Geometry-Aware Spatio-Temporal Context Modeling for 4D Occupancy Forecasting

Sitao Chen[1], Zhuangwei Zhuang[1], Hui Luo[2⋆], Qingyao Wu[1], and Mingkui Tan[1⋆]

[1] South China University of Technology, Guangzhou, China
[2] Institute of Optics and Electronics, CAS, Chengdu, China
chensitao27@gmail.com, mingkuitan@scut.edu.cn

Abstract. 4D occupancy forecasting models the spatio-temporal evolution of 3D scenes and is crucial for autonomous driving, especially for corner-case simulation. Existing methods often rely on discrete tokenization followed by autoregressive prediction, yet struggle with geometric distortion in static structures and inconsistent temporal coherence over the forecasting horizon. In this work, we propose a Geometry-Aware Spatio-Temporal context modeling method (GAST) for 4D occupancy forecasting, built upon progressive explicit-implicit generation and dual-path spatio-temporal modeling. Specifically, the generation module produces per-frame occupancy with high geometric fidelity and semantic plausibility through pose-driven warping, motion-aware feature modulation, and attention-based feature refinement. Subsequently, the spatio-temporal module enhances spatial consistency through global context aggregation while capturing scene evolution through temporal dynamics extraction. This unified design enables joint optimization of historical reconstruction and future forecasting in an end-to-end manner. Extensive experiments on Occ3D-nuScenes demonstrate the superiority of our method, outperforming the state-of-the-art by 7.67% in mIoU and 6.44% in IoU with a 2.84× speedup, while maintaining strong performance in long-term forecasting. Our source code is publicly available at https://github.com/chenst27/GAST.

Keywords: 4D Occupancy Forecasting · Scene Understanding · Autonomous Driving

1 Introduction

3D occupancy represents driving scenes through dense voxel grids with geometric occupancy and semantic labels, offering a holistic and fine-grained 3D representation [1, 31, 33, 47]. Extending this into the temporal domain, 4D occupancy forecasting [7, 16, 23, 45] predicts future 3D scene states from historical observations to model spatio-temporal evolution, supporting critical downstream applications, such as long-horizon motion planning [13, 46], simulation of corner-case

⋆ Corresponding authors.

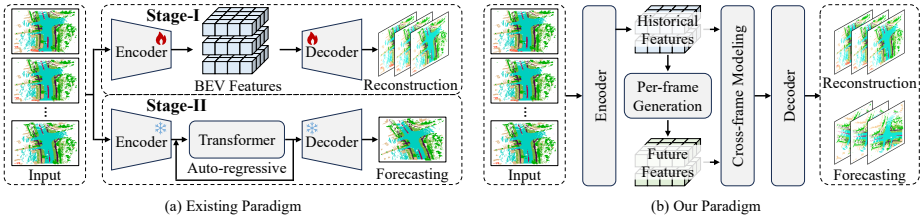

Fig. 1: Comparison with the existing paradigm [16, 45]. It adopts a two-stage pipeline that encodes occupancy into discrete tokens and autoregressively generates future tokens. In contrast, our end-to-end framework directly models spatio-temporal representations in continuous BEV space via per-frame generation and cross-frame modeling.

scenarios [2, 29], and synthesis of high-fidelity data [35, 39]. As a result, 4D occupancy forecasting has emerged as a fundamental task for modern autonomous driving systems.

Driven by advances in generative AI [3], autoregressive-based occupancy world models [16, 32, 34, 45] have become the dominant paradigm for 4D occupancy forecasting (See Figure 1). These methods typically compress 3D semantic occupancy into discrete tokens via VQ-VAE [27], then autoregressively generate future tokens for decoding into occupancy grids. However, they suffer from three major limitations. First, the lack of explicit geometric constraints on static scene structures leads to shape distortion or positional drift (*e.g.*, roads, buildings) during long-term prediction. Second, the causal, step-by-step nature of autoregressive generation hinders the modeling of global spatial context across timestamps, compromising spatio-temporal consistency. Third, the two-stage training pipeline (*i.e.*, separating tokenization from sequence modeling) leaves the predictor inherently constrained by the tokenizer's representational capacity. Addressing these limitations requires a 4D occupancy forecasting framework that jointly ensures geometric consistency of static structures, plausible semantic evolution of dynamic scenes, and cross-frame coherence over the prediction horizon.

In this work, we propose a Geometry-Aware Spatio-Temporal context modeling framework (GAST) that integrates progressive explicit-implicit generation and dual-path spatio-temporal modeling. Given historical occupancy and ego-poses, we first encode compact bird's-eye-view (BEV) features and predict future ego-poses. The generation module then propagates the current BEV structure into future frames via pose-driven rigid transformation, thereby enforcing geometric consistency for static elements. The transformed futures are refined through two implicit stages: motion-aware feature modulation injects sequential ego-dynamics via adaptive affine transformations, and attention-based feature refinement selectively fuses high-fidelity cues from the current observation to enable plausible dynamic evolution. Building on these per-frame representations, the dual-path spatio-temporal module processes spatial layout and temporal dynamics in parallel: a global spatial context aggregator unifies historical and future features in a shared world coordinate system to capture long-range struc-

tural dependencies, while a temporal dynamics extractor regularizes the semantic evolution through sequential encoding. Their outputs are fused to enforce cross-frame coherence in long-horizon predictions. Finally, our GAST is trained end-to-end by jointly optimizing historical reconstruction and future forecasting, which eliminates complex two-stage pipelines and mitigates the error accumulation in autoregressive generation. Our contributions are summarized as follows:

- We propose a geometry-aware spatio-temporal context modeling method (GAST) for 4D occupancy forecasting that constructs per-frame future representations through progressive explicit-implicit generation and enforces cross-frame consistency via dual-path spatio-temporal modeling, achieving geometric fidelity, semantic plausibility, and temporal coherence.
- We design a progressive explicit-implicit generation strategy that explicitly propagates current scene geometry to future frames via pose-driven warping and implicitly refines future representations through motion-aware modulation and attention-based refinement, enhancing both geometric consistency and plausible scene evolution.
- We introduce a dual-path spatio-temporal modeling module that processes spatial layout and temporal evolution in parallel through a global spatial aggregator and a sequential temporal extractor, then fuses their outputs to ensure long-horizon coherence. Extensive experimental results on Occ3D-nuScenes demonstrate the state-of-the-art performance of our method.

2 Related Work

3D occupancy prediction. Semantic occupancy prediction estimates both the occupancy state and semantic label of every voxel in 3D space, offering a more comprehensive and expressive scene representation than object detection [14, 41] or LiDAR segmentation [25, 37]. Existing methods can be mainly categorized into camera-based, LiDAR-based, and multi-sensor fusion methods. Camera-based methods [8, 12, 18, 42] exploit rich appearance cues from images to infer 3D occupancy. CGFormer [42] introduces the context- and geometry-aware voxel Transformer to lift 2D features into 3D volumes. VisHall3D [18] proposes VisFrontierNet to trace the visible frontier and OcclusionMAE to hallucinate plausible geometries for occluded regions. LiDAR-based methods [19,30,36,40] leverage precise spatial measurements of point clouds for volumetric reconstruction. SSC-RS [19] captures multi-level semantic context and multi-scale geometric structures, fusing them through an adaptive representation module. VPNet [30] models semantic uncertainty by introducing confident voxels to represent diverse occupancy possibilities. Multi-sensor methods [6, 20, 28, 47] further exploit complementary strengths from the two modalities. Co-Occ [20] employs implicit volume rendering to fuse explicit LiDAR-camera features. OccGen [28] conditions occupancy generation on multi-modal features and progressively refines predictions. Despite their effectiveness in static perception, these methods fail to model scene dynamics and thus cannot capture temporal evolution.

