FreeOcc: Decoupling Ego-Motion for Efficient 4D Occupancy Forecasting via Continuous Flow Matching

Zeping Zhang
University of Ottawa
Zeping.Zhang@uottawa.ca

Zhuoya Zhao
University of Ottawa
zoezhao0427@gmail.com

Samy Metari
National Research Council Canada
Samy.Metari@nrc-cnrc.gc.ca

Robert Laganière*
University of Ottawa
laganier@uottawa.ca

Abstract

Current 4D occupancy world models use heavy cross-attention or deep recurrence to disentangle ego-motion from scene dynamics. This requires significant computational capacity. However, this effort is largely redundant since the ego-vehicle's intended future motion is provided by the downstream motion planner. We present FreeOcc, a lightweight continuous-time world model that introduces explicit $SE(2)$ geometric decoupling into a conditional flow-matching framework. By mathematically factoring out self-motion, we create a purely dynamic feature space. This space supports attention-free temporal fusion and allows physics-constrained ODEs to operate strictly on residual agent kinematics. To maximize deployment efficiency, FreeOcc introduces an uncertainty-aware blending mechanism. This routes well-observed voxels to a single-pass decoder, reserving the ODE solver exclusively for ambiguous, high-uncertainty regions. Requiring only 28.3M parameters—roughly 40% of comparable models—our optimal ensemble mode (FreeOcc-E) runs at 27 FPS. On nuScenes, FreeOcc achieves state-of-the-art forecasting performance with a 0.47 m Near Field Chamfer Distance (NFCD) and 32.1% mIoU. Furthermore, the model transfers zero-shot to SemanticKITTI. It also proves highly robust, retaining 92.8% of its ideal mIoU under significant sensor packet loss and timestamp jitter.

1 Introduction

Anticipating the future 3D structure of a driving scene is critical for pre-emptive motion planning. Dense 4D occupancy forecasting addresses this by predicting a sequence of future voxelised binary volumes $\{O_\tau\}_{\tau=1}^{N}$ directly from recent sensor history, elegantly capturing complex scene geometry without relying on explicit object detection or tracking.

Recent works tackle this via: autoregressive token generation Zheng et al. [2024], Zhang et al. [2024], continuous flow formulations Shi et al. [2024], geometric proxy representations Yu et al. [2026], and unsupervised implicit fields Agro et al. [2024], Yang et al. [2024]. However, these methods generally rely on cross-attention or deep recurrence to separate the ego-vehicle's motion from observed scene dynamics. This entanglement inflates parameter counts. It also requires expensive temporal attention and complicates error control at long horizons. Crucially, the ego-vehicle's future trajectory is typically available at inference time as the intended route provided by the motion planner. While explicit $SE(2)$ warping in BEV space is well established Jaderberg et al. [2015], recent generative world models have largely favoured learned temporal attention instead. In this setting, learned

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

alignment is unnecessary: a single geometric warp factors out ego-motion and leaves the model to focus on residual dynamics, occlusion evolution, and far-field extrapolation. Driven by this insight, we propose **FreeOcc**. By mathematically factoring out self-motion, our architecture simplifies the forecasting pipeline, yielding three primary contributions:

- **Attention-Free Temporal Fusion:** Removing ego-motion allows a lightweight ConvGRU to effectively capture residual dynamics. By avoiding dense cross-attention, FreeOcc processes a $200 \times 200 \times 16$ voxel resolution at 27 FPS. Furthermore, it achieves this with only 28.3M parameters, which is roughly 40% of the size of comparable world models.
- **Physics-Constrained Generation:** The motion-decoupled feature space stabilizes the conditioning for flow matching. Physical constraints—such as velocity smoothness, mass conservation, and ray-traced free-space—can be applied directly to agent kinematics without relying on restrictive geometric proxies.
- **Uncertainty-Aware Routing:** Decoupling self-motion separates true scene ambiguity—such as dynamic objects and occlusion boundaries—from misalignment artefacts. The decoder's per-voxel confidence is used to route well-observed regions to a single-pass decoder, while the continuous ODE solver handles highly uncertain areas.

On nuScenes, FreeOcc achieves state-of-the-art forecasting performance (0.47 m NFCD, 32.1% mIoU) and transfers zero-shot to SemanticKITTI. The model remains robust under significant sensor dropout and temporal jitter, retaining 92.8% of its ideal mIoU.

2 Related Work

4D occupancy forecasting. Recent world models approach future scene prediction through autoregressive token generation Zheng et al. [2024], discrete diffusion Zhang et al. [2024], implicit fields Agro et al. [2024], or camera-based point cloud forecasting Yang et al. [2024]. However, these paradigms typically dedicate massive capacity to entangling ego-motion and scene dynamics via deep recurrence or cross-attention. FreeOcc bypasses this bottleneck by explicitly factoring out self-motion via algebraic $SE(2)$ warping (see Section 3.4), allowing it to directly predict dense voxel grids via physics-constrained continuous flow matching (CFM) without relying on discrete tokenisation or implicit rendering.

Flow matching and continuous generative models. Continuous normalising flows trained via flow matching Lipman et al. [2023] offer a simulation-free generative objective. Recent works like Streaming Flow Shi et al. [2024], and latent flow models Liu et al. [2025] adapt this to occupancy and LiDAR streams. Yet, they still rely on computationally heavy dense cross-attention for temporal alignment. FreeOcc preserves the continuous expressivity of flow matching but replaces the attention bottleneck with an explicit geometric warp, further regularising the generative trajectory with physics-informed constraints (velocity smoothness, mass conservation, ray-traced free-space).

Physics-informed perception. While integrating physical priors into perception is well-established—such as range-aware encoding Lang et al. [2019] or heteroscedastic uncertainty Kendall and Gal [2017]—applying structural constraints to dense 4D forecasting remains challenging. Recent concurrent works like SuperOcc Yu et al. [2026] and DTT Xu et al. [2025] enforce geometric structure by restricting the output space to predefined proxies (e.g., superquadrics or delta residuals), which inevitably sacrifices fine-grained fidelity on irregular shapes. Rather than simplifying the output representation, FreeOcc guides the full-resolution generative trajectory at the *loss* level. This ensures physically plausible dynamics while retaining exact voxel-level structural fidelity.

3 Methodology

FreeOcc forecasts dense 4D occupancy from LiDAR with an explicit task division: geometry handles ego-motion, while learning handles residual scene dynamics. By mathematically isolating the vehicle's known trajectory via differentiable $\mathrm{SE}(2)$ warping, we remove the need for learned temporal alignment, allowing the generative model to focus entirely on independently moving agents, occlusion evolution, and long-horizon uncertainty.

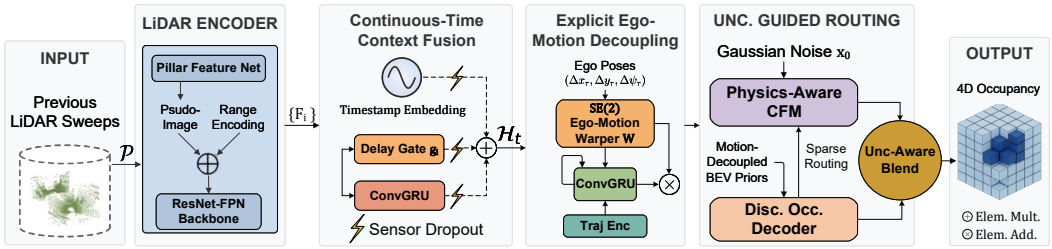

Figure 1: Overview of FreeOcc. Previous sweeps are encoded into BEV, fused by a continuous-time temporal aggregator, propagated to future ego frames by $\mathrm{SE}(2)$ warping, and decoded by the Discriminative Occupancy Decoder plus the Physics-Aware Continuous Flow Matching. The final prediction blends both paths using decoder uncertainty.

3.1 Overall Architecture

The architecture of FreeOcc (depicted in Figure 1) is driven by this principle of explicit geometric decoupling, operating in three distinct phases:

- **Continuous-Time Context Fusion (Sections 3.2 and 3.3):** Given K asynchronous previous sweeps $\mathcal{P} = \{P_{t-K+1}, \ldots, P_t\}$, we encode each sweep into BEV and fuse them via timestamp-aware recurrent aggregation into a unified present-state conditioning map $\mathcal{H}_t \in \mathbb{R}^{C\times H\times W}$.
- **Explicit Ego-Motion Decoupling (Section 3.4):** Rather than learning temporal alignment, an ego-aware propagator explicitly warps $\mathcal{H}_t$ to future coordinate frames. The resulting BEV priors $\{C_\tau\}_{\tau=1}^{N}$ isolate the residual scene changes.
- **Uncertainty-Guided Dual-Path Generation (Sections 3.5 to 3.7):** Each prior C_τ is decoded through two branches to maintain inference speed. Confident, well-observed regions with lower uncertainty pass through a fast discriminative decoder. The resulting uncertainty map then selectively routes only the highly uncertain areas to a physics-aware conditional flow-matching ODE.

We use the following notation throughout. BEV features have spatial size $H \times W$ and channel width C; voxel grids have size $X \times Y \times Z$ at resolution δ (m/voxel). The current time is t, the future index is $\tau \in \{1, \ldots, N\}$, and the internal flow time is $s \in [0, 1]$.

3.2 LiDAR Encoder with Sensor-Range Encoding

Each sweep $P_i = \{p_j\}_{j=1}^{M_i}$, where M_i is the number of points in the sweep and $p_j \in \mathbb{R}^5$ (coordinates, intensity, ring), is first processed by a Pillar Feature Net Lang et al. [2019] (see Figure 1), which encodes points into a sparse pillar representation. Specifically, a shared three-layer MLP maps each point within a pillar to $\mathbb{R}^C$, and per-pillar max-pooling yields a pseudo-image $\tilde{F}_i \in \mathbb{R}^{C\times H\times W}$.

We further add a learned sinusoidal range encoding $\rho(r_{hw})$ to each BEV cell, giving the encoder an explicit distance prior that reflects the drop in LiDAR sampling density with range. A lightweight ResNet–FPN backbone He et al. [2016], Lin et al. [2017a] then produces the per-sweep feature $F_i \in \mathbb{R}^{C\times H\times W}$; the exact encoding formula is given in Appendix A.1.

3.3 Continuous-Time Temporal Aggregator

After encoding the individual sweeps into spatial features F_i, the model must aggregate them over time. Real-world LiDAR data suffers from scheduling jitter, delays and packet loss. Instead of learning temporal alignment with attention, FreeOcc encodes each sweep at its measured arrival time and lets a lightweight ConvGRU absorb the irregular history.

For the i-th sweep at absolute time t_i, we compute $\Delta t_i = t_i - t \leq 0$ and encode it with an exponentially spaced sinusoidal timestamp embedding $\phi(\Delta t_i)$ (detailed in Appendix A). We then refine it with a learned projection $\psi : \mathbb{R}^C \to \mathbb{R}^C$. A delay-aware confidence gate discounts stale observations:

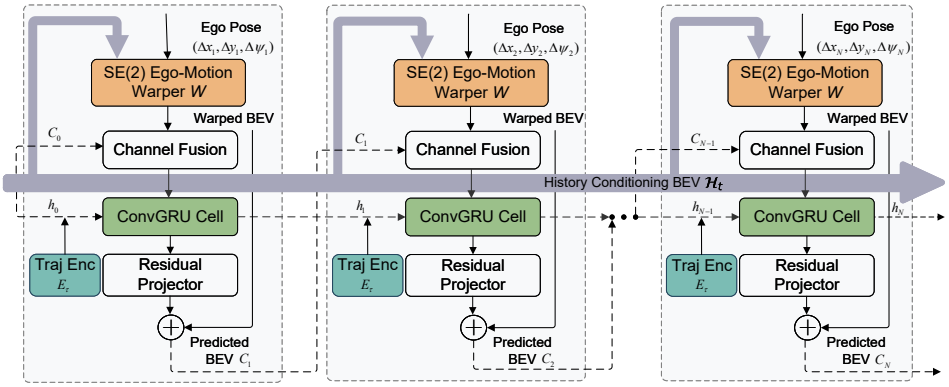

Figure 2: Ego-aware future propagator. At each horizon, the warped BEV—obtained via SE(2) ego-motion alignment—is explicitly added to the residual scene dynamics predicted by a ConvGRU and a residual projector.

$$g_i = \sigma\big(\mathrm{ReLU}(w_1\ \psi(\phi(\Delta t_i))) + b\big)\ \in\ (0,1), \tag{1}$$

where σ denotes the sigmoid activation function, w_1 is a learned weight matrix, and b is an additive bias initialized to $+2$ ($\sigma(2) \approx 0.88$).

