pei, Tianwei Lin, Lichao Huang, and Zhizhong Su. Sparse4d v3: Advancing end-to-end 3d detection and tracking. arXiv preprint arXiv:2311.11722, 2023.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11976–11986, 2022.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

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*, 2024.

Jinhyung Park, Yihan Hu, Chensheng Peng, Wenzhao Zheng, Kris Kitani, and Wei Zhan. S2go: Streaming sparse gaussian occupancy. In *International Conference on Learning Representations*, 2026.

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*, 2026.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, 2025.

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*, 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.

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*, 2024.

Xiaofeng Wang, Zheng Zhu, Wenbo Xu, Yunpeng Zhang, Yi Wei, Xu Chi, Yun Ye, Dalong Du, Jiwen Lu, and Xingang Wang. Openoccupancy: A large scale benchmark for surrounding semantic occupancy perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

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*, 2023.

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 *2025 IEEE International Conference on Robotics and Automation*, pp. 8434–8441, 2025. doi: 10.1109/ICRA55743.2025.11127665.

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 *2025 IEEE International Conference on Robotics and Automation*, pp. 6063–6070, 2025. doi: 10.1109/ICRA55743.2025.11127609.

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*, 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*, 2024.

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*, 2025.

A Additional Method Details

This section supplements the implementation details omitted from the main paper. We focus on probe correspondence, coordinate conventions, lifecycle supervision, and the execution flow of SPLF and LPD. Network dimensions and optimization hyperparameters are reported in Sec. B.

A.1 Probe Initialization and Coordinate Convention

SCOPE inherits N sparse query slots $q_i = (\mathbf{f}_i, \mathbf{c}_i)$ from the base occupancy forecaster. The same probe slots are evaluated across all forecast horizons, providing index-level correspondence for lifecycle modeling without assuming instance-level identity. Given the historical image features $\mathcal{F}_X$ and a target timestamp τ_h, the shared sparse decoder produces $\bar{Z}_i^h = (\bar{\mathbf{f}}_i^h, \bar{P}_i^h, \bar{S}_i^h)$, where $\bar{P}_i^h$ contains R candidate points and $\bar{S}_i^h$ contains their semantic logits. Its candidate center is $\bar{\mathbf{c}}_i^h = R^{-1} \sum_n \bar{\mathbf{p}}_{i,n}^h$.

All predicted probes and future occupancy targets are expressed in the ego frame of the current observation. Future ego poses, when required for annotation alignment, are used only by the offline data-processing pipeline and are never provided to the network. Candidate centers are projected into the input camera views using the known ego poses and camera calibration when sampling multi-view visual evidence.

A.2 Lifecycle Modeling and Weak Supervision

SPLF constructs $\mathbf{u}_i = \Phi_{\text{life}}[\mathbf{f}_i; \mathbf{c}_X; \psi_p(\mathbf{c}_i); \mathbf{h}_{\text{ego}}]$, where $\mathbf{c}_X$ is the pooled scene context. The lifecycle head predicts $(b_i, \ell_i, \alpha_i, \kappa_i^+, \kappa_i^-)$, with $d_i = b_i + \ell_i$. Sigmoid and softplus transformations constrain the activation amplitude, duration, and transition sharpness to valid ranges. Forecast timestamps are normalized over the prediction interval. Although continuous in τ, the support function is evaluated only at the current and benchmark forecast timestamps:

$$e_i(\tau \mid X) = \alpha_i \sigma\big(\kappa_i^+(\tau - b_i)\big)\, \sigma\big(\kappa_i^-(d_i - \tau)\big)\,. \tag{16}$$

We denote these evaluations by $e_i^h = e_i(\tau_h \mid X)$, including e_i^0 at the current timestamp, and derive $p_i^h = \min(e_i^{h-1}, e_i^h)$ and $\gamma_i^h = [e_i^h - e_i^{h-1}]_+$. Decreasing e_i^h suppresses slot support approaching termination.

Since lifecycle annotations are unavailable, we reuse the per-horizon point-to-occupancy assignment $\mathcal{M}_h$ from the base occupancy supervision. Each candidate point is independently assigned by 1-NN to an occupied ground-truth voxel; this detached geometric assignment does not enforce semantic or instance correspondence across horizons. The matched-support target is $y_i^h = R^{-1} \sum_n \mathbf{1}[(i,n) \in \mathcal{M}_h]$. We apply focal binary cross-entropy between e_i^h and y_i^h, and align $\bar{e}^h = N^{-1} \sum_i e_i^h$ with $\bar{y}^h = N^{-1} \sum_i y_i^h$ using SmoothL1 loss. Their sum forms $\mathcal{L}_{\text{life}}$. Reliability filtering is used only by the auxiliary dynamics losses.