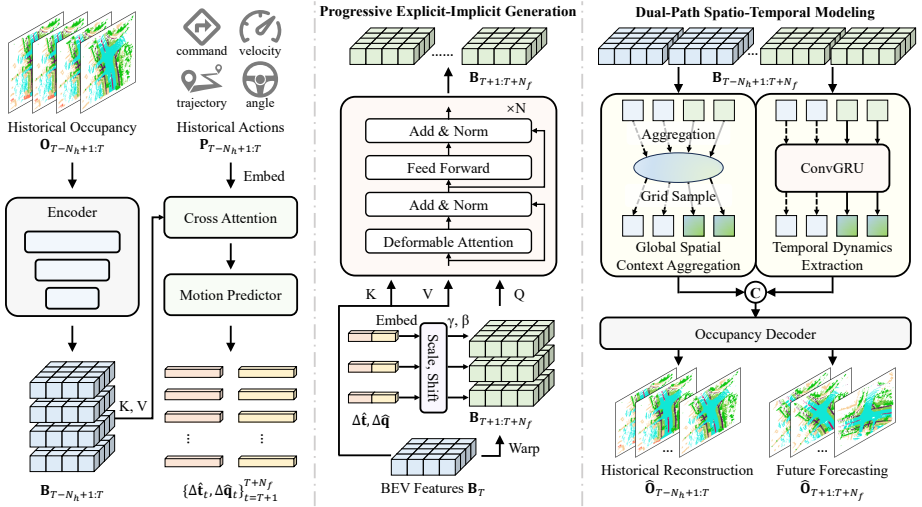

Fig. 2: General architecture of GAST. We first encode historical occupancy into BEV features and predict future ego-poses. The generation module outputs per-frame future features via geometric warping, feature modulation, and deformable cross-attention. The spatio-temporal module enhances global consistency through spatial context aggregation while modeling scene evolution via temporal extraction. Finally, an occupancy decoder jointly reconstructs past observations and forecasts future occupancy.

4D occupancy forecasting. 4D occupancy forecasting aims to predict the future spatio-temporal evolution of 3D scenes in terms of both geometry and semantics. OccWorld [45] pioneers the occupancy world model paradigm that first compresses 3D scenes into discrete tokens via a VQ-VAE tokenizer and then autoregressively generates future tokens with a generative transformer. RenderWorld [38] proposes the Air-Masked VAE that decouples the encoding of air and non-air voxels to preserve the structural sparsity of 3D scenes. Occ-LLM [34] introduces the Motion-Separation VAE that distinguishes between movable and immovable scene elements, and leverages a large language model for occupancy generation. OccLLaMA [32] unifies scene, language, and action tokens into a single vocabulary for joint occupancy-language-action generation. OccTENS [9] decomposes 4D occupancy forecasting into spatial scale-by-scale generation and temporal scene-by-scene prediction. OccSora [29] adopts a 4D scene tokenizer to directly extract spatio-temporal representations from 4D occupancy and generates scenes using a diffusion transformer [22]. DynamicCity [2] further employs HexPlane [5] as a compact 4D representation for both tokenization and generation, enabling efficient modeling of long-horizon scene dynamics. I^2-World [16] decouples scene tokenization into intra-scene and inter-scene components, where the former captures fine-grained static details and the latter models dynamic scene evolution. DOME [7] introduces a diffusion-based world model that compresses occupancy into a continuous latent space and generates future occupancy

conditioned on historical observations. COME [23] further incorporates ControlNet [44] to convert diverse scene conditions into control features that guide the occupancy generation process. However, these methods lack explicit structural priors and cross-frame geometric alignment. In this paper, we propose a geometry-aware spatio-temporal modeling method that constructs future occupancy through progressive explicit-implicit generation and ensures long-horizon coherence via dual-path spatio-temporal modeling.

3 Proposed Method

3.1 Overall Architecture

As illustrated in Figure 2, our GAST consists of an occupancy encoder (Section 3.2), a progressive explicit-implicit generation module (Section 3.3), a dual-path spatio-temporal modeling module (Section 3.4) and an occupancy decoder (Section 3.5). Given historical occupancy and ego-motion information, the encoder extracts compact BEV features and predicts future ego-poses. The generation module produces future BEV features through explicit geometric warping, motion-aware implicit modulation, and attention-based feature refinement. To ensure cross-frame coherence, the spatio-temporal module processes historical and initial future features in parallel: global spatial context aggregation unifies multi-frame features in a shared coordinate system while temporal dynamics extraction models scene evolution via ConvGRU. Their outputs are fused and fed into the decoder, which jointly optimizes historical reconstruction and future forecasting to yield the final 4D occupancy output.

3.2 Occupancy Encoder

Given N_h historical frames, we denote the occupancy grids and ego-poses as $\{\mathbf{O}_t\}_{t=T-N_h+1}^{T}$ and $\{\mathbf{P}_t\}_{t=T-N_h+1}^{T}$, where $\mathbf{P}_t = (\mathbf{t}_t, \mathbf{q}_t)$ with translation $\mathbf{t}_t \in \mathbb{R}^3$ and unit quaternion $\mathbf{q}_t \in \mathbb{R}^4$. Following [45], we compress $\mathbf{O}_t \in \mathbb{R}^{H\times W\times D}$ into a compact BEV representation $\mathbf{B}_t \in \mathbb{R}^{c\times h\times w}$, forming $\{\mathbf{B}_t\}_{t=T-N_h+1}^{T}$. To model ego-motion, we compute relative motions between consecutive frames [16]. For $t \in [T-N_h+1, T-1]$, the translation delta $\Delta\mathbf{t}_t \in \mathbb{R}^3$ and relative rotation $\Delta\mathbf{q}_t \in \mathbb{R}^4$ are embedded with other motion information into an ego-motion token via a lightweight pose encoder. This token is fused with BEV features through cross-attention to predict the future N_f relative pose changes $\{\Delta\hat{\mathbf{t}}_t, \Delta\hat{\mathbf{q}}_t\}_{t=T+1}^{T+N_f}$, which are integrated to obtain absolute future poses $\{\hat{\mathbf{P}}_t\}_{t=T+1}^{T+N_f}$. For fair comparison with methods [7, 16, 23] assuming known future ego-motion, our GAST can also use ground-truth poses, demonstrating remarkable flexibility.

3.3 Progressive Explicit-Implicit Generation

In this section, we present our progressive explicit-implicit generation (PEIG) module that produces initial per-frame future features by combining explicit

geometric propagation with implicit semantic refinement, which preserves geometric fidelity of static structures and captures plausible semantic variations of dynamic scenes.