The time-aware feature is $\tilde{F}_i = g_i \cdot F_i + \psi(\phi(\Delta t_i))$. During training, we additionally perturb sweep timestamps and randomly drop non-current sweeps so the GRU sees both asynchronous and incomplete histories. This combination improves robustness to timing noise and sensor dropout without adding attention parameters; the exact augmentation settings are given in Appendix A.1. The sequence $\{\tilde{F}_i\}$ is fused by a two-layer ConvGRU Cho et al. [2014], and the final hidden state is layer-normalised to yield $\mathcal{H}_t \in \mathbb{R}^{C\times H\times W}$.

3.4 Ego-Aware Future Propagator: Geometric Decoupling

With the historical context fused into the present-state representation $\mathcal{H}_t$, FreeOcc uses explicit geometric warping (Figure 2) to remove the dominant ego-induced motion before prediction. The resulting feature space is easier to learn for both the ConvGRU and the velocity network v_θ: the static structure is aligned by construction, and the remaining updates correspond to residual scene dynamics.

Given the intended future relative ego-pose $(\Delta x_\tau, \Delta y_\tau, \Delta\psi_\tau)$—recorded from odometry during training and provided by the onboard motion planner at deployment—we construct a 2×3 affine matrix:

$$A_\tau = \begin{bmatrix} \cos\Delta\psi_\tau & \sin\Delta\psi_\tau & \Delta y_\tau / R \\ -\sin\Delta\psi_\tau & \cos\Delta\psi_\tau & \Delta x_\tau / R \end{bmatrix}, \tag{2}$$

where R is the BEV half-range. The warping $\mathcal{W}$ applies this via bilinear grid sampling Jaderberg et al. [2015], mapping $\mathcal{H}_t$ to the future coordinate frame.

At each step τ, a ConvGRU cell fuses three inputs (the warped BEV, the previous output $C_{\tau-1}$, and a trajectory encoding $E_\tau = \mathrm{MLP}(\Delta x_\tau, \Delta y_\tau, \Delta\psi_\tau) \in \mathbb{R}^C$) via channel fusion ($1\times 1$ convolution):

$$h_\tau = \mathrm{ConvGRU}\big(\mathrm{Conv}_{1\times1}[\mathcal{W}(\mathcal{H}_t, A_\tau) \,\|\, C_{\tau-1} \,\|\, E_\tau],\ h_{\tau-1}\big), \tag{3}$$

with $h_0 = C_0 = \mathcal{H}_t$. A residual skip from the warped BEV ensures the ConvGRU learns only the *residual deviation*:

$$C_\tau = \mathrm{Proj}(h_\tau) + \mathcal{W}(\mathcal{H}_t, A_\tau), \tag{4}$$

where Proj denotes the Residual Projector (a 3×3 convolution followed by GroupNorm and ReLU), as illustrated in Figure 2. This factorisation separates independent-agent dynamics from ego-induced geometric shift and makes the velocity-field constraints in Section 3.5 well-defined.

3.5 Physics-Aware Continuous Flow Matching

Because the propagator has factored out the ego-vehicle's motion, the resulting priors C_τ isolate the independent agent dynamics. This decoupled feature space provides a highly stable condition for the optimal-transport flow matching problem Lipman et al. [2023]. Let $x_1 \in \{0,1\}^{X\times Y\times Z}$ denote the ground-truth binary occupancy and $x_0 \sim \mathcal{N}(0, I)$. The conditional probability path is:

$$x_s = (1-s)\, x_0 + s\, x_1, \quad s \in [0,1], \tag{5}$$

with target velocity $u(s) = x_1 - x_0$. A 3D U-Net velocity network v_θ is trained to predict this velocity. Let $\tilde{C}_\tau$ be the height-expanded version of the propagated BEV prior C_τ; the velocity network is directly conditioned on $\tilde{C}_\tau$.

Because the propagator already removes ego motion, the velocity field v_θ operates on residual agent dynamics. We regularise it with an anisotropic total variation so nearby voxels on the same rigid body move coherently:

$$\mathcal{L}_{\text{vel}} = \frac{1}{3} \sum_{d\in\{x,y,z\}} \frac{1}{|\Omega_d|} \sum \big| v_\theta^{(d+1)} - v_\theta^{(d)} \big|, \tag{6}$$

where the sum runs over adjacent voxels along each axis.

Cooperative uncertainty-guided training. We couple the decoder's uncertainty to the generative branch so ODE capacity is spent where single-pass prediction is least reliable. Let $\hat{s}_v$ denote the detached per-voxel log-variance predicted by the decoder (Section 3.6). The flow-matching loss is weighted as

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{s,x_0}\big[\, w(x_1, \hat{s}_v) \, \big\| v_\theta(x_s, s, \tilde{C}_\tau) - u(s) \big\|_2^2 \,\big], \tag{7}$$

where

$$\tilde{u}_v = \text{clamp}\left(\frac{\hat{s}_v - \mu_{\text{low}}}{\mu_{\text{high}} - \mu_{\text{low}}},\, 0,\, 1 \right), \qquad w(x_1, \hat{s}_v) = (1 + x_1) \cdot (1 + \tilde{u}_v). \tag{8}$$

Here μ_{low} and μ_{high} are EMA estimates of the 5th and 95th percentiles of $\hat{s}_v$ over training batches. The first factor prevents the trivial all-empty solution—most voxels are air or free space, so an unweighted model can reduce loss by predicting empty almost everywhere. The second factor upweights voxels where the decoder is uncertain, such as boundaries, dynamic objects, and far-field regions. Network details are given in Appendix A.1.

At inference, the ODE is solved from $s = 0$ to $s = 1$:

$$\frac{dx}{ds} = v_\theta(x, s, \tilde{C}_\tau), \quad x(0) \sim \mathcal{N}(0, I), \tag{9}$$

using a fixed-step midpoint integrator with S steps (default $S = 10$), yielding $\hat{O}_\tau^{\text{ode}} = \text{clamp}(x(1), [0,1])$. While continuous flow matching provides exceptional fidelity for complex scene dynamics, solving the ODE across the entire volume is computationally redundant for static, well-observed geometry. Therefore, we couple our flow-matching objective with a fast discriminative decoder (Section 3.6), directing the ODE's capacity strictly toward regions of high uncertainty.

3.6 Discriminative Occupancy Decoder

The second path lifts each BEV prior C_τ to a voxel volume with a shallow BEV-to-3D decoder and predicts both occupancy logits $\hat{y}_\tau \in \mathbb{R}^{X\times Y\times Z}$ and per-voxel log-variance $\hat{s}_\tau \in \mathbb{R}^{X\times Y\times Z}$. We add a future-step embedding before lifting; the exact decoder and head configuration is given in Appendix A.1. The decoder prediction is $\hat{O}_\tau^{\text{dec}} = \sigma(\hat{y}_\tau)$.

The per-voxel log-variance $\hat{s}_\tau$ serves a dual role: it enters the cooperative flow matching loss (Section 3.5) during training, and drives uncertainty-aware blending at inference (Section 3.7).

3.7 Uncertainty-Aware Ensemble Blending

The two generation paths offer complementary strengths: the decoder provides fast, high-recall predictions in well-observed regions, while the ODE yields high-precision samples at the cost of

iterative computation, excelling in ambiguous or dynamic areas. Rather than fixed blending (max or mean), we propose uncertainty-aware blending, which uses the decoder's own per-voxel aleatoric uncertainty to route each voxel to the appropriate path.

We convert the predicted log-variance $\hat{s}_v$ into a bounded confidence weight:

$$w_v = \sigma(-\hat{s}_v) \ \in \ (0, 1). \tag{10}$$

The final occupancy probability is a confidence-weighted interpolation:

$$\hat{O}_\tau(v) = w_v \, \hat{O}_\tau^{\mathrm{dec}}(v) + (1 - w_v) \, \hat{O}_\tau^{\mathrm{ode}}(v). \tag{11}$$

Low log-variance implies high confidence and therefore high w_v, so the blend favours the fast decoder in well-observed regions. Higher log-variance lowers w_v and shifts weight toward the ODE in ambiguous or far-field areas. The uncertainty head is trained jointly with the decoder, so the routing signal improves alongside prediction quality.

FreeOcc supports three inference modes with distinct speed/accuracy trade-offs:

- **FreeOcc-D** (Decoder-only): single forward pass, real-time;
- **FreeOcc-G** (Generative ODE-only): S-step ODE, highest precision;
- **FreeOcc-E** (Ensemble): uncertainty-aware blending, best IoU at acceptable latency;

3.8 Physics-Aware Training Objectives

FreeOcc is trained end-to-end with a composite loss coupling the flow-matching objective ($\mathcal{L}_{\mathrm{FM}}$) with three specialized regularization groups:

$$\begin{aligned}\mathcal{L} = \mathcal{L}_{\mathrm{FM}} &+ \underbrace{\lambda_{\mathrm{occ}}\mathcal{L}_{\mathrm{occ}} + \lambda_{\mathrm{lov}}\mathcal{L}_{\mathrm{Lov}} + \lambda_{\mathrm{dice}}\mathcal{L}_{\mathrm{Dice}} + \lambda_{\mathrm{sp}}\mathcal{L}_{\mathrm{sp}} + \lambda_{\mathrm{tc}}\mathcal{L}_{\mathrm{tc}}}_{\text{Structural Fidelity}} \\ &+ \underbrace{\lambda_{\mathrm{ray}}\mathcal{L}_{\mathrm{ray}} + \lambda_{\mathrm{cons}}\mathcal{L}_{\mathrm{cons}} + \lambda_{\mathrm{vel}}\mathcal{L}_{\mathrm{vel}} + \lambda_{\mathrm{rw}}\mathcal{L}_{\mathrm{rw}}}_{\text{Physical Priors}} + \underbrace{\lambda_{\mathrm{unc}}\mathcal{L}_{\mathrm{unc}}}_{\text{Uncertainty}},\end{aligned} \tag{12}$$

where λ denotes the hyperparameter weights. Structural fidelity supervises the discriminative decoder, while physical priors enforce ray-traced free-space, mass conservation, and piecewise-smooth kinematics (Eq. (6)). For uncertainty calibration, we employ a heteroscedastic classification loss:

$$\mathcal{L}_{\mathrm{unc}} = \frac{1}{|\Omega|} \sum_{ijk} \left[\exp(-\hat{s}_{ijk}) \, \ell_{\mathrm{BCE}}(\hat{y}_{ijk}, y_{ijk}) + \beta \hat{s}_{ijk} \right], \tag{13}$$

where ℓ_{BCE} denotes the binary cross-entropy loss applied to the predicted occupancy logits $\hat{y}_{ijk}$, and $\beta = 0.5$ is a fixed regularization penalty to prevent degenerate solutions. Detailed descriptions for each term are provided in Appendix A.2. We omit Occ3D Tian et al. [2023] because its focus on completing unobservable hidden geometry misaligns with our target task of sensor-grounded forecasting (see Appendix C.3 for a detailed discussion).

4 Experiments

4.1 Experimental Setup

Datasets. We train on nuScenes Caesar et al. [2020] and evaluate both in-domain and zero-shot on SemanticKITTI Behley et al. [2019]. For the cross-domain setting, we apply simple geometric calibration to the SemanticKITTI point clouds before evaluation. Dataset splits and calibration details are given in Appendix C.

Voxel Specification. We use a $[-40, 40]$m spatial range ($200\times200\times16$ grid at $\delta = 0.4$ m resolution) with $N = 6$ future and $K = 4$ history steps, but explicitly train a dedicated $[-70, 70]$m variant to ensure fair spatial comparisons against point-based baselines in Table 1.

Table 1: 3D Geometry Forecasting on nuScenes. Evaluated via Near Field Chamfer Distance (NFCD, m) ↓, L1 Depth Error (m) ↓, and Absolute Relative Error (AbsRel, %) ↓.

Method	1.0s (τ=2)			3.0s (τ=6)		
	NFCD	L1	AbsRel	NFCD	L1	AbsRel
4D-Occ† Khurana et al. [2023]	1.41	1.40	10.37	1.40	1.71	13.48
Copilot4D† Zhang et al. [2024]	**0.36**	1.30	8.58	0.58	1.51	10.38
UnO† Agro et al. [2024]	–	–	–	0.89	1.15	8.50
ViDAR† Yang et al. [2024]	0.85	1.55	11.50	1.18	1.82	14.20
FreeOcc-D (ours)	0.48	1.25	8.15	0.63	1.35	9.20
FreeOcc-G (ours)	0.45	1.20	7.85	0.52	1.15	8.10
FreeOcc-E (ours)	0.41	**1.18**	**7.60**	**0.47**	**1.08**	**7.45**

†Numbers from original papers. L1 Depth Error is an evaluation-only metric computed by projecting predicted occupied voxels to range images and comparing against GT LiDAR depths; no L1 loss is required during training.

Metrics & Evaluation Protocol. We report two complementary metric tracks:

- Voxel metrics: class-agnostic mIoU, precision, recall, and F1 on binarized occupancy grids;
- Point metrics: NFCD, L1 depth error, and AbsRel following prior point-forecasting work Agro et al. [2024], Zhang et al. [2024].