A.3 Lifecycle-Routed Probe Dynamics

For horizon h, LPD infers the ego-conditioned context $\mathbf{z}_{\text{ego}}^h = \Phi_{\text{ego}}[\mathbf{h}_{\text{ego}}; \mathbf{c}_X; \psi_t(\tau_h)]$. It samples local evidence $\mathbf{v}_i^h = \text{MVSample}(\mathcal{F}_X, \bar{\mathbf{c}}_i^h)$ and pools probe semantics as $\bar{\boldsymbol{\pi}}_i^h = R^{-1} \sum_n \sigma(\bar{\mathbf{S}}_{i,n}^h)$. These representations predict the continuous-valued motion tendency m_i^h of the probe-supported region, rather than instance velocity.

A spatial neighborhood $\mathcal{N}_i^h$ is constructed from candidate centers. The pair descriptor $\mathbf{x}_{ij}^h$ combines feature compatibility, relative geometry, semantic similarity, existence, and mobility. MM relations use the neighbor gate $g_j^{\text{MM},h} = e_j^h m_j^h$, whereas MS relations use $g_j^{\text{MS},h} = e_j^h (1 - m_j^h)$. The gated relation coefficients are normalized within $\mathcal{N}_i^h$ and aggregated into $\mathbf{R}_i^{\text{MM},h}$ and $\mathbf{R}_i^{\text{MS},h}$.

LPD refines each horizon-specific candidate bundle rather than directly transporting the previous horizon's output. Lifecycle-derived persistence and emergence signals route the coordinate correction, while the fused probe context refines semantic logits. Existence scores separately weight the semantic scores before voxelization; no additional existence factor is applied to the routed coordinate correction.

Algorithm 1 Trajectory-free SCOPE forecasting pipeline.

Require: Historical images X, calibration, historical ego states, timestamps $\{\tau_h\}_{h=1}^{H}$
Ensure: Future occupancies $\{\hat{\mathcal{O}}^h\}_{h=1}^{H}$
Extract multi-scale features $\mathcal{F}_X$, scene context $\mathbf{c}_X$, and ego representation $\mathbf{h}_{\text{ego}}$
Generate candidate states $\{\bar{Z}_i^h\}_{i=1,h=1}^{N,H}$ with the base sparse decoder
for each probe slot i **do**
    Construct lifecycle feature $\mathbf{u}_i$
    Predict $(b_i, \ell_i, \alpha_i, \kappa_i^+, \kappa_i^-)$
    Evaluate e_i^h, p_i^h, and γ_i^h for all h
end for
for each horizon h in parallel **do**
    for each probe i **do**
        Sample $\mathbf{v}_i^h$ and predict mobility m_i^h
        Build $\mathcal{N}_i^h$ and aggregate MM/MS contexts
        Predict base, mobility, and emergence displacements
        Route $\Delta\mathbf{c}_i^h$ using p_i^h, γ_i^h, and m_i^h
        Refine $\bar{P}_i^h, \bar{S}_i^h$ into $\hat{P}_i^h, \hat{S}_i^h$
    end for
    Materialize $\hat{\mathcal{O}}^h = \text{Vox}(\{(\hat{P}_i^h, e_i^h \sigma(\hat{S}_i^h))\}_{i=1}^{N})$
end for
if training **then**
    Construct detached per-horizon 1-NN assignments $\mathcal{M}_h$ and support targets y_i^h
    Optimize occupancy, lifecycle, mobility, stability, and emergence objectives
end if
return $\{\hat{\mathcal{O}}^h\}_{h=1}^{H}$

A.4 Training and Inference

A slot is support-reliable at horizon h when at least 5% of its points (four of 64) lie within $0.8\,\text{m}$ of occupied ground truth. Mobility supervision uses slots reliable at both adjacent horizons, while emergence supervision uses slots reliable at the current horizon with $[y_i^h - y_i^{h-1}]_+ > 0$. These filters remove poorly supported slots but do not establish instance-level correspondence. The mobility target is obtained by normalizing adjacent-horizon matched-support centroid displacement with $s_m = 2.0\,\text{m}$: $\tilde{m}_i^h = \text{clip}(\|\mathbf{g}_i^h - \mathbf{g}_i^{h-1}\|_2/2.0, 0, 1)$. Displacements of $2.0\,\text{m}$ or larger therefore saturate the target at one.