Explicit Geometric Transformation. Given that static scene structures undergo deterministic geometric transformations under ego-motion, we design the explicit geometric transformation (EGT) that leverages ego-poses as strong geometric priors to perform rigid spatial warping on the current BEV features, yielding geometrically consistent initial states for future occupancy. Specifically, given the current BEV feature $\mathbf{B}_T$ and the future absolute pose sequence $\{\hat{\mathbf{P}}_{T+1}, ..., \hat{\mathbf{P}}_{T+N_f}\}$, for each future time step $t \in [T+1, T+N_f]$, we warp $\mathbf{B}_T$ to the future ego-coordinate system via

$$\mathbf{B}_t^{geo} = \text{Warp}(\mathbf{B}_T; \mathbf{P}_T^{-1} \circ \hat{\mathbf{P}}_t), \tag{1}$$

where $\circ$ denotes the pose composition operation, and Warp performs differentiable grid sampling. This parameter-free module provides a geometrically faithful static background for subsequent dynamic modeling.

Implicit Feature Modulation. To model non-rigid scene dynamics, we propose implicit feature modulation (IFM) that adaptively refines the warped features conditioned on ego-motion priors. Specifically, we leverage the future relative pose changes $\{\Delta\hat{\mathbf{t}}_t, \Delta\hat{\mathbf{q}}_t\}_{t=T+1}^{T+N_f}$ as prior information for dynamic modeling. For each future time step $t \in [T+1, T+N_f]$, we concatenate the translation changes $\Delta\hat{\mathbf{t}}_t \in \mathbb{R}^3$ and rotation changes $\Delta\hat{\mathbf{q}}_t \in \mathbb{R}^4$ along the feature dimension, forming complete relative pose change vectors. We encode all temporal relative pose changes through a multi-layer perceptron (MLP), and generate scaling parameters $\gamma_t \in \mathbb{R}^c$ and offset parameters $\beta_t \in \mathbb{R}^c$ via two separate linear layers. For each future time step t, we apply a channel-wise affine transformation to the explicitly transformed feature $\mathbf{B}_t^{geo}$:

$$\mathbf{B}_t^{mod} = \gamma_t \odot \mathbf{B}_t^{geo} + \beta_t, \tag{2}$$

where $\odot$ denotes channel-wise multiplication. This module adaptively modulates features using relative pose changes, enhancing dynamic scene representation.

Feature Refinement. To refine local spatial details and dynamic scene coherence, we propose a feature refinement (FR) module that employs deformable cross-attention to ground future predictions in high-fidelity context from the current observation. Specifically, given the feature map F, each query q will aggregate the sampled features around the corresponding 2D reference point p through

$$\text{DeformAttn}(q, p, F) = \sum_{i=1}^{N_{head}} W_i \sum_{j=1}^{N_{point}} A_{ij} W_i^{'} F(p + \triangle p_{ij}), \tag{3}$$

where N_{head} and N_{point} denote the number of attention heads and sampled points, respectively. $A_{ij} \in [0, 1]$ and $\triangle p_{ij} \in \mathbb{R}^2$ are attention weight and sampling offset for the j^{th} point in the i^{th} head. $F(p + \triangle p_{ij})$ represents the feature of the

sampled point $p + \triangle p_{ij}$ obtained by bilinear interpolation. For each future time step $t \in [T+1, T+N_f]$, we denote the real-world coordinates of the BEV grid centers at t as $\mathbf{C}_t \in \mathbb{R}^{h \times w \times 2}$. Leveraging the relative ego pose between t and the current time step T, we apply a coordinate transformation $\mathcal{T}_{t \to T}$ to project $\mathbf{C}_t$ into the current ego-centric coordinate system, yielding 2D reference points $\mathcal{T}_{t \to T}(\mathbf{C}_t) \in \mathbb{R}^{h \times w \times 2}$. We treat the modulated future BEV feature $\mathbf{B}_t^{mod}$ as queries, and the current-time BEV feature $\mathbf{B}_T$ as keys and values. The refined features are computed as

$$\mathbf{B}_t^{ref} = \text{DeformAttn}(\mathbf{B}_t^{mod}, \mathcal{T}_{t \to T}(\mathbf{C}_t), \mathbf{B}_T). \tag{4}$$

The resulting features are then passed through a feed-forward network. This mechanism allows each future BEV location to adaptively attend to semantically and spatially relevant regions in the current scene.

3.4 Dual-Path Spatio-Temporal Modeling

Given per-frame future features, we propose a dual-path spatio-temporal modeling (DPSTM) module for cross-frame modeling via (1) global spatial context aggregation that unifies multi-frame features in a shared coordinate system for geometric consistency, and (2) temporal dynamics extraction that models semantic evolution through sequential processing for temporal coherence.

Global Spatial Context Aggregation. To capture long-range spatial dependencies and enrich scene-level semantic understanding, we design the global spatial context aggregation (GSCA) that leverages both historical and future ego-poses to unify multi-frame BEV features into a shared global coordinate system. Specifically, we transform the historical features $\{\mathbf{B}_t\}_{t=T-N_h+1}^{T}$ and the refined future features $\{\mathbf{B}_t^{ref}\}_{t=T+1}^{T+N_f}$ from their ego-centric frames into a fixed global coordinate system using their corresponding poses $\{\mathbf{P}_t\}_{t=T-N_h+1}^{T}$ and $\{\hat{\mathbf{P}}_t\}_{t=T+1}^{T+N_f}$, yielding a temporally extended sequence of spatially aligned feature maps $\{\mathbf{B}_t^{align}\}_{t=T-N_h+1}^{T+N_f}$. To capture multi-granularity context, we project the aligned feature maps onto BEV grids at K decreasing resolutions $\{r_k\}_{k=1}^{K}$, where r_1 is the coarsest and r_k the finest, forming a multi-scale BEV representation. At each scale k, a lightweight convolutional block aggregates local spatial information, producing scale-specific global features $\mathbf{F}_k^{global}$. Low-resolution features ($k = 1$) encode large-scale scene topology, while high-resolution features ($k = K$) preserve fine-grained details. We then fuse these features hierarchically in a coarse-to-fine manner. Starting with $\tilde{\mathbf{F}}_1 = \mathbf{F}_1^{global}$, we iteratively upsample, concatenate, and refine via a residual convolution block:

$$\tilde{\mathbf{F}}_k = \text{Conv}(\text{Upsample}(\tilde{\mathbf{F}}_{k-1}) \oplus \mathbf{F}_k^{global}), \quad k = 2, \ldots, K, \tag{5}$$

where $\oplus$ denotes channel-wise concatenation. The final full-resolution global contextual feature is defined as $\mathbf{F}^{global} := \tilde{\mathbf{F}}_K$. We map it back to the ego-centric BEV grid of each future time step $t \in [T+1, T+N_f]$ via differentiable grid sampling:

$$\mathbf{B}_t^{spatial} = \text{GridSample}(\mathbf{F}^{global},\ \phi(\hat{\mathbf{P}}_t)), \tag{6}$$

where $\phi(\hat{\mathbf{P}}_t)$ denotes the sampling grid obtained by transforming the ego-centric BEV coordinates at time t into the global frame using the predicted pose $\hat{\mathbf{P}}_t$, followed by normalization to $[-1, 1]$. This process yields globally consistent, context-aware future BEV features $\{\mathbf{B}_t^{spatial}\}_{t=T+1}^{T+N_f}$.