For point metrics, we convert predicted occupied voxels to voxel-centre point clouds. Metric alignment details are given in Appendix C.

Implementation. FreeOcc (28.3M parameters) is trained end-to-end on 4 NVIDIA RTX 4090 GPUs for 10 epochs with AdamW ($\mathrm{lr}=3\times10^{-4}$, cosine decay, 500-step warmup, $\mathrm{wd}=0.01$), batch size 2 with $2\times$ gradient accumulation, and FP16 mixed precision. ODE inference: $S=10$ midpoint steps.

Baselines. We compare against representative voxel- and point-based methods:

- 4D-Occ Khurana et al. [2023], OccWorld Zheng et al. [2024], and ViDAR Yang et al. [2024] for dense occupancy comparison;
- Copilot4D Zhang et al. [2024] and UnO Agro et al. [2024] for point-based structural geometry;
- **Static Copy**: Current-frame GT ego-warped to each future step, serving as a no-learned-dynamics lower bound.

For cross-modality fairness, point outputs are voxelized when evaluated on occupancy metrics, and our voxel predictions are converted to point clouds when evaluated on NFCD/L1. OccWorld is trained from scratch using the released implementation. Copilot4D and UnO are compared only on point metrics using values reported in their original papers†.

Note on recent concurrent work. Concurrent works SuperOcc Yu et al. [2026] and DTT Xu et al. [2025] rely on predefined geometric proxies. Crucially, SuperOcc targets current-frame 3D prediction rather than 4D forecasting, and DTT currently lacks publicly released code. Therefore, neither is directly applicable as a forecasting baseline under our protocol, and we include them strictly as conceptual comparisons.

4.2 Main Results on nuScenes

We report point-based geometry forecasting and voxel-based occupancy forecasting separately.

Geometry forecasting (Table 1). On the point-based track, FreeOcc-E reaches the strongest NFCD at τ=6 (0.47 m vs. 0.58 m for Copilot4D and 0.89 m for UnO). At τ=2, Copilot4D remains best on NFCD (0.36 vs. 0.41), while FreeOcc-E is best on L1 depth error and AbsRel. The gap closes at longer horizons: Copilot4D rises from 0.36 to 0.58 NFCD, whereas FreeOcc-E rises from 0.41 to 0.47. FreeOcc-D achieves the highest recall (49.4%) but struggles with false positives (41.3% precision). FreeOcc-E uses uncertainty-aware blending to explicitly suppress noise, a visualization is provided in Figure 8. This trade-off slightly lowers recall but boosts precision to 54.6%, securing the

Table 2: 4D Occupancy Forecasting on nuScenes. Evaluated via Class-Agnostic 3D Voxel mIoU (%) ↑.

Method	Forecasting Horizon mIoU (%)						Average Metrics (%)			
	τ=1	τ=2	τ=3	τ=4	τ=5	τ=6	mIoU	Prec.	Rec.	F1
Static Copy	31.5	22.1	14.8	10.5	8.8	7.5	15.9	34.0	23.5	27.8
4D-Occ* Khurana et al. [2023]	35.8	27.5	22.4	19.0	16.5	14.2	22.6	39.1	31.8	35.1
OccWorld Zheng et al. [2024]	39.7	34.2	27.9	26.4	23.6	21.6	28.9	50.5	40.2	44.8
ViDAR* Yang et al. [2024]	30.5	24.2	20.5	17.8	15.9	14.3	20.5	39.5	29.8	34.1
FreeOcc-D (ours)	40.8	34.3	29.1	25.1	21.8	19.2	28.4	41.3	**49.4**	44.9
FreeOcc-G (ours)	41.9	35.2	30.2	26.9	24.5	23.1	30.3	52.5	42.5	47.4
FreeOcc-E (ours)	**43.6**	**36.3**	**32.2**	**28.9**	**26.5**	**24.9**	**32.1**	**54.6**	45.5	**49.6**

*Indicates the models whose output point clouds we voxelize for evaluation.

best F1 score. FreeOcc-E also improves over FreeOcc-G, showing that uncertainty-aware blending is stronger than either branch alone.

Occupancy forecasting (Table 2). On the voxel-based track, FreeOcc-E achieves the highest average mIoU at 32.1%, improving over OccWorld by +3.2 points and over FreeOcc-G by +1.8 points. The ensemble is best at every horizon, reaching 24.9% mIoU at τ=6. FreeOcc-G consistently improves over FreeOcc-D, showing the benefit of continuous generation. The additional gain from FreeOcc-E indicates that uncertainty-aware blending balances the decoder's recall with the ODE's sharper long-horizon structure. Static Copy and 4D-Occ degrade much faster, and ViDAR remains limited by the point-to-voxel conversion used for dense occupancy evaluation.

4.3 Efficiency Comparison

Table 3 shows that FreeOcc keeps a strong accuracy–speed trade-off. FreeOcc-D runs at 45.4 FPS, FreeOcc-E sustains 27.0 FPS while improving accuracy over all baselines, and FreeOcc-G remains compact but slower because of iterative ODE integration. Detailed latency analysis under adverse conditions is deferred to Appendix H.

Table 3: Efficiency comparison on nuScenes. We evaluate the parameter count and inference speed of our three variants: D (Decoder-only), G (Generation-only), and E (Ensemble) against representative baselines. All measurements are conducted on a single NVIDIA RTX 4090.

Method	#Params	Latency (ms)	FPS
4D-Occ Khurana et al. [2023]	8.59M	73	13.7
OccWorld Zheng et al. [2024]	72.4M	56	17.8
ViDAR Yang et al. [2024]	69.1M	189	5.3
FreeOcc-D (ours)	20.7M	22	45.4
FreeOcc-G (ours)	7.5M	168	6.0
FreeOcc-E (ours)	28.3M	37	27.0

4.4 Zero-Shot Cross-Domain Generalisation

Table 4 evaluates cross-domain generalisation by deploying nuScenes-trained models directly on SemanticKITTI with no fine-tuning. Despite the substantial domain gap (different LiDAR beam count, sensor mounting height, and geographic driving context), FreeOcc-E remains clearly ahead of the baselines, outperforming OccWorld by +5.8 points at short term and +5.9 points at long term. The gap is larger at long horizons, where FreeOcc-E reaches 18.2% mIoU versus 12.3% for OccWorld and 8.9% for 4D-Occ. FreeOcc-E also improves over FreeOcc-G, again indicating that the ensemble is more robust than either branch alone under domain shift. Notably, while FreeOcc-D's aggressive generation yields higher recall at long horizons (33.2% compared to 30.8% for FreeOcc-E), it suffers from reduced precision. Additional SemanticKITTI visualizations are provided in Appendix F.

Table 4: Zero-shot transfer from nuScenes to SemanticKITTI (no fine-tuning). We report metrics across short-term and long-term horizons to assess structural preservation under distribution shift.

Method	Short-term ($\tau \leq 3$)			Long-term ($\tau > 3$)		
	mIoU	Precision	Recall	mIoU	Precision	Recall
Static Copy	14.5	30.2	22.1	7.5	22.5	13.0
4D-Occ Khurana et al. [2023]	15.2	31.8	24.4	8.9	24.1	15.2
OccWorld Zheng et al. [2024]	18.5	38.1	30.1	12.3	31.5	22.1
ViDAR Yang et al. [2024]	6.5	12.4	8.3	2.3	1.5	2.2
FreeOcc-D (ours)	22.0	42.5	**39.8**	15.5	34.1	**33.2**
FreeOcc-G (ours)	23.5	44.0	36.5	17.2	38.5	29.8
FreeOcc-E (ours)	**24.3**	**45.0**	37.5	**18.2**	**39.5**	30.8

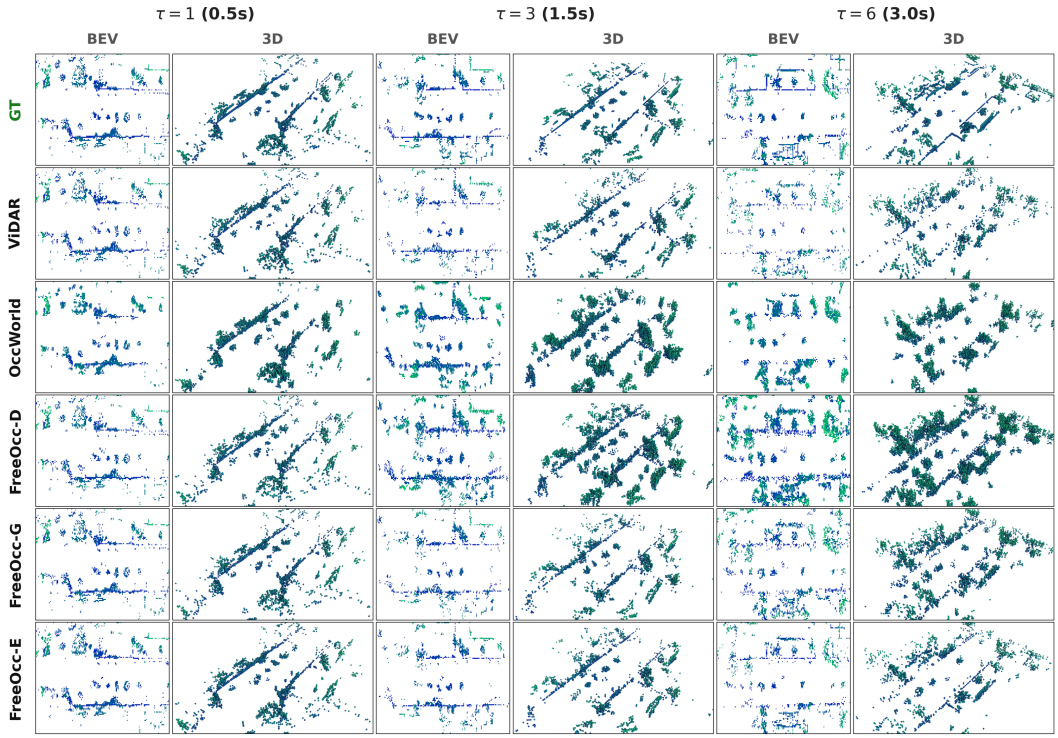

Figure 3: 4D occupancy forecasting on nuScenes(sample 1720) in BEV and 3D views across six horizons. FreeOcc-E is compared against OccWorld and ViDAR; more qualitative examples are given in Appendix D.

4.5 Robustness and Ablation Summary

Under timestamp jitter and 30% history dropout, FreeOcc-E maintains a 29.8% mIoU, compared with 19.0% for OccWorld, 14.2% for 4D-Occ, and 12.8% for ViDAR. In ablations, the largest single drop comes from removing ego-motion warping (32.1→25.8 mIoU), followed by removing flow matching (32.1%→28.4% mIoU). FreeOcc-E also remains stronger than FreeOcc-G, confirming that uncertainty-aware blending adds measurable gains beyond the ODE alone. Full quantitative results are reported in Appendix C.4.

4.6 Qualitative Results

Figure 3 shows the qualitative trend behind the tables. FreeOcc-E preserves cleaner occupancy boundaries and more stable long-horizon structure than OccWorld and ViDAR, especially in dynamic and far-field regions. Additional visualizations, uncertainty maps, and robustness examples can be found in Appendix D–Appendix G.

5 Conclusion

We introduce FreeOcc, a continuous-time world model for 4D occupancy forecasting. By explicitly decoupling ego-motion via SE(2) warping, we isolate residual scene dynamics. This formulation enables attention-free temporal fusion and physics-constrained continuous flow matching. To reduce inference latency, an uncertainty-aware routing mechanism processes well-observed geometry with a single-pass decoder, reserving ODE integration strictly for ambiguous regions. Operating with roughly 40% of the parameters of comparable models, FreeOcc achieves 32.1% mIoU on nuScenes at 27 FPS. Furthermore, the model demonstrates zero-shot transfer to SemanticKITTI and retains 92.8% of its mIoU under simulated sensor degradation. These results indicate that explicit geometric decoupling and physical priors provide an effective, robust foundation for autonomous driving perception.

Limitations and future directions are discussed in Appendix B.

Acknowledgements

This work was partially funded by the National Research Council Canada (NRC) and the Natural Sciences and Engineering Research Council (NSERC) of Canada.

References

Ben Agro, Quinlan Sykora, Sergio Casas, Thomas Gilles, and Raquel Urtasun. Uno: Unsupervised occupancy fields for perception and forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14487–14496, 2024.

Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Jürgen Gall. SemanticKITTI: A dataset for semantic scene understanding of LiDAR sequences. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 9297–9307, 2019.

Maxim Berman, Amal Rannen Triki, and Matthew B Blaschko. The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4413–4421, 2018.

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*, pages 11621–11631, 2020.