The stability and emergence losses respectively suppress unnecessary motion residuals and guide newly activated support. The complete occupancy objective is

$$\mathcal{L} = \mathcal{L}_{\text{occ}} + \mathcal{L}_{\text{mob}} + \mathcal{L}_{\text{stb}} + \mathcal{L}_{\text{emg}} + \lambda_{\text{life}}\mathcal{L}_{\text{life}}. \tag{17}$$

For trajectory evaluation only, we train a lightweight head with a multi-horizon L_2 objective. This head and loss are disabled for all occupancy results, and its predictions are never fed back as forecasting conditions.

Algorithm 1 summarizes the complete pipeline. At inference, all forecast horizons are processed in parallel using only historical observations, without future trajectories, future poses, or autoregressive rollout.

B Implementation Details

B.1 Network Configuration

We use nuScenes (Caesar et al., 2020) with Occ3D occupancy annotations (Tian et al., 2023), following the main-paper protocol. Table 5 summarizes the principal configuration. Each sample contains the current timestamp and its four adjacent historical timestamps, with six surround-view cameras per timestamp. Images are processed at 256×704. The occupancy volume covers $[-40, -40, -1, 40, 40, 5.4]$ meters with a voxel size of 0.4 meters, resulting in a $200 \times 200 \times 16$ grid. The decoder predicts 17 occupied semantic classes, while the empty label is excluded from the occupied logits.

Table 5: Principal network configuration of SCOPE.

Configuration	Value
Input sequence	5 frames, 6 cameras
Image resolution	256×704
Occupancy grid	$200 \times 200 \times 16$
Voxel size	$0.4 \times 0.4 \times 0.4$ m
Probe number N	1,340
Points per probe R	64
Decoder layers / dimension	$6/256$
Attention heads / FFN dimension	$8/512$
FPN levels / channels	$4/256$
Refine points per layer	$1, 4, 16, 24, 32, 64$
Relation construction	top-768, 12-NN
Visual refresh	top-384 probes

We evaluate two image encoders. The ResNet-50 (He et al., 2016) variant uses stage channels $[256, 512, 1024, 2048]$, with its stem and first residual stage frozen. Its backbone is initialized from a Cascade Mask R-CNN model pretrained on COCO and NuImages (Caesar et al., 2020). The DINOv3 (Siméoni et al., 2025) variant uses ConvNeXt-Base (Liu et al., 2022) with stage channels $[128, 256, 512, 1024]$, initialized from the DINOv3 LVD-1689M checkpoint, and keeps all stages trainable. For both variants, only the image backbone is initialized from pretrained weights; the FPN, sparse decoder, and SCOPE modules are trained from scratch.

Both encoders feed a standard four-level FPN (Lin et al., 2017) with 256 channels per level. The resulting feature resolutions are approximately 64×176, 32×88, 16×44, and 8×22. The sparse decoder contains six layers with four sampling groups and four sampling points per group, frame, and FPN level. The relation module uses a hidden dimension of 256. Visual refresh samples the centers of the top-384 probes from all four FPN levels and all 30 input camera views.

B.2 Optimization

Both variants are trained for 40 epochs on four NVIDIA A100 GPUs with a batch size of four per GPU. We use AdamW (Loshchilov & Hutter, 2019) with weight decay 0.01, a base learning rate of 2×10^{-4}, linear warmup, and cosine annealing to a minimum learning-rate ratio of 10^{-3}. The ResNet-50 variant uses 500 warmup iterations. The DINOv3 variant uses 1,500 warmup iterations and a backbone learning-rate multiplier of 0.005, while its FPN and prediction heads retain the base learning rate. Gradients are clipped using a global L_2 norm of 5, without gradient accumulation.

Training uses BF16 autocast, while model parameters and optimizer states remain in FP32. No gradient scaler or loss scaling is applied. Numerically sensitive operations, including interpolation, `grid_sample`, KNN, and multi-scale multi-view sampling, are locally evaluated in FP32. TF32 is enabled for CUDA matrix multiplication and cuDNN, and activation checkpointing is used for both backbones.

We maintain an exponential moving average updated after every optimizer step, with $d_t = 0.999(1 - \exp(-u_t/2000))$. The update counter is initialized to 10,560, yielding an initial decay of approximately 0.9939, which gradually approaches 0.999. Training follows a shared curriculum: the first five epochs optimize current-scene reconstruction, after which full future occupancy forecasting is enabled.The selection of the lifecycle-loss weight λ_{life} is analyzed in Sec. C.3.