Temporal Dynamics Extraction. To model the dynamic evolution of scene semantics, we introduce the temporal dynamics extraction (TDE) that propagates contextual information across historical and future frames. Specifically, we concatenate the historical BEV features $\{\mathbf{B}_t\}_{t=T-N_h+1}^{T}$ and the refined future features $\{\mathbf{B}_t^{ref}\}_{t=T+1}^{T+N_f}$ along the time dimension to form a continuous spatio-temporal sequence $\mathcal{B}_{T-N_h+1:T+N_f} = [\mathbf{B}_{T-N_h+1}, \ldots, \mathbf{B}_T, \mathbf{B}_{T+1}^{ref}, \ldots, \mathbf{B}_{T+N_f}^{ref}]$. We then process $\mathcal{B}$ in chronological order using a lightweight ConvGRU encoder. Denoting the hidden state at time step t as h_t, the recurrence is defined as:

$$\mathbf{h}_t = \text{ConvGRU}(\mathcal{B}_t, \mathbf{h}_{t-1}), \quad t = T - N_h + 1, \ldots, T + N_f, \tag{7}$$

with $\mathbf{h}_{T-N_h} = \mathbf{0}$. The hidden state $\mathbf{h}_t$ serves as a temporally aware representation, encoding the contextual evolution from the start of the sequence up to time t. For future steps ($t > T$), it propagates historical dynamics forward in a causal manner. Finally, we project each hidden state to the BEV feature space to obtain dynamic-aware representations $\{\mathbf{B}_t^{temp}\}_{t=T+1}^{T+N_f}$. This module enhances the temporal coherence and dynamic plausibility of future predictions.

Global spatial context aggregation and temporal dynamics extraction serve as complementary branches: the former enforces cross-frame geometric consistency while the latter ensures temporal smoothness. Their outputs are fused at each future time step $t \in [T + 1, T + N_f]$ via channel-wise concatenation followed by a convolutional block to produce the final future BEV features:

$$\hat{\mathbf{B}}_t = \text{Conv}(\mathbf{B}_t^{spatial} \oplus \mathbf{B}_t^{temp}), \tag{8}$$

where $\oplus$ denotes channel-wise concatenation. We further enhance future BEV features with deformable self-attention to exploit intra-frame semantics.

3.5 Occupancy Decoder

To facilitate spatio-temporal representation learning, our decoder jointly optimizes two objectives: reconstructing historical occupancy from past BEV features and predicting future occupancy from enhanced future features. Specifically, we construct a unified spatio-temporal feature volume by concatenating historical BEV features $\{\mathbf{B}_t\}_{t=T-N_h+1}^{T}$ and enhanced future features $\{\hat{\mathbf{B}}_t\}_{t=T+1}^{T+N_f}$ along the time dimension. This volume is upsampled to the original resolution via bilinear interpolation and subsequently processed by residual convolution blocks to predict 4D occupancy $\hat{\mathbf{O}}_{T-N_h+1:T+N_f} \in \mathbb{R}^{(N_h+N_f)\times S\times H\times W\times D}$, with S semantic classes. The historical segment ($t \leq T$) serves as a reconstruction target, while the future segment ($t > T$) provides the final predictions. This design encourages consistent, high-fidelity representations across both observed and unobserved time steps, bridging perception and prediction in a unified framework.

3.6 Optimization and Training Details

We train our GAST in an end-to-end manner with an objective that combines occupancy prediction loss and ego-pose regression loss. For 4D occupancy, we supervise both historical reconstruction ($t \leq T$) and future forecasting ($t > T$) against ground-truth labels $\mathbf{O}_{T-N_h+1:T+N_f} \in \mathbb{R}^{(N_h+N_f)\times H\times W\times D}$, using a weighted cross-entropy loss and a Lovász-Softmax loss for each segment. The total occupancy loss is the sum of the historical and future components, each weighted by the same hyperparameters λ_{wce} and λ_{lov}. For ego-motion, the encoder predicts future relative pose changes (translation deltas and rotation quaternions). The pose loss combines an L2 loss for translation and a quaternion-based angular loss, with weights $\lambda_{\text{trans}} = 0.01$ and $\lambda_{\text{rot}} = 1.0$ following [16]. The overall objective is the sum of the occupancy and pose losses.

4 Experiments

4.1 Experimental Setup

Dataset. Following [7, 16, 23, 45], we evaluate our method on the widely used Occ3D-nuScenes [26], a semantic occupancy prediction benchmark derived from the large-scale nuScenes [4]. Occ3D-nuScenes provides semantic annotations across 18 categories, comprising 17 semantic classes and 1 empty class. The spatial coverage of the dataset ranges from $[-40m, -40m, -1m]$ to $[40m, 40m, 5.4m]$ along the X, Y, and Z axes, respectively. Given a voxel size of $0.4m \times 0.4m \times 0.4m$, this volume is discretized into a $200 \times 200 \times 16$ voxel grid. This dataset consists of 700 training scenes and 150 validation scenes. Moreover, we perform zero-shot evaluation on the Occ3D-Waymo [26] validation set derived from the Waymo Open Dataset [24], which comprises 202 validation scenes and has the same spatial range and voxel resolution as Occ3D-nuScenes.

Evaluation metrics. We adopt the intersection over union (IoU) as the geometric metric, distinguishing between occupied and empty voxels. Meanwhile, we report the mean intersection over union (mIoU) across all semantic classes to evaluate the quality of semantic segmentation on occupied voxels. Following the evaluation protocol of [16, 45], we present results for each future timestamp and the average performance across all timestamps on the validation set.

Implementation details. We implement our method using PyTorch [21]. Following prior work [7,16,23,45], we leverage 2 seconds of historical data ($N_h = 4$) to predict the future occupancy for the next 3 seconds ($N_f = 6$). We employ one deformable cross-attention layer (with 4 attention heads and 8 sampled points) for feature refinement and two ConvGRU layers for temporal dynamics extraction. The number of BEV resolutions K for global spatial context aggregation is set to 3, corresponding to spatial grids of 50×50, 100×100, and 200×200. The feature dimension is set to 128 empirically. We train the model for 30 epochs using the AdamW optimizer [17] with a weight decay of 0.01. The initial learning rate is set to 0.001 and decays to zero following a cosine annealing schedule. All experiments are conducted with a batch size of 8 across 4 NVIDIA RTX 3090 GPUs. The loss weighting coefficients λ_{wce} and λ_{lov} are all set to 1.0.

Table 1: 4D occupancy forecasting performance on Occ3D-nuScenes validation set. The settings vary by input modality and whether the ego trajectory is predicted (Pred.) or provided as ground truth (GT). "Avg." denotes the average performance across 1s, 2s, and 3s horizons. The **bold** numbers indicate the best results.