Kyunghyun Cho, Bart Van Merriënboer, Çağlar Gulçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1724–1734, 2014.

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*, pages 7425–7434, 2026.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 28, 2015.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

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*, pages 1116–1124, 2023.

Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12697–12705, 2019.

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*, pages 2117–2125, 2017a.

Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017b.

Yaron Lipman, Ricky T.Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

Tianran Liu, Shengwen Zhao, and Nicholas Rhinehart. Towards foundational lidar world models with efficient latent flow matching. *arXiv preprint arXiv:2506.23434*, 2025.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In *Proceedings of the International Conference on 3D Vision (3DV)*, pages 565–571, 2016.

Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *European conference on computer vision*, pages 194–210. Springer, 2020.

Yining Shi, Kun Jiang, Ke Wang, Jiusi Li, Yunlong Wang, Mengmeng Yang, and Diange Yang. Streamingflow: Streaming occupancy forecasting with asynchronous multi-modal data streams via neural ordinary differential equation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14833–14842, 2024.

Xiaoyu Tian, Tao Jiang, Longfei Yun, Yucheng Mao, Huitong Yang, Yue Wang, Yilun Wang, and Hang Zhao. Occ3D: A large-scale 3D occupancy prediction benchmark for autonomous driving. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Haoran Xu, Peixi Peng, Guang Tan, Yiqian Chang, Yisen Zhao, and Yonghong Tian. Delta-triplane transformers as occupancy world models. *arXiv preprint arXiv:2503.07338*, 2025.

Zetong Yang, Li Chen, Yanan Sun, and Hongyang Li. Visual point cloud forecasting enables scalable autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Zichen Yu, Quanli Liu, Wei Wang, Liyong Zhang, and Xiaoguang Zhao. Superocc: Toward cohesive temporal modeling for superquadric-based occupancy prediction. *arXiv preprint arXiv:2601.15644*, 2026.

Lunjun Zhang, Yuwen Xiong, Ze Yang, Sergio Casas, Rui Hu, and Raquel Urtasun. Copilot4d: Learning unsupervised world models for autonomous driving via discrete diffusion. In *The Twelfth International Conference on Learning Representations*, 2024.

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*, pages 55–72. Springer, 2024.

A Additional Method Details

A.1 Temporal augmentation and decoder details

Range encoding. For BEV cell (h, w) with radial distance $r_{hw} = \sqrt{x_h^2 + y_w^2}$, we use

$$\rho_{hw} = W_2 \, \mathrm{ReLU}\big(W_1 \, [\sin(2^0 r), \cos(2^0 r), \ldots, \sin(2^{B_r - 1} r), \cos(2^{B_r - 1} r)]\big), \tag{14}$$

and add it to the pillar pseudo-image as $\bar{F}_i \leftarrow \bar{F}_i + \rho$.

Timestamp jitter and sensor dropout. To train the temporal aggregator on irregular input streams, each history sweep P_i $(i < t)$ is assigned a perturbed offset

$$\Delta t_i' = \Delta t_i + \eta_i, \qquad \eta_i \sim \mathcal{U}(-\delta_t, \delta_t), \tag{15}$$

with $\delta_t = 0.1\,\mathrm{s}$. The current frame is never perturbed. In parallel, every non-current sweep is dropped independently with probability $p_{\mathrm{drop}} = 0.3$, while forcing at least one earlier sweep to remain. Jitter trains continuous interpolation over residual timing noise, whereas dropout trains recovery from incomplete histories.

The timestamp embedding itself uses exponentially spaced sinusoidal frequencies:

$$\phi(\Delta t_i) = \big[\sin(\omega_1 \Delta t_i), \, \cos(\omega_1 \Delta t_i), \, \ldots, \, \sin(\omega_{C/2} \Delta t_i), \, \cos(\omega_{C/2} \Delta t_i)\big], \tag{16}$$

where $\omega_j = \exp(-j \ln \Lambda / (C/2))$.

Decoder architecture. Each BEV prior C_τ is refined by two 3×3 BEV convolutions, lifted to $C \cdot Z$ channels by a 1×1 convolution, reshaped to $\mathbb{R}^{C \times H \times W \times Z}$, augmented with a height positional encoding z_{pos}, and processed by two $3 \times 3 \times 3$ volume convolutions. A future-step embedding e_τ is added before lifting. The occupancy head maps $C \to C/2 \to C/4 \to 1$, while the uncertainty head predicts $\hat{s}_\tau$ with a two-layer MLP whose bias is initialised to -0.5.

Velocity network details. The velocity network v_θ is a 3D U-Net with channel widths $\{c, 2c, 4c\}$, a bottleneck, and symmetric skip connections. Flow time s is injected with a sinusoidal embedding, and each future step τ has a learned embedding. For diversity, a fraction $\alpha = 0.2$ of initial Gaussian samples is replaced by slightly noised decoder outputs.

Uncertainty normalisation for cooperative flow matching. The normalisation bounds μ_{low} and μ_{high} in Eq. (7) are maintained as exponential moving averages of the 5th and 95th percentiles of detached $\hat{s}_v$ computed over training batches. This ensures that the uncertainty weight w remains stable while effectively spending generative capacity strictly in regions where the discriminative path is least reliable.

A.2 Full loss definitions

Focal BCE $\mathcal{L}_{\mathrm{occ}}$ Lin et al. [2017b]. We use focal BCE with $\alpha = 0.75$ to upweight the minority occupied class and suppress easy free-space gradients.

Lovász hinge $\mathcal{L}_{\mathrm{Lov}}$ Berman et al. [2018]. This term is a differentiable surrogate for IoU. Its weight is warmed up as

$$\lambda_{\mathrm{lov}}(n) = \bar{\lambda}_{\mathrm{lov}} \cdot \min\left(1, \max\left(0, \frac{n - n_{\mathrm{start}}}{n_{\mathrm{ramp}}}\right)\right). \tag{17}$$

Dice $\mathcal{L}_{\mathrm{Dice}}$ Milletari et al. [2016]. Dice complements BCE and Lovász with a smooth overlap objective.

Sparsity $\mathcal{L}_{\mathrm{sp}}$. We penalise deviation from a target occupied ratio $\bar{r} = 0.05$ using $|\bar{\sigma}(\hat{y}) - \bar{r}|$, with a weight that grows linearly with horizon as $(1 + 0.5\tau)$.

Temporal consistency $\mathcal{L}_{\mathrm{tc}}$. We match predicted and ground-truth temporal derivatives:

$$\mathcal{L}_{\mathrm{tc}} = \frac{1}{N - 1} \sum_{\tau} \big\| \big(\sigma(\hat{y}_{\tau+1}) - \sigma(\hat{y}_\tau)\big) - \big(y_{\tau+1} - y_\tau\big) \big\|_2^2 . \tag{18}$$

Ray-traced free-space $\mathcal{L}_{\text{ray}}$. LiDAR rays define observed free space before the first hit. With traversed voxel set $\mathcal{F}$, we use

$$\mathcal{L}_{\text{ray}} = \frac{1}{|\mathcal{F}|} \sum_{(i,j,k)\in\mathcal{F}} \sigma(\hat{y}_{ijk}). \tag{19}$$

Voxels behind the first surface hit remain unconstrained.

Conservation $\mathcal{L}_{\text{cons}}$. Under a closed-world approximation, total occupied mass should vary smoothly:

$$\mathcal{L}_{\text{cons}} = \frac{1}{N-1} \sum_{\tau} \big(\Delta\hat{M}_\tau - \Delta M_\tau^{\text{gt}}\big)^2, \tag{20}$$

where $\Delta\hat{M}_\tau$ is measured over the interior region Ω.

Range-weighted loss $\mathcal{L}_{\text{rw}}$. We modulate supervision by sensor distance using

$$w(r) = w_{\text{near}}(1 - r/r_{\max}) + w_{\text{far}}\, r/r_{\max}, \tag{21}$$

with $w_{\text{near}} = 1.5$ and $w_{\text{far}} = 0.5$.

Velocity smoothness $\mathcal{L}_{\text{vel}}$. Eq. (6) imposes piecewise-smooth motion on the ego-decoupled flow field.

Uncertainty calibration $\mathcal{L}_{\text{unc}}$. Following the heteroscedastic classification approach, we supervise the uncertainty head using:

$$\mathcal{L}_{\text{unc}} = \frac{1}{|\Omega|} \sum_{ijk} \left[\exp(-\hat{s}_{ijk})\, \ell_{\text{BCE}}(\hat{y}_{ijk}, y_{ijk}) + \beta\hat{s}_{ijk}\right], \tag{22}$$

where ℓ_{BCE} denotes the binary cross-entropy loss applied to the predicted occupancy logits $\hat{y}_{ijk}$. We fix $\beta = 0.5$ as a penalty to prevent the model from trivializing the objective by predicting infinite variance.

Hyperparameter configurations. For the composite objective defined in Eq. 12, the base scalar weights are empirically set as follows: $\lambda_{\text{occ}} = 2.0$, $\lambda_{\text{lov}} = 1.0$, $\lambda_{\text{dice}} = 1.0$, $\lambda_{\text{sp}} = 0.5$, $\lambda_{\text{tc}} = 1.0$, $\lambda_{\text{ray}} = 1.0$, $\lambda_{\text{cons}} = 0.1$, $\lambda_{\text{vel}} = 0.5$, $\lambda_{\text{rw}} = 1.0$, and $\lambda_{\text{unc}} = 1.0$. For the range encoding (Eq. 14), the frequency band limit is $B_r = 6$, and for the timestamp embedding (Eq. 16), the base period $\Lambda = 10000$. The scheduling parameters for the Lovász hinge warmup are $n_{\text{start}} = 1000$ and $n_{\text{ramp}} = 2000$ iterations.

B Limitations and Future Directions

FreeOcc does not model semantic classes; extending the framework to class-conditional flow matching remains future work. The ODE branch still adds non-trivial inference cost, which motivates future work on adaptive solvers or distillation.

Furthermore, the explicit geometric decoupling relies on a planar SE(2) ego-motion warp. This assumes flat-world dynamics, which may introduce minor spatial misalignments under severe pitch, roll, or steep inclines. However, enforcing SE(2) decoupling rather than a full 3D volumetric SE(3) warp drastically reduces the computational overhead of grid sampling. In practice, our evaluations demonstrate that the residual ConvGRU effectively absorbs minor Z-axis translations and pitch deviations, confirming that SE(2) alignment strikes the optimal balance between geometric accuracy and real-time inference latency for urban driving scenarios. Furthermore, the model treats the motion planner's trajectory as a known input; however, experiments show that a self-predicted trajectory extrapolated from the vehicle's recent speed and yaw rate, using no future information, still achieves 31.2% mIoU (vs. 32.1% with the recorded route), indicating that the method does not critically depend on an accurate external plan. Sensitivity to trajectory noise is analysed in Appendix I. Mitigating vulnerability to local odometry or planning drift remains an area for future refinement.

Finally, the point-based track retains a quantisation floor induced by the voxel resolution ($\delta = 0.4\,\text{m}$). Since δ is a preprocessing choice rather than an architectural constraint, finer grids or a learned sub-voxel refinement head could reduce this error floor at additional compute cost.

C Experimental Details

C.1 Dataset and transfer details

We train on the standard nuScenes split (700/150/150 train/val/test) with 32-beam LiDAR, 20 Hz sweeps, and 2 Hz keyframes. For zero-shot transfer, we evaluate on SemanticKITTI. Before evaluation, SemanticKITTI point clouds are range-clipped to $[-40, 40]$ m in the horizontal plane, shifted by 0.11 m along Z, and beam-downsampled to reduce the 64-to-32 line density mismatch.

C.2 Metric Alignment with Baselines

In the point-based evaluation track (Table 1), we report Near Field Chamfer Distance (NFCD) to align with UnO Agro et al. [2024].

Copilot4D Zhang et al. [2024] does not use the term "NFCD", but it reports Chamfer Distance within a fixed Region of Interest (ROI). This evaluation protocol is equivalent to the NFCD computation used by UnO. We therefore report Copilot4D's ROI-restricted Chamfer Distance in the NFCD column.

C.3 Discussion: Sensor-Grounded vs. Completion-Based Evaluation

We do not evaluate on Occ3D Tian et al. [2023] because it measures a different task. Occ3D targets *occupancy completion*: reconstructing the full 3D scene, including regions never observed by LiDAR, such as building interiors and heavily occluded surfaces. FreeOcc instead addresses *sensor-grounded occupancy forecasting*: predicting the 3D structure that the sensor will observe at future timesteps.

This distinction matters for planning. The relevant target is future sensor-observable geometry, not the interior of occluded structures. Accordingly, our supervision is derived from future LiDAR returns rather than dense completion labels. Training against completion labels would force the model to infer geometry behind walls and inside objects, which is underdetermined and unnecessary for the forecasting setting considered here.