B.3 Data Processing and Augmentation

We always use the current frame and the four nearest historical frames, without temporal dropout or randomized frame intervals. Original 900×1600 images are resized using a ratio sampled uniformly from $[0.38, 0.55]$, followed by a bottom-aligned 256×704 crop with random horizontal position. Zero padding is applied on the right when the resized width is insufficient. Horizontal flipping is applied with probability 0.5, while image rotation is disabled.

Table 6: Effect of probe budget N and bundle size R.

Setting	N	R	Avg mIoU	Avg IoU	FPS
Probe budget	1,100	64	19.66	39.52	12.06
	1,220	64	20.78	42.17	11.24
	1,340	64	**21.02**	<u>42.99</u>	10.79
	1,460	64	<u>20.97</u>	**43.10**	9.44
	1,580	64	**21.02**	42.81	7.58
Point bundle	1,340	48	20.11	41.26	12.69
	1,340	56	20.68	42.58	11.50
	1,340	64	**21.02**	**42.99**	10.79
	1,340	72	20.89	<u>42.90</u>	9.60
	1,340	80	<u>20.93</u>	42.83	8.02

Table 7: Effect of the relation-neighborhood size.

K	mIoU@3s	Avg mIoU	Avg IoU	FPS
4	17.39	20.55	42.14	11.07
8	17.56	20.60	42.57	10.86
12	<u>17.87</u>	**21.02**	**42.99**	10.79
16	**17.89**	20.93	**42.99**	10.63
24	17.82	<u>21.00</u>	<u>42.91</u>	10.38

The same geometric transformation is shared by all 30 images in a sample, and the corresponding projection matrices are updated accordingly, preserving cross-view and cross-time geometric consistency. Photometric transformations are sampled independently for each image with probability 0.5, including brightness $[-32, 32]$, contrast and saturation $[0.5, 1.5]$, hue $[-18°, 18°]$, and random channel permutation. Images are then converted from BGR to RGB, normalized using ImageNet statistics, and padded to a divisor of 32.

C Additional Ablation Studies

C.1 Number of Probes and Points

Table 6 examines the accuracy–efficiency trade-off of the probe budget N and point-bundle size R. Increasing N from 1,100 to 1,340 improves average mIoU and IoU by 1.36 and 3.47 points, respectively, indicating that a small probe budget limits scene coverage. Further increasing N yields inconsistent accuracy changes: $N = 1{,}460$ improves average IoU by only 0.11 points, while $N = 1{,}580$ matches the average mIoU of $N = 1{,}340$ and reduces average IoU by 0.18 points. Meanwhile, FPS decreases from 10.79 to 9.44 and 7.58, respectively.

C.2 Relation Neighborhood Size

Table 7 examines the number of nearest neighbors K used for relational aggregation. Increasing K from 4 to 12 improves mIoU@3s, average mIoU, and average IoU by 0.48, 0.47, and 0.85 points, respectively, showing that an overly small neighborhood provides insufficient interaction context. Beyond $K = 12$, performance largely saturates: $K = 16$ yields only a 0.02-point improvement in mIoU@3s while reducing average mIoU and FPS, and $K = 24$ introduces further computation without accuracy gains. We therefore use $K = 12$, which provides the best overall accuracy–efficiency trade-off.

C.3 Lifecycle Hyperparameters

Table 8 analyzes the lifecycle-loss weight λ_{life} and mobility normalization scale s_m. Increasing λ_{life} from 0.6 to 0.8 consistently improves all metrics, whereas larger weights gradually reduce perfor-

Table 8: Sensitivity to lifecycle-loss weighting and mobility normalization.

Parameter	Value	mIoU@3s	Avg mIoU	Avg IoU
λ_{life}	0.6	17.42	20.38	41.24
	0.7	17.54	20.60	41.86
	0.75	17.66	20.78	42.16
	0.8	**17.87**	**21.02**	**42.99**
	0.85	17.78	20.96	42.85
	0.9	17.78	20.95	42.74
	1.0	17.69	20.86	42.48
s_m (m)	1.0	17.69	20.80	42.61
	2.0	**17.87**	**21.02**	**42.99**
	4.0	17.10	20.33	42.16

Table 9: Sensitivity to future-trajectory conditions. Perturbations are applied only to the predicted trajectory at inference. SW-TC denotes SparseWorld-TC (Du et al., 2026).