Method	Input	Ego traj.	mIoU (%) ↑				IoU (%) ↑			
			1s	2s	3s	Avg.	1s	2s	3s	Avg.
OccWorld-D [45]	Camera	Pred.	11.55	8.10	6.22	8.62	18.90	16.26	14.43	16.53
OccWorld-T [45]	Camera	Pred.	4.68	3.36	2.63	3.56	9.32	8.23	7.47	8.34
OccWorld-S [45]	Camera	Pred.	0.28	0.26	0.24	0.26	5.05	5.01	4.95	5.00
OccWorld-F [45]	Camera	Pred.	8.03	6.91	3.54	6.16	23.62	18.13	15.22	18.99
RenderWorld [38]	Camera	Pred.	2.83	2.55	2.37	2.58	14.61	13.61	12.98	13.73
OccLLaMA [32]	Camera	Pred.	10.34	8.66	6.98	8.66	25.81	23.19	19.97	22.99
PreWorld [11]	Camera	Pred.	12.27	9.24	7.15	9.55	23.62	21.62	19.63	21.62
Occ-LLM [34]	Camera	Pred.	11.28	10.21	9.13	10.21	27.11	24.07	20.19	23.79
DFIT-OccWorld [43]	Camera	Pred.	13.38	10.16	7.96	10.50	19.18	16.85	15.02	17.02
GAST-STC (Ours)	Camera	Pred.	**19.32**	**13.96**	**10.62**	**14.84**	**27.97**	**23.35**	**20.22**	**23.87**
DOME-STC [7]	Camera	GT	17.79	14.23	11.58	14.53	26.39	23.20	20.42	23.33
I^2-World-STC [16]	Camera	GT	21.67	18.78	16.47	18.97	30.55	28.76	26.99	28.77
GAST-STC (Ours)	Camera	GT	**22.90**	**19.93**	**17.17**	**20.16**	**31.62**	**29.92**	**28.00**	**29.87**
Copy&Paste [45]	3D-Occ	Pred.	14.91	10.54	8.52	11.33	24.47	19.77	17.31	20.52
OccWorld [45]	3D-Occ	Pred.	25.78	15.14	10.51	17.14	34.63	25.07	20.18	26.63
OccLLaMA [32]	3D-Occ	Pred.	25.05	19.49	15.26	19.93	34.56	28.53	24.41	29.17
RenderWorld [38]	3D-Occ	Pred.	28.69	18.89	14.83	20.80	37.74	28.41	24.08	30.08
Occ-LLM [34]	3D-Occ	Pred.	24.02	21.65	17.29	20.99	36.65	32.14	28.77	32.52
COME [23]	3D-Occ	Pred.	30.57	19.91	13.38	21.29	36.96	28.26	21.86	29.03
DFIT-OccWorld [43]	3D-Occ	Pred.	31.68	21.29	15.18	22.71	40.28	31.24	25.29	32.27
GAST (Ours)	3D-Occ	Pred.	**38.60**	**24.79**	**18.54**	**27.38**	**44.69**	**34.21**	**28.85**	**35.90**
DOME [7]	3D-Occ	GT	35.11	25.89	20.29	27.10	43.99	35.36	29.74	36.36
UniScene [10]	3D-Occ	GT	35.37	29.59	25.08	31.76	38.34	32.70	29.09	34.84
COME [23]	3D-Occ	GT	42.75	32.97	26.98	34.23	50.57	43.47	38.36	44.13
I^2-World [16]	3D-Occ	GT	47.62	38.58	32.98	39.73	54.29	49.43	45.69	49.80
GAST (Ours)	3D-Occ	GT	**55.54**	**46.33**	**40.18**	**47.40**	**60.77**	**55.97**	**51.96**	**56.24**

4.2 Quantitative Comparison

4D occupancy forecasting. We evaluate our method on the Occ3D-nuScenes validation set. Following [7,16,23,45], we assess our GAST under various configurations, determined by two key factors: (1) whether the model input consists of ground-truth 3D occupancy or raw sensor inputs (*e.g.*, camera images), and (2) whether the ego trajectory is ground truth or generated by a planning module. The results are presented in Table 1. When camera images are used as model input, we follow the established approach of I^2-World [16] by employing STCOcc [15] to process the visual data and generate semantic occupancy predictions. For scenarios where the ego trajectory is predicted, we forecast the future ego poses by leveraging historical scene features and ego-motion information. Under this configuration, our GAST achieves the best performance, with an average mIoU of 14.84% and an average IoU of 23.87%. For example, it surpasses DFIT-OccWorld [43] by 4.34% in mIoU and by 6.85% in IoU. When the ground-truth ego trajectory is provided, we perform a comparative

Table 2: Long-term 4D occupancy forecasting performance on Occ3D-nuScenes validation set. Ground-truth 3D occupancy and trajectories are used as inputs. "Avg." denotes the average performance over the 1-8 second prediction horizon. The **bold** numbers indicate the best results.

Method	mIoU (%) ↑								
	1s	2s	3s	4s	5s	6s	7s	8s	Avg.
DOME [7]	30.10	21.35	17.36	14.86	12.61	11.03	10.00	9.34	15.83
COME [23]	33.78	24.57	21.35	18.25	15.84	13.85	12.99	11.96	19.07
GAST (Ours)	**36.23**	**30.61**	**27.13**	**24.36**	**21.98**	**19.81**	**17.83**	**15.96**	**24.12**
Method	IoU (%) ↑								
	1s	2s	3s	4s	5s	6s	7s	8s	Avg.
DOME [7]	39.04	31.20	27.14	24.73	22.32	20.28	19.05	17.97	25.21
COME [23]	44.20	36.25	32.86	30.03	26.93	24.70	23.30	21.44	29.96
GAST (Ours)	**49.09**	**45.11**	**42.11**	**39.52**	**37.19**	**34.99**	**32.87**	**30.87**	**39.06**

evaluation with DOME-STC [7] and I^2-World-STC [16], with all methods using predictions from STCOcc [15]. Specifically, across the 1s, 2s, and 3s prediction horizons, our method consistently delivers the highest accuracy, yielding mIoU of 22.90%, 19.93%, 17.17%, and IoU of 31.62%, 29.92%, 28.00%, respectively. When the model is provided with ground-truth 3D occupancy along with a predicted ego trajectory, our GAST achieves an average mIoU of 27.38% and an average IoU of 35.90%, substantially outperforming all existing approaches. Further improvements are observed when ground-truth 3D occupancy and ground-truth ego trajectory are used simultaneously, elevating performance to 47.40% mIoU and 56.24% IoU, corresponding to gains of 20.02% mIoU and 20.34% IoU, respectively. Compared to state-of-the-art methods, including diffusion-based approaches [7, 10, 23] and autoregressive frameworks [16], our GAST demonstrates consistently stronger predictive accuracy. Specifically, it surpasses I^2-World [16] by a notable margin of 7.67% in mIoU and 6.44% in IoU. These consistent and significant improvements across diverse evaluation settings highlight not only the robustness of our design under practical conditions but also its superior capacity to capture spatio-temporal dependencies in dynamic 3D scenes, thereby validating the effectiveness and advancement of the proposed method.

Long-term forecasting. Long-term occupancy forecasting plays a crucial role in autonomous driving, enabling safer planning and decision-making in dynamic environments. To evaluate the long-term predictive capability of our method, we extend the prediction horizon from 3 seconds to 8 seconds and use both ground-truth 3D occupancy and ground-truth ego trajectory as model input, following the setup of COME [23]. As shown in Table 2, our GAST achieves an average mIoU of 24.12% and an average IoU of 39.06% over the full 8-second horizon, consistently surpassing all compared baselines at every future timestamp. Specifically, it outperforms DOME [7] and COME [23] by 8.29%, 5.05% in mIoU and 13.85%, 9.1% in IoU, respectively. These results thus validate the effectiveness and scalability of our method in long-term forecasting.

Table 3: Zero-shot performance on Occ3D-Waymo validation set.