The evaluation objectives also differ. Occ3D rewards filling unobserved regions, whereas our protocol measures binary occupied/free IoU against future scans and therefore rewards accurate prediction of observable structure. These objectives need not align: a model tuned for completion may hallucinate geometry that improves completion scores but is undesirable for downstream planning.

C.4 Additional experimental results

C.4.1 Robustness to sensor temporal imperfections

Table 5: Robustness to sensor temporal imperfections on nuScenes. The imperfect setting adds severe timestamp jitter and 30% history-frame dropout. All metrics are voxel-based.

Method	Condition	mIoU ↑	Prec. ↑	Rec. ↑	F1 ↑
4D-Occ Khurana et al. [2023]	Ideal	22.6	39.1	31.8	35.1
	Imperfect	14.2	30.5	21.0	24.9
OccWorld Zheng et al. [2024]	Ideal	28.9	50.5	40.2	44.8
	Imperfect	19.0	38.5	28.5	32.8
ViDAR Yang et al. [2024]	Ideal	20.5	39.5	29.8	34.1
	Imperfect	12.8	28.2	19.5	23.1
FreeOcc-D (ours)	Ideal	28.4	41.3	**49.4**	44.9
	Imperfect	26.0	37.5	**46.5**	41.3
FreeOcc-G (ours)	Ideal	30.3	52.5	42.5	47.4
	Imperfect	28.1	49.8	39.0	43.7
FreeOcc-E (ours)	Ideal	**32.1**	**54.6**	45.5	**49.6**
	Imperfect	**29.8**	**51.5**	42.0	**46.3**

FreeOcc retains the strongest performance under temporal noise. FreeOcc-E keeps 92.8% of its ideal mIoU (29.8/32.1), while OccWorld retains 65.7%, 4D-Occ 62.8%, and ViDAR 62.4%. This is

consistent with the timestamp-jitter and sensor-dropout training described in Section 3.3. Detailed results are shown in Table 6.

C.4.2 Ablation studies

Table 6: Ablation study on nuScenes validation. Metrics are voxel-based. The results demonstrate that explicit geometric decoupling and physics-aware regularization are the primary drivers of forecasting quality.

Variant	mIoU ↑	Precision ↑	F1 ↑
FreeOcc-E (full)	32.1	54.6	49.6
Generation Path Ablations			
w/o Flow Matching (FreeOcc-D)	28.4	41.3	44.9
w/o Decoder (FreeOcc-G)	30.3	52.5	47.4
Max blending	31.2	53.4	48.4
Mean blending	30.5	52.6	47.5
Architecture Ablations			
w/o Ego-Motion Warping	25.8	46.2	40.5
w/o Residual Connection	30.2	52.1	47.2
w/o Delay-Aware Confidence Gate	31.1	53.3	48.3
w/o Sensor Dropout	31.4	53.8	48.8
w/o Timestamp Jitter ($\delta_t = 0$)	31.6	54.0	49.0
Fixed-stride temporal encoding	30.8	52.9	48.0
Physics-Aware Regularization Ablations			
w/o **Structural Fidelity** (Group)	28.1	49.5	44.1
w/o **Physical Priors** (Group)	28.7	50.8	44.8
w/o $\mathcal{L}_{\mathrm{vel}}$ (Smoothness)	30.4	52.3	47.4
w/o $\mathcal{L}_{\mathrm{ray}}$ (Free-space)	31.1	53.2	48.3
w/o $\mathcal{L}_{\mathrm{cons}}$ (Mass conservation)	31.4	53.8	48.8
w/o $\mathcal{L}_{\mathrm{unc}}$ (Calibration)	31.5	53.9	48.9
w/o $\mathcal{L}_{\mathrm{rw}}$ (Range weighting)	31.6	54.1	49.0
Uncertainty Ablations			
w/o cooperative unc.-guided FM	30.6	52.7	47.7
w/o uncertainty-aware blending	31.2	53.4	48.4

The ablations in Table 6 support our core design philosophy. Removing Ego-Motion Warping causes the single largest drop (32.1 $\rightarrow$ 25.8 mIoU), confirming that explicit geometric decoupling is fundamental to forecasting accuracy. Within our loss system, the Structural Fidelity group is essential for grounding the discriminative path, while the Physical Priors (specifically velocity smoothness $\mathcal{L}_{\mathrm{vel}}$) ensure that the generative ODE produces kinematically plausible trajectories. Notably, removing either entire regularization group leads to a performance drop significantly larger than removing any individual term, proving that our physics-aware objectives act as a cohesive system rather than a collection of independent penalties.

D Additional Qualitative Results

This appendix provides additional visualizations supporting the qualitative analysis in Section 4.6.

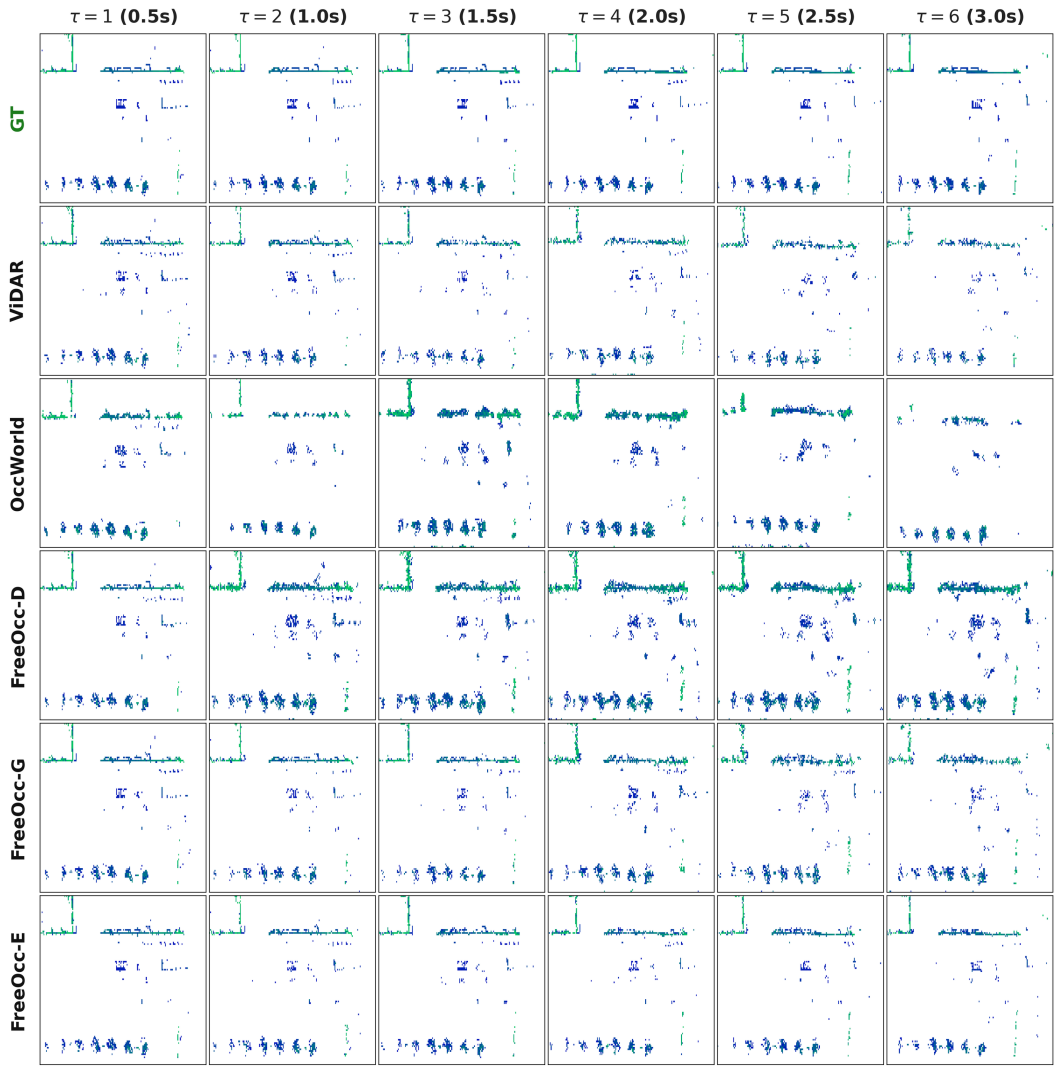

Figure 4: **BEV visualization of 4D occupancy forecasting (additional scene, sample 0970).** Columns span τ=1 to τ=6, comparing FreeOcc against OccWorld and ViDAR on a dense urban scene with multiple interacting vehicles. FreeOcc-E preserves scene layout and object extent more faithfully over the full 3.0 s horizon, whereas OccWorld and ViDAR drift more noticeably after τ=3.

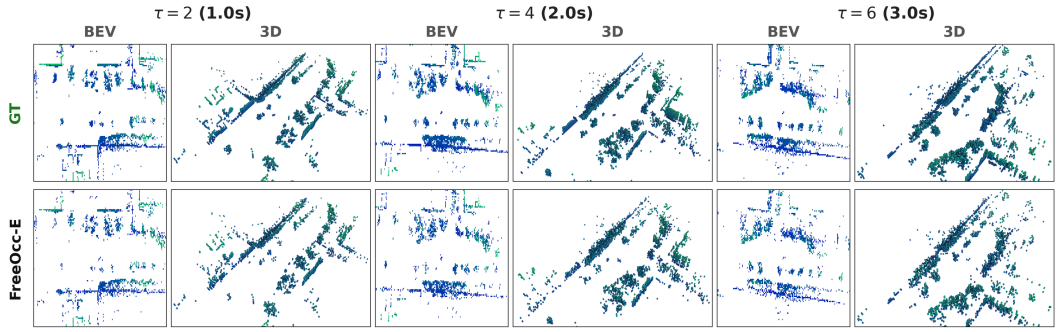

Figure 5: **Temporal evolution of GT vs. FreeOcc-E (nuScenes sample 1730).** Top row: ground-truth occupancy at τ=2 (1.0 s), τ=4 (2.0 s), and τ=6 (3.0 s), each shown as BEV (left) and 3D voxel rendering (right). Bottom row: corresponding FreeOcc-E predictions. Agreement is strongest at short horizons. By τ=6, the main discrepancies appear in the far field and around dynamic actors, while large static structures remain well preserved.

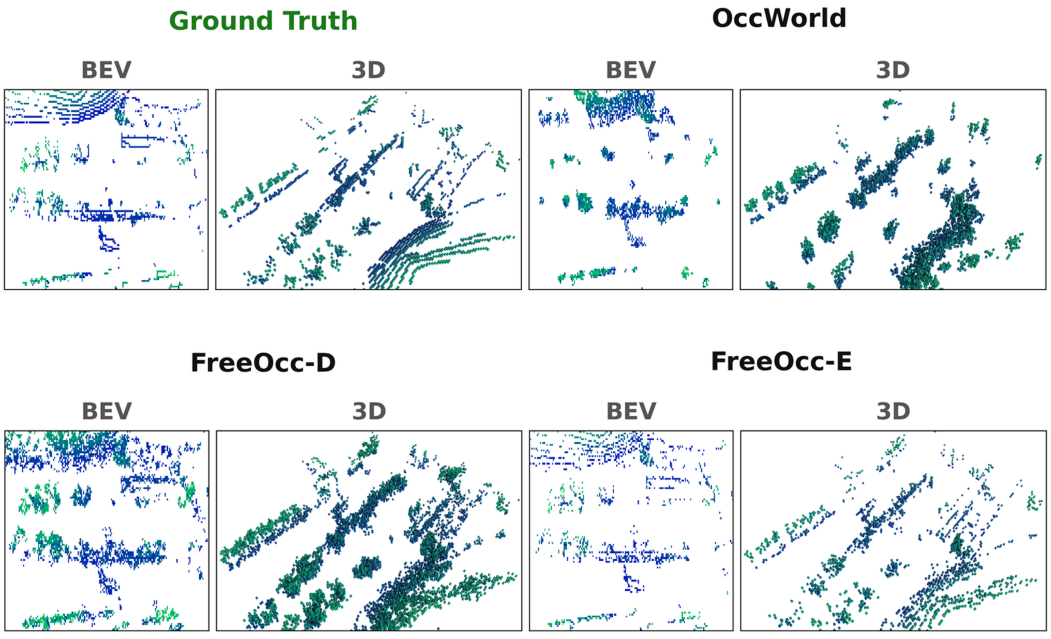

Figure 6: **Long-horizon comparison at τ=6 (3.0 s, nuScenes sample 3380).** Ground Truth, OccWorld, FreeOcc-D, and FreeOcc-E are shown in paired BEV and 3D views. At 3.0 s, OccWorld exhibits boundary blur and noise around moving objects. FreeOcc-D slightly over-predicts occupancy, whereas FreeOcc-E preserves sharper boundaries and cleaner free space.