Setting	mIoU		IoU	
	3s	Avg	3s	Avg
SW-TC + GT Traj.	23.35	25.60	47.23	49.02
SW-TC + Pred. Traj.	18.06	21.60	41.83	44.91
+ Trans. 0.25 m	16.93	19.82	38.47	42.13
+ Trans. 0.50 m	15.27	18.22	35.62	39.88
+ Trans. 1.00 m	13.49	16.36	32.86	38.03
+ Yaw 2°	15.90	18.77	36.12	40.09
+ Yaw 5°	12.37	16.08	32.25	37.61
SW-TC (no Traj.)	12.16	15.99	32.97	36.91
SCOPE (no Traj.)	17.87	21.02	39.98	42.99

mance, suggesting that excessive lifecycle supervision can interfere with occupancy optimization. The results remain relatively stable around 0.75–0.9, with $\lambda_{\text{life}} = 0.8$ achieving the best overall accuracy. For mobility normalization, $s_m = 2.0\,\text{m}$ outperforms both 1.0 and $4.0\,\text{m}$. A smaller scale causes moderate displacements to saturate the mobility target too early, while a larger scale compresses mobility differences. We therefore use $\lambda_{\text{life}} = 0.8$ and $s_m = 2.0\,\text{m}$ in all experiments.

C.4 Trajectory-Condition Sensitivity

Table 9 evaluates the sensitivity of SparseWorld-TC (Du et al., 2026) to future-trajectory conditions. Replacing the ground-truth trajectory with a predicted one already reduces average mIoU and IoU by 4.00 and 4.11 points, respectively. Additional translation or yaw perturbations lead to monotonic degradation. A $0.25\,\text{m}$ translation error decreases the predicted-trajectory results by 1.78 average mIoU and 2.78 average IoU, while a $1.0\,\text{m}$ error causes drops of 5.24 and 6.88 points. Similarly, a 5° yaw error reduces the two metrics by 5.52 and 7.30 points. At the 3-second horizon, the 1.0 m translation and 5° yaw perturbations reduce IoU by 8.97 and 9.58 points, respectively, relative to the unperturbed predicted-trajectory setting.

This sensitivity is consistent with the architecture of SparseWorld-TC (Du et al., 2026): the future-to-current transformation both modulates all future-query features and transforms query coordinates for multi-view feature sampling. An inaccurate trajectory therefore introduces a confident but incorrect geometric prior into two critical pathways, potentially sampling unrelated objects, cameras, or out-of-view regions. Subsequent temporal attention may further propagate these errors across horizons. In contrast, SCOPE requires no trajectory condition and achieves 21.02 average mIoU and 42.99 average IoU, outperforming all perturbed SparseWorld-TC settings and narrowing the gap to its unperturbed predicted-trajectory result to 0.58 mIoU and 1.92 IoU.

Table 10: Efficiency and complexity under the camera-only setting. Training time is measured on four A100 GPUs with a per-GPU batch size of 2; all methods use the same hardware and precision. Baselines are COME (Shi et al., 2026), I^2-World-STC (Liao et al., 2025), SparseWorld (Dang et al., 2026), and SparseWorld-TC (Du et al., 2026).

Method	Params (M)	FLOPs (G)	Mem. (GB)	FPS ↑	Train (min/ep.)
COME	520.80	2066.6	10.55	2.37	50.79
I^2-World-STC	22.68	1154.31	2.92	9.40	11.78
SparseWorld	76.03	1615.52	9.74	8.00	47.32
SparseWorld-TC	79.43	5105.38	19.74	9.44	84.76
SCOPE	79.34	1726.68	13.69	10.79	64.92

D Efficiency and Complexity

All models are evaluated under the camera-only setting using the same implementation environment and numerical precision. Training time is measured on four A100 GPUs with a per-GPU batch size of 2, while inference efficiency follows an identical timing protocol for all methods.

Table 10 compares model complexity and runtime efficiency. SCOPE contains 79.34M parameters, remaining close to SparseWorld (Dang et al., 2026) (76.03M) and SparseWorld-TC (Du et al., 2026) (79.43M), while using 84.77% fewer parameters than COME (Shi et al., 2026). Compared with SparseWorld-TC, SCOPE reduces FLOPs by 66.18%, peak memory by 30.65%, and training time by 23.41%, while increasing inference speed from 9.44 to 10.79 FPS. This substantial reduction indicates that avoiding trajectory-conditioned query transformation and repeated trajectory-guided feature sampling yields a considerably lighter forecasting pipeline.