Method	Rate	mIoU (%)	IoU (%)
Copy&Paste	10Hz	28.34	40.09
I^2-World [16]	10Hz	43.73	60.97
GAST (Ours)	10Hz	**57.47**	**70.66**
Copy&Paste	2Hz	17.17	28.21
I^2-World [16]	2Hz	36.38	52.36
GAST (Ours)	2Hz	**46.70**	**60.09**

Table 4: Comparisons of model efficiency on the Occ3D-nuScenes validation set. The **bold** numbers indicate the best results.

Method	#Params.↓	#FLOPs↓	Latency↓	mIoU↑	IoU↑
DOME [7]	444.07M	2928.66G	1899.15ms	27.10%	36.36%
COME [23]	692.97M	4461.46G	4377.67ms	34.23%	44.13%
I^2-World [16]	22.67M	494.58G	227.09ms	39.73%	49.80%
GAST (Ours)	**12.09M**	**416.27G**	**80.03ms**	**47.40%**	**56.24%**

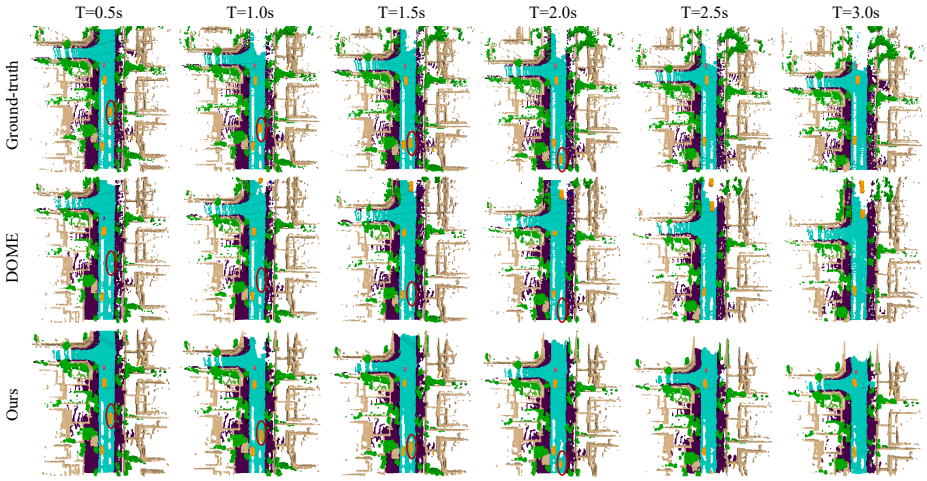

Fig. 3: Qualitative results of GAST on Occ3D-nuScenes.

Zero-shot forecasting. To evaluate the generalization ability of our method, we conduct zero-shot evaluation on the Occ3D-Waymo validation set following I^2-World [16]. As shown in Table 3, our GAST delivers superior performance under different sampling rates. Specifically, at the 10 Hz setting, our GAST surpasses I^2-World [16] by 13.74% in mIoU and 9.69% in IoU. When the sampling rate drops to 2 Hz, the margins are 10.32% and 7.73% in mIoU and IoU, respectively. These improvements demonstrate the strong generalization of our method.

4.3 Qualitative Evaluation

To better understand the benefits of our method, we provide qualitative evaluations on Occ3D-nuScenes. As shown in Figure 3, our GAST achieves superior geometric fidelity and temporal consistency, particularly in preserving fine-grained structural details and capturing dynamic object motions, which demonstrates its effectiveness in modeling both static structure and dynamic motion.

4.4 Efficiency Analysis

We evaluate the efficiency of our method on the Occ3D-nuScenes validation set using a single GeForce RTX 3090. As shown in Table 4, our GAST achieves better

Table 5: Ablation study of the proposed model components on Occ3D-nuScenes validation set. The **bold** numbers denote the best results.

	PEIG			DPSTM		mIoU (%)	IoU (%)
	EGT	IFM	FR	GSCA	TDE		
1						18.31	29.01
2	✓					31.78	41.65
3	✓	✓				32.55	42.63
4	✓	✓	✓			40.47	49.15
5	✓	✓	✓	✓		46.08	55.50
6	✓	✓	✓	✓	✓	**47.40**	**56.24**

Table 6: Effect of the number of historical frames on the Occ3D-nuScenes validation set.

Number	Ego traj.	mIoU (%)	IoU (%)
1	Pred.	17.01	30.46
2	Pred.	24.51	35.11
3	Pred.	26.32	35.50
4	Pred.	**27.38**	**35.90**
1	GT	34.88	51.52
2	GT	46.18	55.91
3	GT	46.97	55.85
4	GT	**47.40**	**56.24**

performance with fewer model parameters, lower FLOPs, and reduced inference latency, demonstrating its effectiveness and efficiency. For instance, our GAST is 2.84× faster than I^2-World [16] while delivering a 7.67% mIoU improvement.

4.5 Ablation Study

Effect of the proposed model components. We study the effect of the proposed model components of our GAST on the Occ3D-nuScenes validation set, including explicit geometric transformation (EGT), implicit feature modulation (IFM), feature refinement (FR) of progressive explicit-implicit generation (PEIG), and global spatial context aggregation (GSCA), temporal dynamics extraction (TDE) of dual-path spatio-temporal modeling (DPSTM). Note that the baseline model uses independent per-frame convolutions to predict future occupancy from current BEV features. The experimental results are shown in Table 5. Comparing the first and second rows, introducing explicit geometric transformation significantly improves the mIoU by 13.47% (31.78% vs. 18.31%) and IoU by 12.64% (41.65% vs. 29.01%), confirming the importance of geometric priors in preserving static structure consistency. Further integrating implicit feature modulation yields additional gains (32.55% mIoU, 42.63% IoU), indicating the benefit of motion-aware feature adaptation. The subsequent feature refinement substantially boosts model performance to 40.47% mIoU and 49.15% IoU, which validates its role in fusing contextual cues for local geometry enhancement. Crucially, the progressive explicit-implicit generation module, through the synergy of rigid propagation and elastic modulation, achieves clear improvements over the baseline in per-frame occupancy forecasting. Moreover, from the fourth and fifth rows, the proposed global spatial context aggregation further contributes 5.61% mIoU (46.08% vs. 40.47%) and 6.35% IoU (55.50% vs. 49.15%), highlighting the significance of modeling long-range spatial dependencies in a unified world coordinate system for coherent scene reconstruction. Finally, when combined with temporal dynamics extraction, the full model achieves the best performance at 47.40% mIoU and 56.24% IoU, verifying the effectiveness of progressive explicit-implicit generation and dual-path spatio-temporal modeling in producing high-fidelity and temporally consistent 4D occupancy prediction.

Table 7: Effect of the number of DCA layers in feature refinement on Occ3D-nuScenes validation set.

Number	1	2	3	4
mIoU (%)	47.40	47.62	**47.71**	47.55
IoU (%)	56.24	**56.98**	56.88	56.68

Table 8: Effect of the number of ConvGRU layers in temporal dynamics extraction on Occ3D-nuScenes validation set.