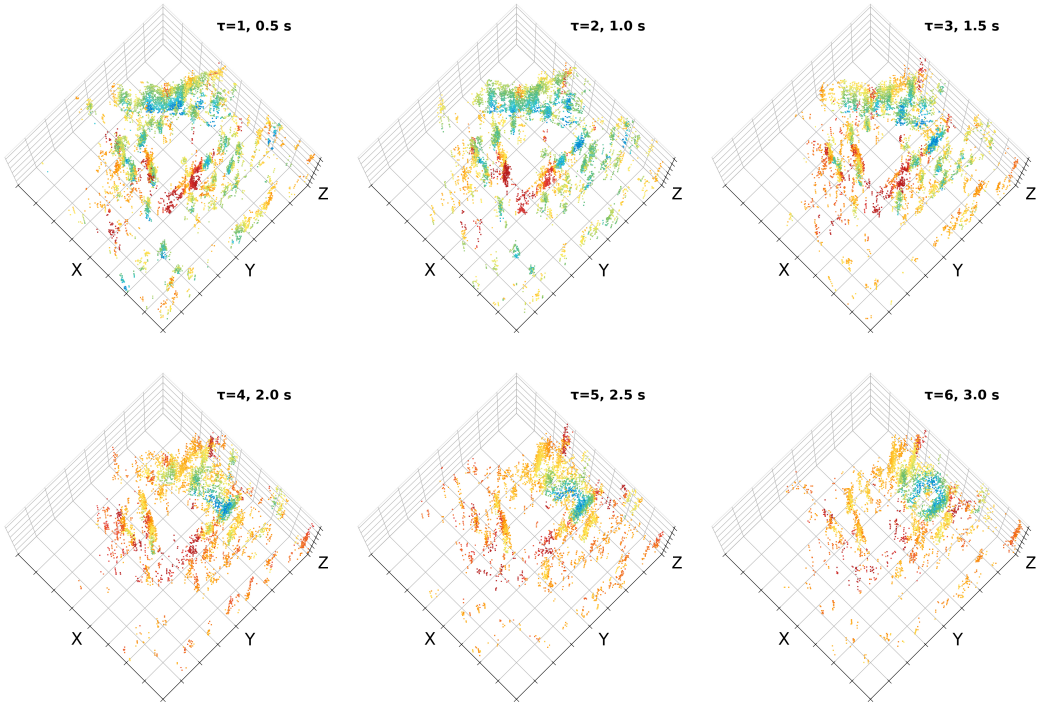

Figure 7: **3D voxel-level uncertainty across prediction horizons (scene-0274, τ=1 to τ=6).** Each panel renders occupied voxels coloured by the decoder's predicted aleatoric uncertainty $\hat{s}_v$; cool tones indicate low uncertainty and warm tones indicate high. At short horizons, uncertainty is concentrated near dynamic actors and occupancy boundaries. As the horizon increases, uncertainty expands into the far field and intensifies around moving objects. By τ=6 (3.0 s), most remaining structures are assigned substantially higher uncertainty. This behaviour is consistent with the ensemble design in Section 3.7, which assigns more weight to the ODE path where decoder confidence weakens.

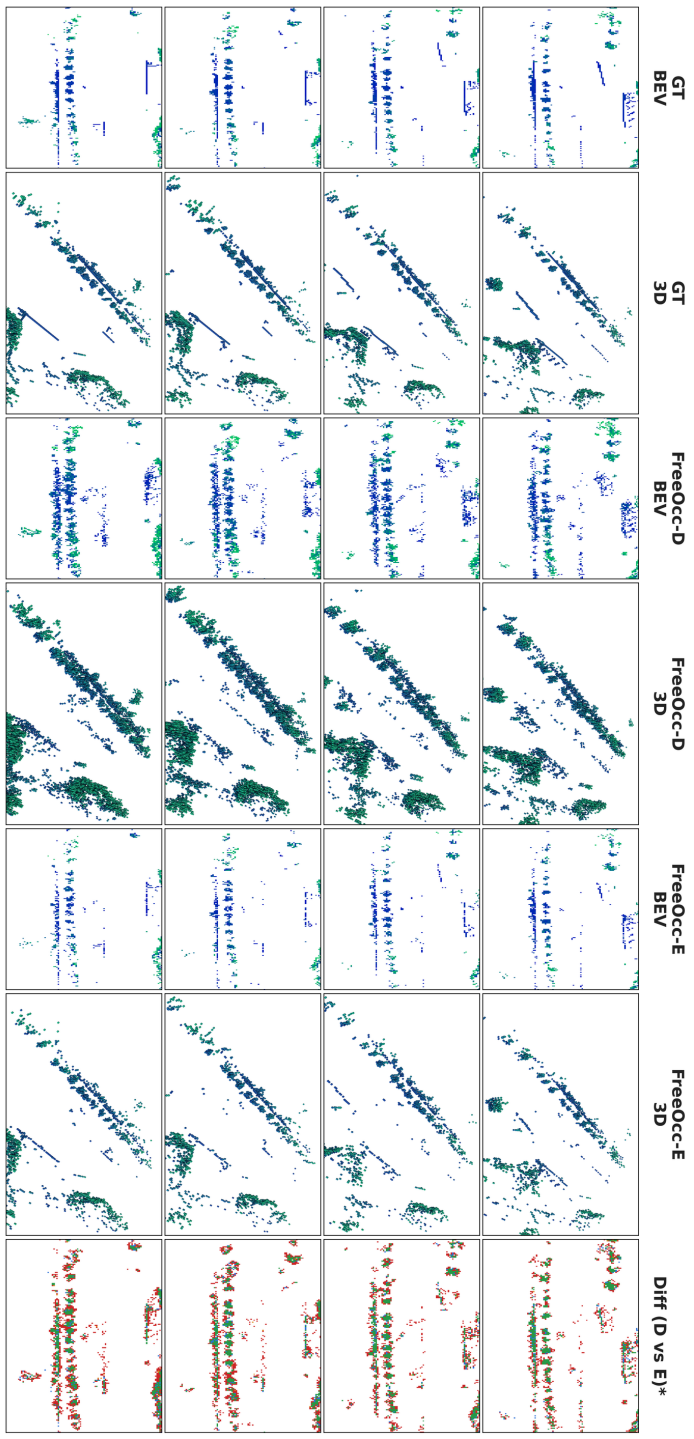

Figure 8: **Noise suppression analysis (nuScenes sample 1290).** Paired BEV and 3D views highlighting hallucinated occupancy at long prediction horizons between FreeOcc-D and FreeOcc-E. FreeOcc-E clears vacated space more effectively while preserving plausible future object positions.

E Uncertainty Visualisation

This appendix visualises the per-voxel aleatoric uncertainty $\hat{s}_v$ predicted by the decoder head, complementing the quantitative ablation of the uncertainty-aware blending mechanism (Section 3.7 and Table 6).

BEV uncertainty evolution. Figure 9 shows BEV projections of predicted occupancy coloured by $\hat{s}_v$ for scene-0220 across all six prediction horizons. Three trends are consistent across the sequence:

- **Near-field confidence.** At τ=1 (0.5 s), uncertainty is concentrated on dynamic objects and occupancy boundaries, while large static structures remain comparatively confident.
- **Horizon-dependent growth.** As τ increases from 1 to 6 (0.5 s to 3.0 s), high-uncertainty regions expand and intensify. By τ=4, a turning vehicle near the centre of scene-0220 dominates the uncertainty map, reflecting the increasing ambiguity of its future trajectory.
- **Far-field diffusion.** Sparse peripheral voxels become uncertain earlier than dense near-field regions, consistent with the weaker LiDAR support available at long range.

These patterns are consistent with the ensemble rule in Section 3.7: short-horizon predictions rely primarily on the decoder, whereas long-horizon and dynamic regions receive greater weight from the ODE branch.

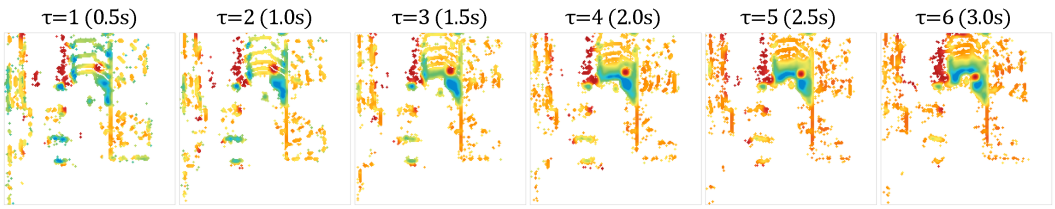

Figure 9: **BEV uncertainty evolution (nuScenes scene-0220, τ=1 to τ=6).** Each panel shows the BEV occupancy coloured by the predicted aleatoric uncertainty $\hat{s}_v$ at horizon τ. Cool colours (cyan/green) indicate low uncertainty; warm colours (orange/red) indicate high uncertainty. Uncertainty increases with horizon and concentrates on dynamic objects and far-field voxels. Similar behaviour appears in Figure 7.

3D voxel-level analysis. Figure 7 renders the same uncertainty signal in 3D for scene-0274, which makes the vertical distribution easier to inspect. At short horizons, the predicted structure is vertically coherent and densely populated; as τ increases, dropout and insertion artefacts become more visible, especially in thin or weakly observed structures. The uncertainty map tracks this degradation closely: voxels that remain at τ=6 are predominantly assigned high uncertainty, indicating that decoder confidence deteriorates in step with forecast difficulty. This behaviour is precisely what the ensemble mechanism exploits when assigning more weight to the ODE path at longer horizons.

F Cross-Domain Qualitative Results on SemanticKITTI

The following figures visualise zero-shot transfer results on SemanticKITTI sequences, using models trained exclusively on nuScenes with no fine-tuning.

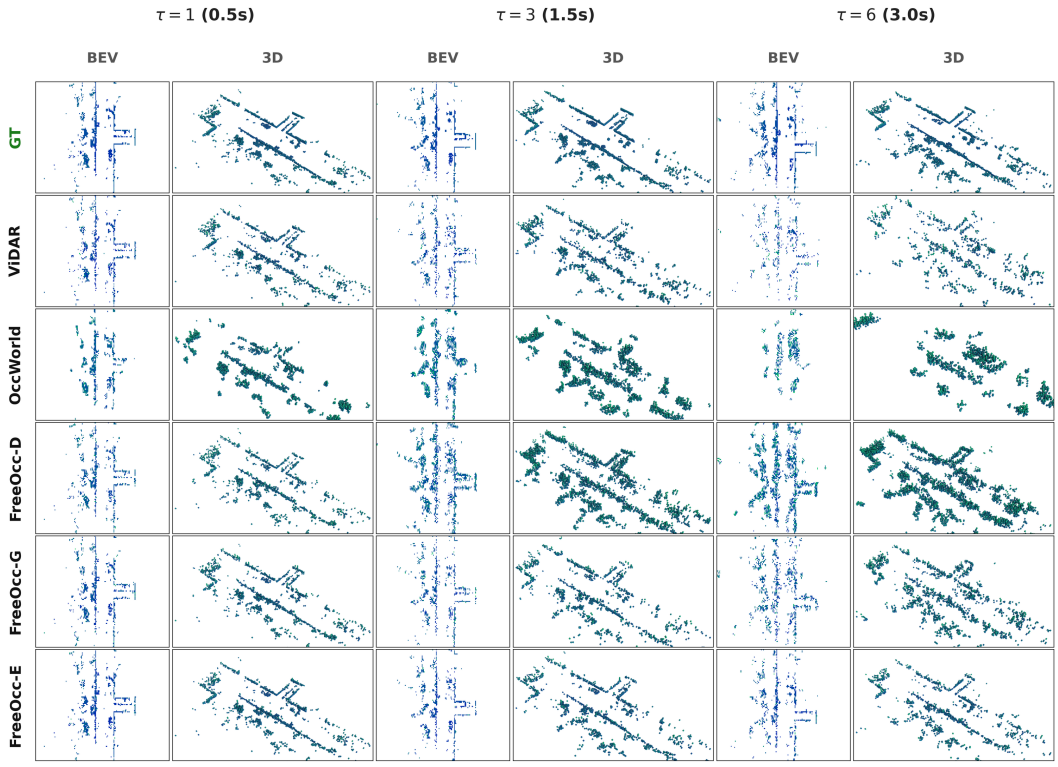

Figure 10: **Zero-shot SemanticKITTI predictions (sample 2070).** BEV and 3D occupancy predictions from a nuScenes-trained model evaluated on SemanticKITTI, comparing FreeOcc-E against OccWorld and ViDAR. Despite domain differences (64-beam LiDAR, different driving context, no fine-tuning), FreeOcc-E preserves road-level structure and major obstacles. The denser 64-beam input provides richer geometric detail than nuScenes's 32-beam data, and FreeOcc adapts without retraining, an advantage of operating on raw 3D geometry rather than appearance-dependent features.