Relative to SparseWorld, SCOPE introduces only 4.35% more parameters and 6.88% more FLOPs, despite adding lifecycle modeling and MM/MS relational dynamics. These modules increase peak memory and training time, but SCOPE still improves inference speed from 8.00 to 10.79 FPS. I^2-World-STC (Liao et al., 2025) remains the most compact model and the fastest to train, whereas COME has substantially more parameters and the lowest inference throughput. Overall, SCOPE achieves the highest FPS among the compared methods while maintaining moderate model complexity, providing a favorable accuracy–efficiency trade-off when combined with the forecasting results in the main paper.

E Additional Qualitative Results

E.1 Additional Probe Lifecycle Visualizations

Figure 5 extends the lifecycle analysis to six additional scenes with different road layouts and occupancy structures. Persistent probes consistently retain high support across the forecasting interval, indicating stable representation demand. Emerging probes start with weak support and become strongly activated at later horizons, although their activation may plateau or slightly decrease after the relevant region has been established. Terminating probes generally remain active through the intermediate horizon and are then sharply suppressed toward 3 s; their support may first increase when the associated region remains useful before entering the termination phase. The spatial probe activations agree with the predicted curves: newly supported probes become visible at later timestamps, whereas decaying probes are faded or removed. The varying transition times and curve shapes across scenes demonstrate that SPLF learns scene-dependent temporal participation rather than assigning a shared horizon-wise schedule. These lifecycles describe representation support and do not imply object-level identity, emergence, or termination.

E.2 Additional Occupancy Forecasts

Figure 6 presents additional results across diverse road layouts and scene structures. SCOPE accurately reconstructs the current scene and preserves the main road geometry, buildings, vegetation, barriers, and occupied regions over future horizons. The predictions remain structurally consistent

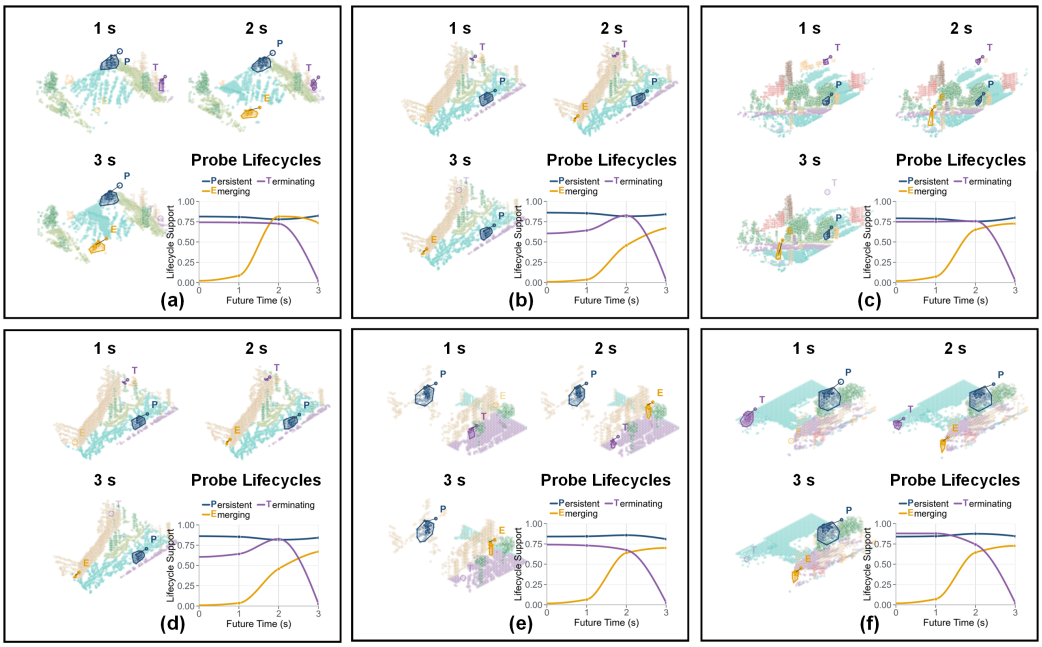

Figure 5: Additional probe lifecycle visualizations across six scenes. In each scene, persistent (P), emerging (E), and terminating (T) probe slots are tracked at 1–3 s, with their predicted lifecycle-support curves shown in the lower-right panel.

with the ground truth as the forecast horizon increases, indicating that the learned probe support generalizes beyond the single example shown in the main paper. Most long-horizon errors occur in distant boundaries, thin structures, small dynamic objects, and sparsely observed regions, where historical visual evidence and matched probe support are limited. Overall, these examples further demonstrate stable trajectory-free forecasting over different scenes while also revealing the expected degradation at the 3-second horizon.