Number	1	2	3	4
mIoU (%)	47.03	47.40	47.31	**47.45**
IoU (%)	56.04	56.24	56.74	**56.85**

Effect of the number of historical frames. We investigate the impact of the number of historical frames on the Occ3D-nuScenes validation set. As shown in Table 6, increasing the number of historical frames consistently improves prediction accuracy by providing a richer temporal context for modeling scene dynamics. Specifically, when using the predicted trajectory, mIoU rises from 17.01% with one frame to 27.38% with four frames. When the ground-truth trajectory is provided, the model performance further increases to 47.40% mIoU with four frames. These results demonstrate that expanding the historical horizon enables the model to better capture scene evolution and motion patterns, leading to more accurate future occupancy prediction.

Effect of the number of DCA layers in feature refinement. We investigate the impact of the number of deformable cross-attention layers in feature refinement on the Occ3D-nuScenes validation set. As shown in Table 7, deformable cross-attention is essential for adaptively aggregating high-fidelity contextual cues from the current observation, thereby enhancing both the local geometry and semantics of the predicted occupancy. Although stacking more layers yields consistent gains in both mIoU and IoU, we adopt a single layer for feature refinement to balance model performance with computational efficiency.

Effect of the number of ConvGRU layers in temporal dynamics extraction. We investigate the impact of the number of ConvGRU layers in temporal dynamics extraction on the Occ3D-nuScenes validation set. As shown in Table 8, the temporal dynamics extraction module, built upon ConvGRU layers, is crucial for capturing scene evolution by modeling sequential dependencies across frames. Performance improves progressively as the number of ConvGRU layers increases from one to four, with four layers achieving the peak results (47.45% mIoU, 56.85% IoU). Considering the gains beyond two layers are marginally incremental, we use two ConvGRU layers as an optimal trade-off between temporal modeling capacity and inference efficiency.

Effect of BEV resolution in global spatial context aggregation. We study the effect of BEV resolution in global spatial context aggregation on the Occ3D-nuScenes validation set. This module adopts a multi-scale design with three hierarchical BEV resolutions 50×50, 100×100, and 200×200. As shown in Table 9, the multi-scale BEV aggregation achieves the overall best performance (47.40% mIoU, 56.24% IoU), consistently outperforming all single-resolution variants. The results demonstrate that combining features from coarse-level scene layout and fine-grained structural details enables more comprehensive spatial context modeling, thereby improving occupancy prediction accuracy.

Table 9: Effect of BEV resolution in global spatial context aggregation on the Occ3D-nuScenes validation set.

BEV Resolution 50×50	100×100	200×200	mIoU (%)	IoU (%)
✓			46.90	55.63
	✓		47.18	55,64
		✓	47.17	55.73
✓	✓	✓	**47.40**	**56.24**

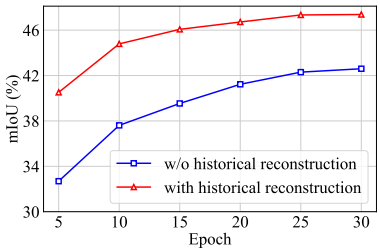

Fig. 4: Validation curves of mIoU with and without historical occupancy reconstruction on the Occ3D-nuScenes.

Effect of historical occupancy reconstruction. We study the effect of historical occupancy reconstruction on the Occ3D-nuScenes validation set. As illustrated in Figure 4, incorporating the reconstruction loss consistently improves mIoU throughout training, indicating effective learning of spatio-temporal dynamics. The results demonstrate that jointly optimizing historical reconstruction and future prediction yields more coherent spatio-temporal representations, enhancing geometric fidelity and temporal consistency in long-horizon forecasting.

5 Conclusion

In this work, we present GAST, a geometry-aware spatio-temporal context modeling method for 4D occupancy forecasting built on progressive explicit-implicit generation and dual-path spatio-temporal modeling. Different from the existing two-stage tokenize-then-generate paradigm, our method unifies historical scene reconstruction and future occupancy forecasting in a continuous BEV space. Specifically, the generation module ensures per-frame geometric consistency and semantic plausibility through pose-driven warping, motion-aware feature modulation, and attention-based feature refinement. The spatio-temporal module further enforces cross-frame coherence via global spatial context aggregation and temporal dynamics extraction, capturing long-range spatial dependencies and smooth scene evolution. The experimental results on the Occ3D-nuScenes demonstrate the superior effectiveness and efficiency of the proposed method.

Acknowledgements

This work was partially supported by the Joint Funds of the National Natural Science Foundation of China (Grant No.U24A20327), Guangdong S&T Program under Grant 2026B0101110001.

References

1. Behley, J., Garbade, M., Milioto, A., Quenzel, J., Behnke, S., Stachniss, C., Gall, J.: Semantickitti: A dataset for semantic scene understanding of lidar sequences. In: ICCV. pp. 9297–9307 (2019)
2. Bian, H., Kong, L., Xie, H., Pan, L., Qiao, Y., Liu, Z.: Dynamiccity: Large-scale 4d occupancy generation from dynamic scenes. arXiv preprint arXiv:2410.18084 (2024)
3. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. NeurIPS **33**, 1877–1901 (2020)
4. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: CVPR. pp. 11621–11631 (2020)
5. Cao, A., Johnson, J.: Hexplane: A fast representation for dynamic scenes. In: CVPR. pp. 130–141 (2023)
6. Duan, Z., Dang, C., Hu, X., An, P., Ding, J., Zhan, J., Xu, Y., Ma, J.: Sdgocc: Semantic and depth-guided bird's-eye view transformation for 3d multimodal occupancy prediction. In: CVPR. pp. 6751–6760 (2025)
7. Gu, S., Yin, W., Jin, B., Guo, X., Wang, J., Li, H., Zhang, Q., Long, X.: Dome: Taming diffusion model into high-fidelity controllable occupancy world model. arXiv preprint arXiv:2410.10429 (2024)
8. Huang, Y., Zheng, W., Zhang, Y., Zhou, J., Lu, J.: Tri-perspective view for vision-based 3d semantic occupancy prediction. In: CVPR. pp. 9223–9232 (2023)
9. Jin, B., Gu, S., Hu, X., Zheng, Y., Guo, X., Zhang, Q., Long, X., Yin, W.: Occtens: 3d occupancy world model via temporal next-scale prediction. arXiv preprint arXiv:2509.03887 (2025)
10. Li, B., Guo, J., Liu, H., Zou, Y., Ding, Y., Chen, X., Zhu, H., Tan, F., Zhang, C., Wang, T., et al.: Uniscene: Unified occupancy-centric driving scene generation. In: CVPR. pp. 11971–11981 (2025)
11. Li, X., Li, P., Zheng, Y., Sun, W., Wang, Y., Chen, Y.: Semi-supervised vision-centric 3d occupancy world model for autonomous driving. arXiv preprint arXiv:2502.07309 (2025)
12. Li, Y., Yu, Z., Choy, C., Xiao, C., Alvarez, J.M., Fidler, S., Feng, C., Anandkumar, A.: Voxformer: Sparse voxel transformer for camera-based 3d semantic scene completion. In: CVPR. pp. 9087–9098 (2023)
13. Li, Y., Fan, L., He, J., Wang, Y., Chen, Y., Zhang, Z., Tan, T.: Enhancing end-to-end autonomous driving with latent world model. arXiv preprint arXiv:2406.08481 (2024)
14. Li, Z., Wang, W., Li, H., Xie, E., Sima, C., Lu, T., Yu, Q., Dai, J.: Bevformer: learning bird's-eye-view representation from lidar-camera via spatiotemporal transformers. IEEE TPAMI **47**(3), 2020–2036 (2024)
15. Liao, Z., Wei, P., Chen, S., Wang, H., Ren, Z.: Stcocc: Sparse spatial-temporal cascade renovation for 3d occupancy and scene flow prediction. In: CVPR. pp. 1516–1526 (2025)
16. Liao, Z., Wei, P., Zhang, R., Chen, S., Wang, H., Ren, Z.: I2-world: Intra-inter tokenization for efficient dynamic 4d scene forecasting. In: ICCV. pp. 25810–25819 (2025)
17. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)