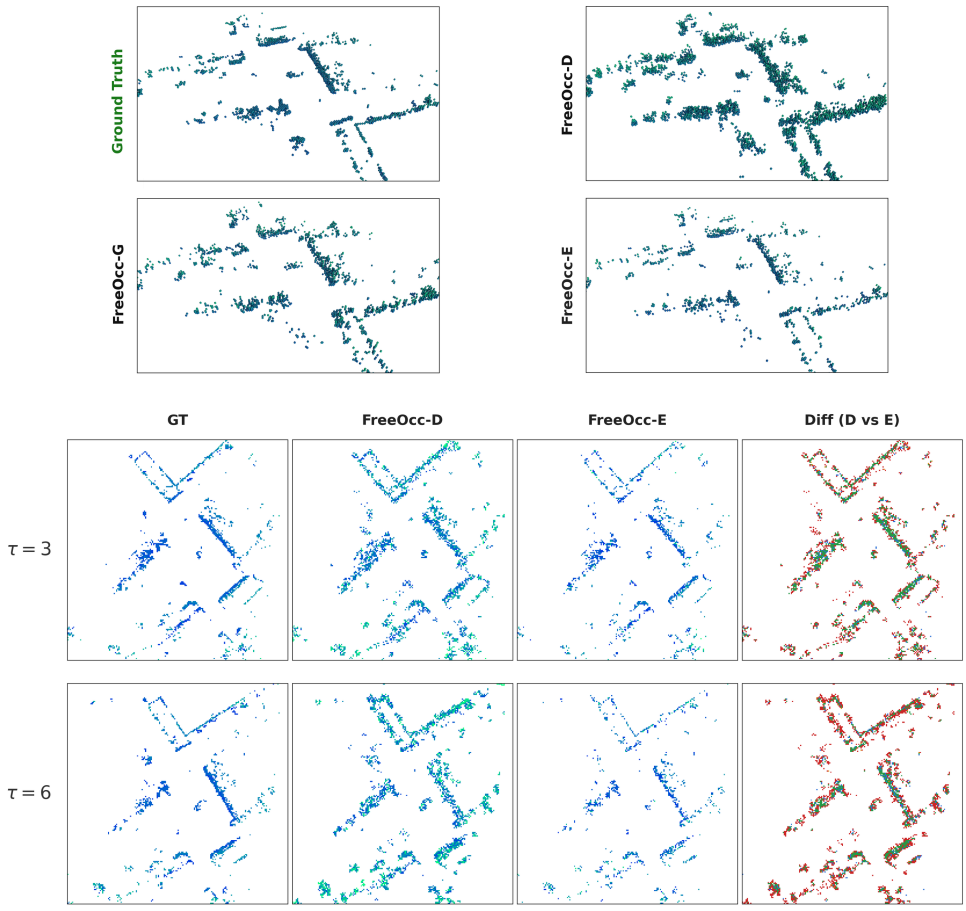

Figure 11: **Additional zero-shot SemanticKITTI scene (sample 1080) and noise suppression analysis.** Top: 3D occupancy predictions at the maximum forecasting horizon ($\tau = 6$, 3.0s), comparing our model variants (FreeOcc-D, -G, -E) against Ground Truth. Bottom: BEV temporal evolution at $\tau = 3$ and $\tau = 6$. Road geometry, vegetation boundaries, and building walls are consistently captured, reinforcing the cross-domain robustness reported in Table 4.

G Robustness Qualitative Results

These visualizations illustrate the effect of degraded sensor conditions (timestamp jitter + 30% frame dropout) on model predictions, corresponding to the quantitative analysis in Table 5.

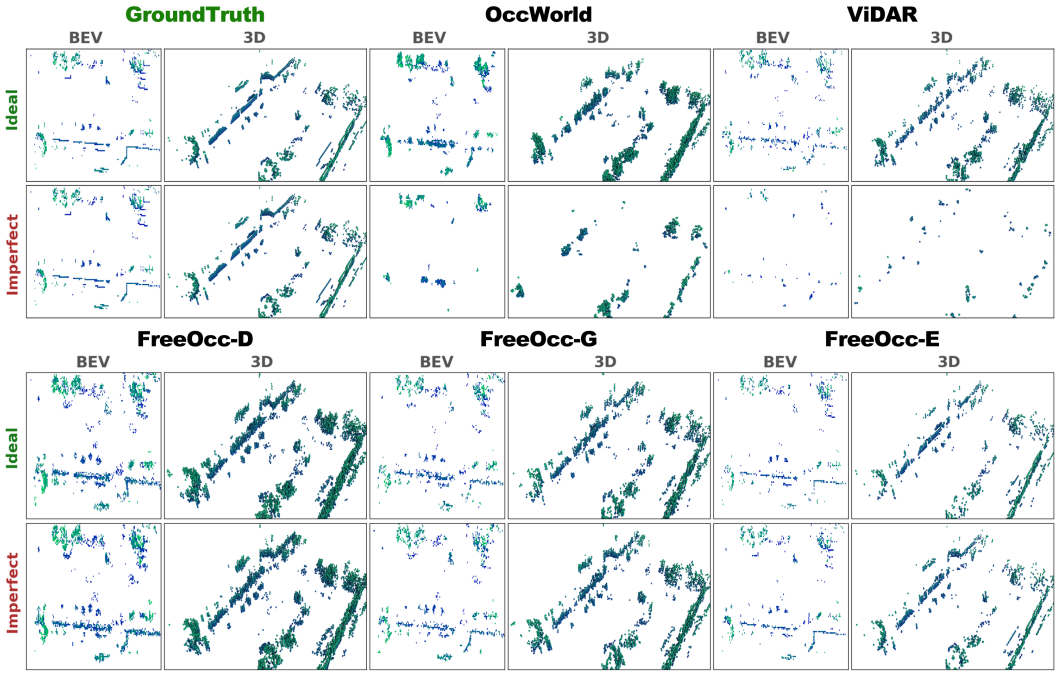

Figure 12: **Robustness on nuScenes ($\tau = 4$, 2.0s, sample 2500).** Occupancy predictions under ideal and imperfect sensor conditions. FreeOcc-E retains its predictions under severe temporal perturbations, while OccWorld and ViDAR break into fragments and hallucinated clusters when temporal input is degraded.

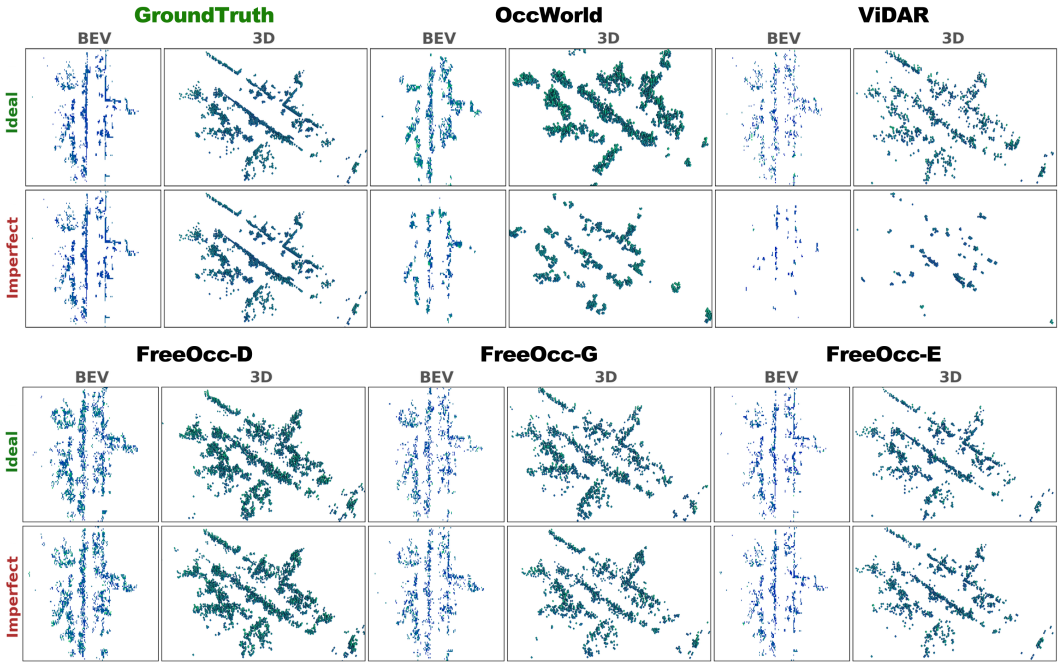

Figure 13: **Cross-domain robustness on SemanticKITTI ($\tau = 4$, 2.0s, sample 2090).** Zero-shot predictions under ideal (top two rows) and imperfect (bottom two rows) sensor conditions. The cross-domain robustness gap compounds both the sensor-domain shift (nuScenes→SemanticKITTI) and the temporal perturbation, yet FreeOcc-E retains recognisable road and object geometry in both cases, while OccWorld and ViDAR suffer severe structural collapse.

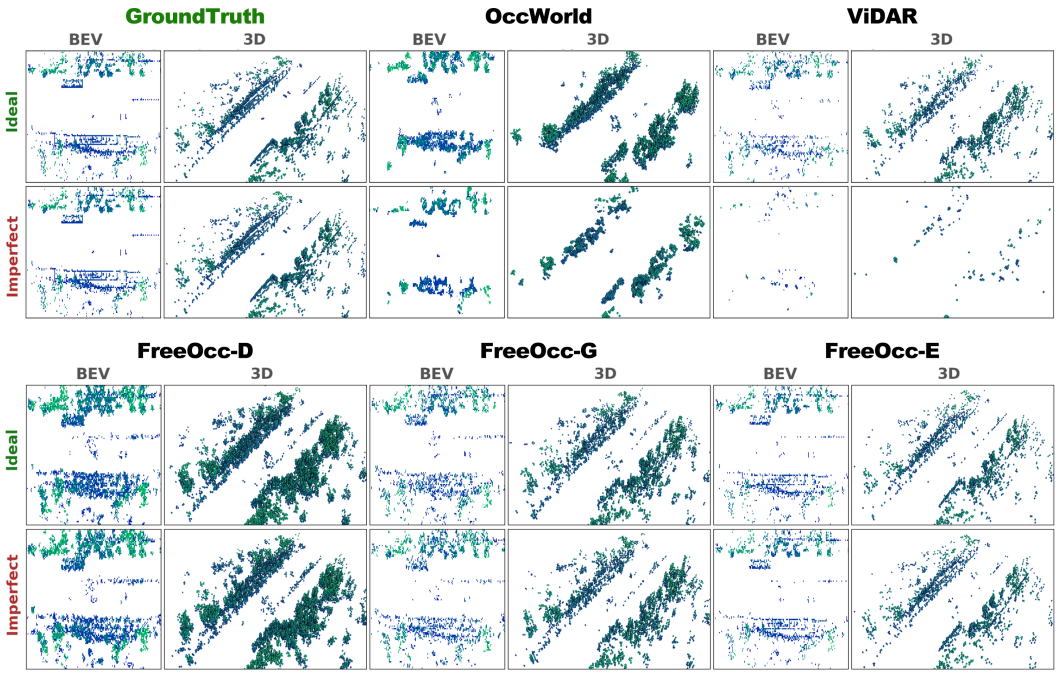

Figure 14: **Robustness on nuScenes at τ=6 (3.0 s, sample 1600).** Long-horizon predictions under ideal (top) and imperfect (bottom) sensor conditions. At the maximum forecasting horizon, OccWorld and ViDAR dissolve into discontiguous fragments, whereas FreeOcc-E preserves the road layout and principal object silhouettes. The gap between ideal and imperfect rows is noticeably smaller for FreeOcc-E than for OccWorld or ViDAR, consistent with the quantitative retention rate in Table 5.