F Additional Discussion

F.1 How Lifecycles Compensate for Missing Trajectory Guidance

Future ego trajectories benefit occupancy forecasting through two closely related effects. First, they provide temporal anchoring by associating each forecast horizon with an explicit future ego state. Second, they provide spatial guidance by transforming future representations and sampling locations according to the anticipated ego motion (Du et al., 2026). These signals substantially reduce the ambiguity of forecasting in an ego-centered coordinate system, but their usefulness depends on the availability and accuracy of an upstream planner or trajectory predictor. As shown in Table 9, perturbing the trajectory supplied to a trajectory-conditioned model progressively degrades both mIoU and IoU because the same erroneous geometric prior affects feature modulation and multi-view sampling.

SCOPE does not explicitly reconstruct a future ego trajectory. Instead, it learns internal representation-control signals from historical images, historical ego states, scene context, and the target timestamp. SPLF determines the temporal support of each probe slot through e_i^h, p_i^h, and γ_i^h, thereby deciding which representation capacity should be inherited, newly activated, or suppressed at each horizon. LPD then combines these signals with inferred ego-conditioned context, local visual evidence, mobility, and MM/MS relations to determine how supported probes should evolve spatially. The resulting mechanism recovers part of the temporal organization and spatial guidance otherwise provided by trajectory conditioning, while remaining independent of future trajectory inputs.