18. Lu, H., Su, Y., Zhang, X., Gao, L., Xue, Y., Wang, L.: Vishall3d: Monocular semantic scene completion from reconstructing the visible regions to hallucinating the invisible regions. In: ICCV. pp. 28674–28684 (2025)
19. Mei, J., Yang, Y., Wang, M., Huang, T., Yang, X., Liu, Y.: Ssc-rs: Elevate lidar semantic scene completion with representation separation and bev fusion. In: IROS. pp. 1–8. IEEE (2023)
20. Pan, J., Wang, Z., Wang, L.: Co-occ: Coupling explicit feature fusion with volume rendering regularization for multi-modal 3d semantic occupancy prediction. IEEE RAL pp. 5687–5694 (2024)
21. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. NeurIPS **32** (2019)
22. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: ICCV. pp. 4195–4205 (2023)
23. Shi, Y., Jiang, K., Meng, Q., Wang, K., Wang, J., Sun, W., Wen, T., Yang, M., Yang, D.: Come: Adding scene-centric forecasting control to occupancy world model. arXiv preprint arXiv:2506.13260 (2025)
24. Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., et al.: Scalability in perception for autonomous driving: Waymo open dataset. In: CVPR. pp. 2446–2454 (2020)
25. Tan, M., Zhuang, Z., Chen, S., Li, R., Jia, K., Wang, Q., Li, Y.: Epmf: Efficient perception-aware multi-sensor fusion for 3d semantic segmentation. IEEE TPAMI **46**(12), 8258–8273 (2024)
26. Tian, X., Jiang, T., Yun, L., Mao, Y., Yang, H., Wang, Y., Wang, Y., Zhao, H.: Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving. NeurIPS **36**, 64318–64330 (2023)
27. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. NeurIPS **30** (2017)
28. Wang, G., Wang, Z., Tang, P., Zheng, J., Ren, X., Feng, B., Ma, C.: Occgen: Generative multi-modal 3d occupancy prediction for autonomous driving. In: ECCV. pp. 95–112. Springer (2024)
29. Wang, L., Zheng, W., Ren, Y., Jiang, H., Cui, Z., Yu, H., Lu, J.: Occsora: 4d occupancy generation models as world simulators for autonomous driving. arXiv preprint arXiv:2405.20337 (2024)
30. Wang, L., Lin, D., Yang, K., Liu, R., Guo, Q., Xie, W., Wang, M., Liang, L., Wang, Y., Li, P.: Voxel proposal network via multi-frame knowledge distillation for semantic scene completion. NeurIPS **37**, 101096–101115 (2024)
31. Wang, X., Zhu, Z., Xu, W., Zhang, Y., Wei, Y., Chi, X., Ye, Y., Du, D., Lu, J., Wang, X.: Openoccupancy: A large scale benchmark for surrounding semantic occupancy perception. In: ICCV. pp. 17850–17859 (2023)
32. Wei, J., Yuan, S., Li, P., Hu, Q., Gan, Z., Ding, W.: Occllama: An occupancy-language-action generative world model for autonomous driving. arXiv preprint arXiv:2409.03272 (2024)
33. Wei, Y., Zhao, L., Zheng, W., Zhu, Z., Zhou, J., Lu, J.: Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving. In: ICCV. pp. 21729–21740 (2023)
34. Xu, T., Lu, H., Yan, X., Cai, Y., Liu, B., Chen, Y.: Occ-llm: Enhancing autonomous driving with occupancy-based large language models. arXiv preprint arXiv:2502.06419 (2025)

35. Yan, T., Wu, D., Han, W., Jiang, J., Zhou, X., Zhan, K., Xu, C.z., Shen, J.: Drivingsphere: Building a high-fidelity 4d world for closed-loop simulation. In: CVPR. pp. 27531–27541 (2025)
36. Yan, X., Gao, J., Li, J., Zhang, R., Li, Z., Huang, R., Cui, S.: Sparse single sweep lidar point cloud segmentation via learning contextual shape priors from scene completion. In: AAAI. vol. 35, pp. 3101–3109 (2021)
37. Yan, X., Gao, J., Zheng, C., Zheng, C., Zhang, R., Cui, S., Li, Z.: 2dpass: 2d priors assisted semantic segmentation on lidar point clouds. In: ECCV. pp. 677–695. Springer (2022)
38. Yan, Z., Dong, W., Shao, Y., Lu, Y., Liu, H., Liu, J., Wang, H., Wang, Z., Wang, Y., Remondino, F., et al.: Renderworld: World model with self-supervised 3d label. In: ICRA. pp. 6063–6070. IEEE (2025)
39. Yang, X., Wen, L., Wei, T., Ma, Y., Mei, J., Li, X., Lei, W., Fu, D., Cai, P., Dou, M., et al.: Drivearena: A closed-loop generative simulation platform for autonomous driving. In: ICCV. pp. 26933–26943 (2025)
40. Yang, X., Zou, H., Kong, X., Huang, T., Liu, Y., Li, W., Wen, F., Zhang, H.: Semantic segmentation-assisted scene completion for lidar point clouds. In: IROS. pp. 3555–3562. IEEE (2021)
41. Yin, T., Zhou, X., Krahenbuhl, P.: Center-based 3d object detection and tracking. In: CVPR. pp. 11784–11793 (2021)
42. Yu, Z., Zhang, R., Ying, J., Yu, J., Hu, X., Luo, L., Cao, S.Y., Shen, H.L.: Context and geometry aware voxel transformer for semantic scene completion. NeurIPS **37**, 1531–1555 (2024)
43. Zhang, H., Xue, Y., Yan, X., Zhang, J., Qiu, W., Bai, D., Liu, B., Cui, S., Li, Z.: An efficient occupancy world model via decoupled dynamic flow and image-assisted training. arXiv preprint arXiv:2412.13772 (2024)
44. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: ICCV. pp. 3836–3847 (2023)
45. Zheng, W., Chen, W., Huang, Y., Zhang, B., Duan, Y., Lu, J.: Occworld: Learning a 3d occupancy world model for autonomous driving. In: ECCV. pp. 55–72. Springer (2024)
46. Zheng, Y., Yang, P., Xing, Z., Zhang, Q., Zheng, Y., Gao, Y., Li, P., Zhang, T., Xia, Z., Jia, P., et al.: World4drive: End-to-end autonomous driving via intention-aware physical latent world model. In: ICCV. pp. 28632–28642 (2025)
47. Zhuang, Z., Wang, Z., Chen, S., Liu, L., Luo, H., Tan, M.: Robust 3d semantic occupancy prediction with calibration-free spatial transformation. arXiv preprint arXiv:2411.12177 (2024)