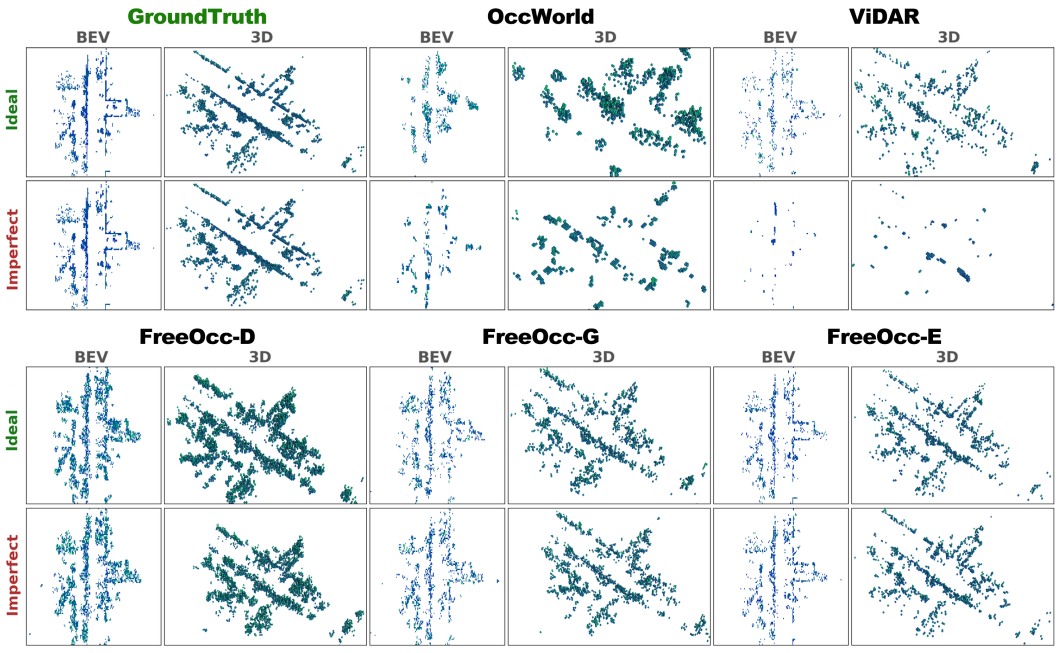

Figure 15: **Cross-domain robustness on SemanticKITTI at τ=6 (3.0 s, sample 2090).** The most challenging evaluation setting: simultaneous cross-domain transfer, long-horizon prediction, and degraded sensor input. FreeOcc-E still recovers the overall road layout and major object boundaries, whereas OccWorld and ViDAR suffer near-complete structural collapse.

H Latency Under Adverse Conditions

A practical question is whether latency remains stable on visually and geometrically difficult scenes, such as rain, reflections, and dense intersections. Here we evaluate runtime directly on a challenging nuScenes rain sequence and report the observed spread.

Evaluation protocol. Figure 16 shows a representative keyframe from nuScenes `scene-0441` (rain, cross-intersection, lane change), one of the most demanding sequences in the dataset. We measure per-frame latency over all 40 keyframes and report the mean and worst-case runtime for FreeOcc-D and FreeOcc-E in Table 7. OccWorld is included as an ideal-condition reference from our benchmark run.

Observed latency stability. FreeOcc-E maintains a mean latency of 37.3 ms (26.8 FPS), with worst-case latency both at 39.1 ms. This spread is small and remains within roughly 2 ms of the dry-weather average reported in the main paper. FreeOcc-D is similarly stable at 22.1 ms (45.2 FPS). Even in this adverse sequence, FreeOcc-E remains faster than our OccWorld reference measured under ideal conditions (56 ms, 17.8 FPS).

Interpretation. These results suggest that, in our implementation, adverse weather and dense traffic do not materially bottleneck runtime. In FreeOcc-E, the fast discriminative decoder first processes the scene and generates the uncertainty map, which then selectively routes only the highly ambiguous regions to the continuous ODE solver. The stability shown in Table 7 indicates that even in challenging conditions like rain, the decoder remains confident enough to resolve the vast majority of the environment (e.g., road layout and free space). Consequently, the additional computational overhead from routing more voxels to the ODE solver is tightly bounded, resulting in a worst-case latency penalty of only about 2 ms. We report these empirical statistics to demonstrate that our uncertainty-aware routing is robust and avoids unbounded computational spikes in unpredictable environments.

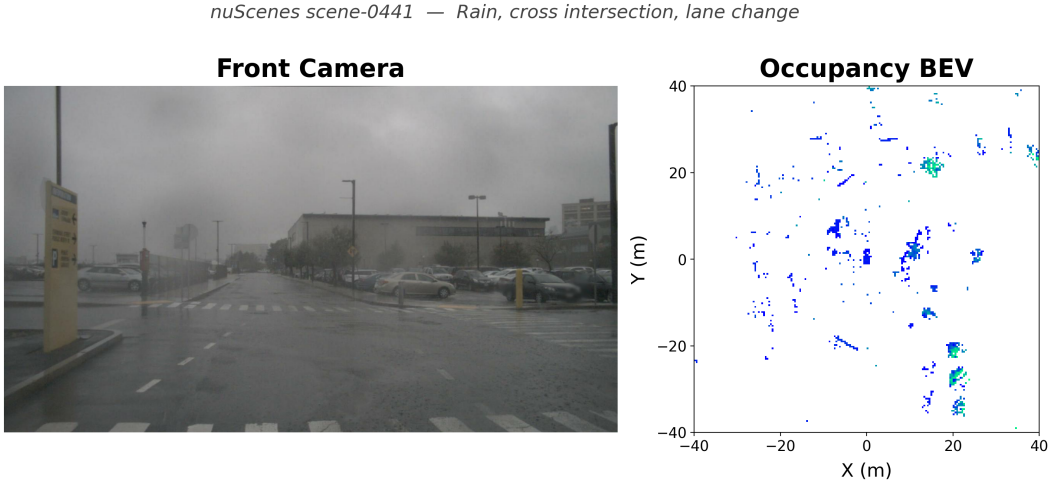

Figure 16: **Rain scene overview (nuScenes scene-0441).** Left: front camera image showing wet road, reflections, and rain droplets. Right: LiDAR BEV coloured by height; the intersection geometry and parked vehicles are clearly visible despite rain-induced noise in the point cloud.

Table 7: **Inference latency on rain sequence** (nuScenes scene-0441, 40 keyframes). Both FreeOcc variants maintain stable latency indistinguishable from dry-weather conditions. FreeOcc-E's worst-case is still 1.4× faster than our OccWorld run under ideal conditions.

Method	Mean (ms)	Worst (ms)	FPS (mean)
OccWorld (ideal, ref.)	56.0	—	17.8
FreeOcc-D (rain)	22.1	23.8	45.2
FreeOcc-E (rain)	37.3	39.1	26.8

I Trajectory Robustness Analysis

To verify that FreeOcc's advantage does not arise from privileged access to the exact future route, we conduct two analyses.

Sensitivity to trajectory error. We inject Gaussian noise into the future ego-poses, reporting the accumulated L2 position error at 3 s.

Table 8: FreeOcc-E mIoU under injected future-trajectory error on nuScenes.

Traj. error @3 s (L2)	0.0 m	0.3 m	0.5 m	1.0 m	2.0 m	4.0 m
FreeOcc-E mIoU	32.1	31.8	31.5	31.1	30.2	28.6

FreeOcc-E degrades gracefully. Across the typical planner-error range ($<$2 m at 3 s), it stays at 30.2 mIoU or above, well above OccWorld's 28.9.

Self-predicted trajectory. We replace the recorded route with a constant-velocity / constant-yaw-rate extrapolation from the recent history, using no future information.

Table 9: FreeOcc-E with a self-predicted trajectory (no future information).

Trajectory	mIoU	NFCD@3s
Recorded route (paper setting)	32.1	0.47
Self-predicted from history	31.2	0.51
OccWorld (no external traj.)	28.9	—

Even without access to the future route, FreeOcc-E reaches 31.2 mIoU, still above OccWorld.

J Information-Matched Warp Baselines

To bound the contribution of the trajectory warp, we build two baselines on FreeOcc's own temporally aggregated present state (not on OccWorld). Warp+Fuse warps this state to the future ego frame and refines it with a light fusion head. Warp+Fuse+FOV additionally masks the warped result to the future sensor field of view. Neither uses the flow-matching branch or the uncertainty ensemble.

Table 10: Information-matched warp baselines versus FreeOcc on nuScenes.

Method	mIoU	Prec.	Rec.	F1	NFCD@3s
Static Copy	15.9	34.0	23.5	27.8	1.11
Warp+Fuse	20.4	35.3	33.1	34.2	1.04
Warp+Fuse+FOV	20.8	36.5	33.4	34.9	0.98
FreeOcc-D	28.4	41.3	49.4	44.9	0.63
FreeOcc-E	32.1	54.6	45.5	49.6	0.47

Both warp baselines trail FreeOcc-E by at least 11.3 mIoU, attributing the gap to the flow-matching branch and the uncertainty ensemble rather than the trajectory warp.

K Static vs. Dynamic Breakdown

We define a ground-truth occupied voxel as *dynamic* if it falls inside the BEV footprint of a nuScenes 3D box of a movable class (vehicle, human, cycle) moving faster than 0.5 m/s, and *static* otherwise. According to our statistics on the nuScenes dataset, dynamic voxels constitute only about 4% of occupied voxels.

FreeOcc and OccWorld are essentially tied on static voxels (37.5 vs 37.2), so the trajectory warp gives no meaningful static advantage. The entire FreeOcc gain is on dynamic voxels ($+5.1$), where the ego-warp offers no help.

Table 11: Static and dynamic mIoU on nuScenes.

Method	Static mIoU	Dynamic mIoU	Overall	Dyn. margin vs OccWorld
Warp+Fuse+FOV	26.8	4.2	20.8	-10.3
OccWorld	37.2	14.5	28.9	—
FreeOcc-E	37.5	19.6	32.1	$+5.1$

L Comparison with SparseWorld-TC

SparseWorld-TC Du et al. [2026] is a camera-based method designed for Occ3D semantic completion. We run the released model and evaluate it at the $\{1, 2, 3\}$ s horizons common to both methods, using our sensor-grounded metrics. We also include FreeOcc-Cam, a camera variant using a Lift-Splat-Shoot (LSS) encoder Philion and Fidler [2020] with our forecasting modules unchanged.

Table 12: Comparison with SparseWorld-TC on our metrics ($\{1, 2, 3\}$ s horizons).

Method	Modality	mIoU	NFCD@3s	FPS
SparseWorld-TC	Camera	23.7	0.82	9.8
FreeOcc-Cam (ours)	Camera	25.6	0.69	19.4
FreeOcc-E (ours)	LiDAR	30.2	0.47	27.0

FreeOcc-E leads on all metrics. The same-modality FreeOcc-Cam also surpasses SparseWorld-TC while running about twice as fast.

M Camera Modality Ablation

To disentangle the LiDAR modality from the forecasting modules, we build FreeOcc-Cam by swapping only the LiDAR encoder for an LSS Philion and Fidler [2020] image-to-BEV encoder. All other components are unchanged. As a control, the camera baseline shares the same LSS encoder but replaces our forecasting stack with a plain feed-forward occupancy head (no ego-motion warp, temporal aggregation, flow matching, or ensemble). Both are trained on the same data.

Table 13: Isolating the modality from the forecasting modules (6-horizon average on nuScenes).

Configuration	Modality	mIoU
FreeOcc (full)	LiDAR	32.1
FreeOcc-Cam	Camera	26.5
Camera baseline (no warp/CFM/ensemble)	Camera	21.9

LiDAR exceeds camera by 5.6 mIoU, and FreeOcc-Cam exceeds the camera baseline by 4.6 mIoU. The forecasting modules thus add value independent of modality.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We explicitly state our performance metrics (e.g., 0.47 m NFCD, 32.1% mIoU), inference speeds (27 FPS), parameter counts, and structural contributions in the Abstract and Section 4.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have dedicated Appendix B to discussing the limitations of our approach.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: Our work is an empirical and architectural contribution for 4D occupancy forecasting rather than a theoretical paper introducing new mathematical proofs.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide comprehensive details on network architectures, augmentation strategies, loss formulations, and hyperparameter settings in Section 4.1 and Appendix A and Appendix C to ensure independent reproducibility.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The mandatory institutional and IP clearance process with our external research partner ultimately prohibited the public release of the source code due to proprietary restrictions. To ensure full methodological reproducibility without the raw code, we have provided comprehensive architectural specifications, loss definitions, and training hyperparameters in Section 4.1, Appendix A, and Appendix C.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 4.1 and Appendix C comprehensively specify dataset splits, voxel specifications, evaluation protocols, and training hyperparameters (optimizer, learning rate, batch size).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Due to the high computational cost of training 4D occupancy forecasting world models, we report performance from single evaluation runs, aligning with the standard evaluation practices of our baselines (e.g., OccWorld, ViDAR).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We explicitly document the hardware used in Section 4.1 and provide detailed inference latency and FPS profiling in Section 4.3 and Appendix H.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our research adheres strictly to the NeurIPS Code of Ethics, focusing on foundational improvements to autonomous driving perception without introducing specific ethical risks.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [N/A]

Justification: Our work focuses on foundational 3D perception and forecasting for autonomous driving. While improvements in this domain generally enhance vehicle safety, the foundational nature of this architecture does not introduce new, direct negative societal impacts.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: Our model predicts 3D occupancy from LiDAR for autonomous driving research and does not generate high-risk content that requires specific deployment safeguards.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We properly cite the nuScenes and SemanticKITTI datasets used in our experiments, which are standard open-access benchmarks in the autonomous driving community.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: Our paper introduces a new model architecture but does not release new datasets or re-packaged assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The research relies entirely on publicly available, pre-collected driving datasets and does not involve human subjects or crowdsourcing.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The research does not involve human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: No LLMs were used as an important, original, or non-standard component of the core methods in this research.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.