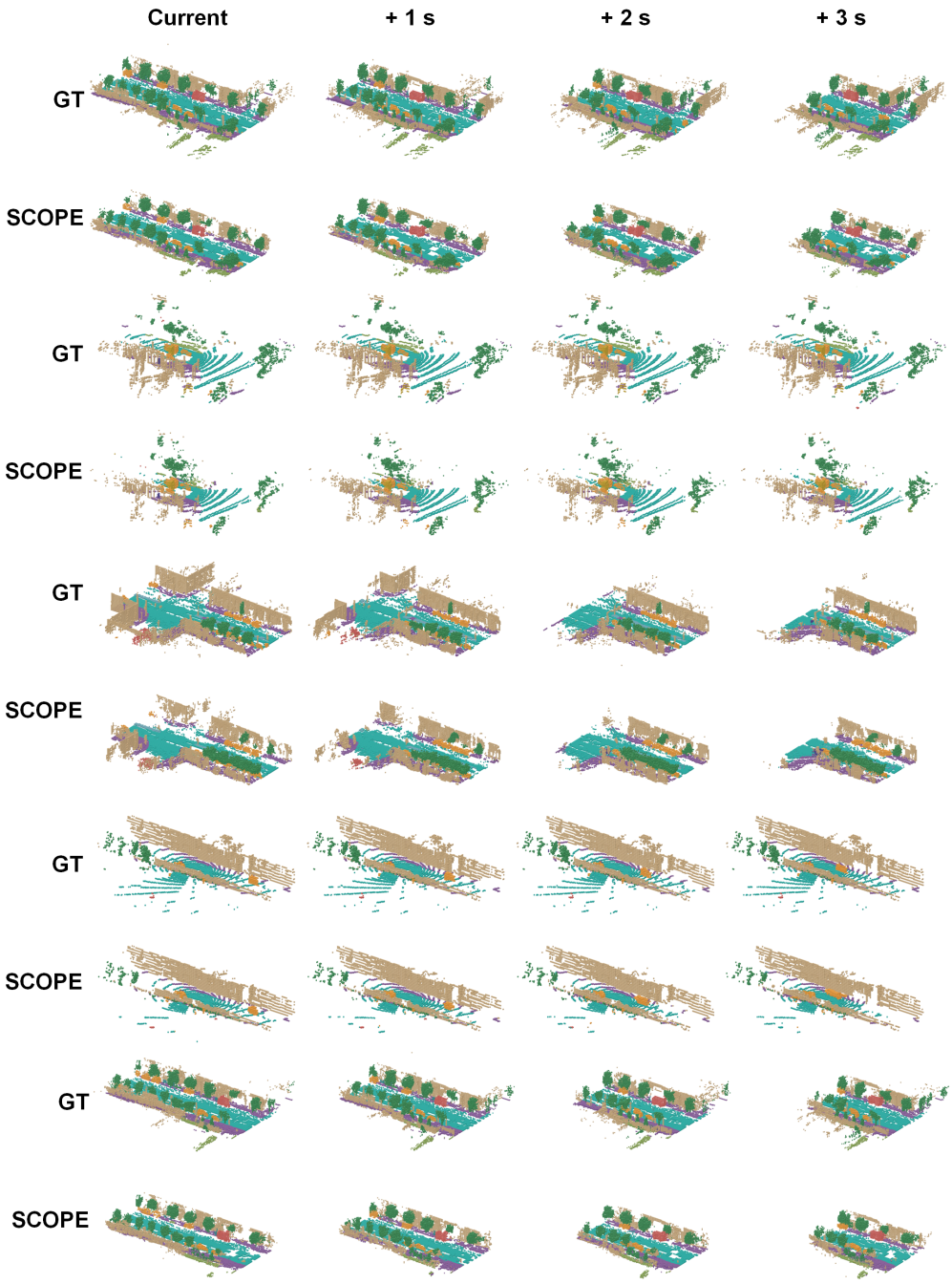

Figure 6: Additional qualitative occupancy forecasts on Occ3D-nuScenes (Tian et al., 2023). For each scene, ground truth and SCOPE results are shown at the current timestamp and 1–3 s future horizons using identical viewpoints and semantic colors.

This compensation is necessarily partial. A ground-truth future trajectory directly reveals future ego motion, whereas SCOPE must infer scene evolution from incomplete historical evidence. Accordingly, the strongest ground-truth-trajectory methods remain more accurate. Nevertheless, the ResNet-50 SCOPE variant substantially narrows the gap to predicted-trajectory forecasting, and the DINOv3 variant reaches comparable performance without receiving trajectories at inference. The intended conclusion is therefore not that probe lifecycles are equivalent to future trajectories, but

that scene-conditioned temporal support and routed dynamics can recover a meaningful portion of their forecasting benefit.

F.2 Interpretation of Probe Lifecycles

The lifecycle variables describe the temporal participation of representation slots rather than object-level birth, death, or identity. Probe indices are shared across horizons to provide architectural slot correspondence, but the point-to-occupancy assignments are constructed independently at each horizon and do not enforce instance tracking. Consequently, a persistent probe denotes inherited representation support, an emerging probe denotes newly activated support, and a terminating probe denotes support whose contribution is being suppressed. Similarly, the mobility score represents the motion tendency of the probe-supported region rather than the velocity of a specific tracked object.

This distinction is important for interpreting both the supervision and the visualizations. The matched-support target y_i^h is a weak proxy derived from geometric occupancy assignment, not an object-existence label. The qualitative lifecycle examples therefore illustrate how the model allocates sparse representation capacity over time; they should not be interpreted as instance-level trajectories or object lifecycle annotations.

F.3 Limitations

SCOPE has several limitations. First, lifecycle supervision relies on weak point-to-occupancy matching. Although support-reliability filtering removes poorly matched probes, it does not guarantee semantic or instance consistency across horizons. Slot reassignment or changing support composition may therefore affect the mobility and emergence targets, especially in crowded or ambiguous scenes.

Second, SCOPE uses a fixed maximum probe budget. SPLF adaptively changes the effective activation mass, but all candidate probes are still decoded before materialization; the current implementation therefore does not convert lifecycle sparsity into fully dynamic computation. Future work may introduce conditional decoding or early pruning to translate learned support into additional runtime savings.

Third, evaluation is currently limited to Occ3D-nuScenes and the benchmark horizons at 1, 2, and 3 seconds. The temporal support function is continuous in form but is evaluated only at these target timestamps, and the current results do not establish arbitrary-time forecasting. Validation across different sensor layouts, traffic densities, geographic domains, and annotation protocols remains necessary.

Finally, parallel horizon prediction avoids autoregressive error accumulation but does not explicitly represent multimodal long-term uncertainty or enforce full cross-horizon consistency. Future extensions could combine uncertainty-aware lifecycle prediction, adaptive probe allocation, and probabilistic scene evolution while preserving the trajectory-free protocol.

F.4 Broader Impact

Trajectory-free camera-only occupancy forecasting may reduce dependence on privileged future information and simplify deployment in systems where planning and perception operate asynchronously. Sparse scene-conditioned representations can also support lower-latency forecasting and provide interpretable indicators of where representation capacity is allocated over time.

However, occupancy forecasts remain susceptible to occlusion, distant objects, rare classes, adverse weather, calibration errors, and domain shift. Incorrect forecasts may affect downstream planning in safety-critical autonomous systems. SCOPE should therefore be treated as a predictive perception component rather than a safety guarantee, and should be combined with uncertainty estimation, redundant sensing, runtime monitoring, and conservative planning. Dataset bias may also cause uneven performance across geographic regions, object categories, and traffic conditions. Broader evaluation and explicit reporting of failure cases are required before real-world deployment.