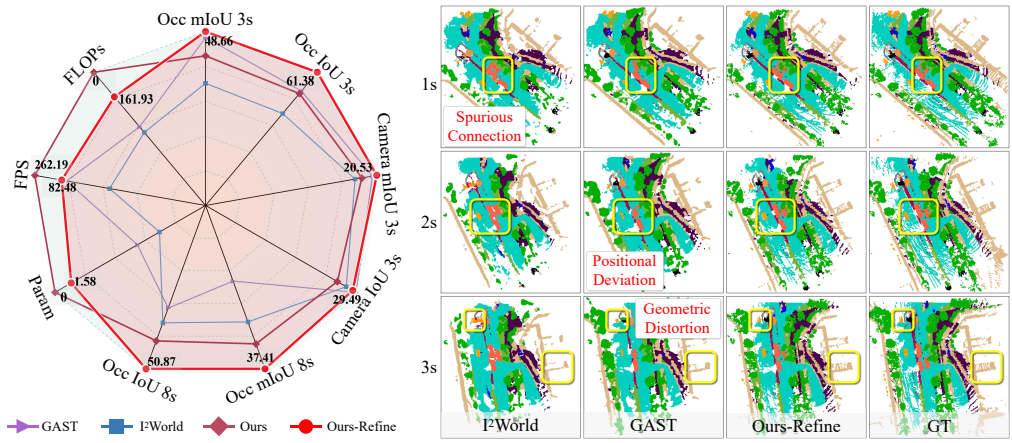
MOTIONWORLD: RETHINKING 4D OCCUPANCY FORECASTING AS STRUCTURED SCENE EVOLUTION

Anonymous authors
Paper under double-blind review



(a) Performance Across Metrics (b) Qualitative results on Occ3D-nuScenes

Figure 1: **Overall comparison on Occ3D-nuScenes.** (a) Quantitative results on forecasting accuracy, efficiency, and model complexity at short- and long-term horizons. (b) Qualitative results at 1s/2s/3s, with highlights marking typical failures of existing methods. Ours-Refine reduces spurious inter-instance connections and positional deviations while preserving fine-grained geometry.

ABSTRACT

We introduce **MotionWorld**, a simple yet effective training-free analytical occupancy world model that decomposes 4D forecasting into static-scene propagation and instance-level motion extrapolation. For the static background, we leverage ego-motion priors to warp historical background and present *Boundary-aware Conservative Completion* (BCC) to conservatively fill warping-induced holes. For dynamic objects, we propose *Trajectory Instance Association* (TIA) to recover a sparse trajectory per instance, and *Self-Validated Motion Extrapolation* (SVME) to select among stationary, constant-velocity, constant-acceleration, exponentially smoothed motion and turning hypotheses via historical replay. The extrapolated instances are then fused with the propagated background to produce future semantic occupancy. We further present **MotionWorld-Refine**, a compact variant that learns targeted corrections using the analytical forecast as a structured prior, improving accuracy while mitigating the structural distortions of unconstrained full-scene generation. On Occ3D-nuScenes, MotionWorld attains highly ***competitive accuracy*** at **265.19** FPS on a dual-core CPU, achieving a **3.5**× speedup over the previous fastest approach running on an RTX 4090 GPU. Under the GT occupancy setting, MotionWorld-Refine establishes a new state of the art, surpassing the previous best method by **1.26%** semantic mIoU and **5.14%** geometric IoU with only **13%** of its parameters. Over an 8-second horizon, it maintains the average gains of **13.29%** and **11.81%** in semantic mIoU and geometric IoU, respectively.

1 INTRODUCTION

4D occupancy world models forecast the voxel-wise geometric and semantic evolution of driving scenes from historical 3D occupancy observations (Park et al., 2026; Chen et al., 2025; Liu et al.,

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

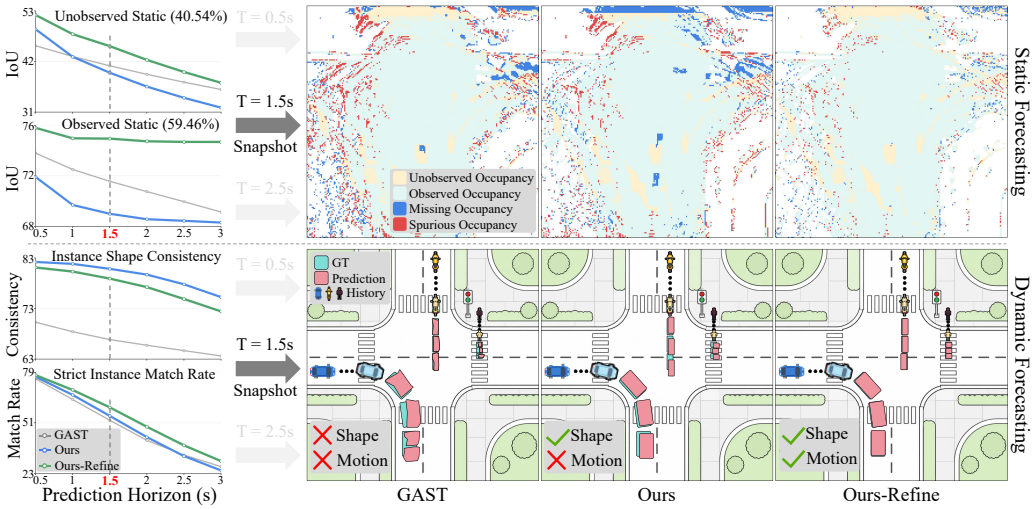


Figure 2: **Static and dynamic forecasting analysis. Top:** Future static occupancy splits into historically observed (59.46%) and unobserved (40.54%) regions. The consistently lower IoU in unobserved regions highlights the difficulty of static completion; the T=1.5 s snapshot visualizes missing and spurious occupancy across methods. **Bottom:** Instance shape consistency and strict match rate degrade with the forecasting horizon. Our training-free core preserves geometry via rigid propagation but is limited by analytical motion extrapolation, whereas Ours-Refine improves motion accuracy while largely retaining shape consistency.

2024). Accurate and temporally consistent forecasts of future occupancy are essential for motion planning (Zheng et al., 2024; Yang et al., 2025; Zhang et al., 2024), out-of-distribution scene generation (Li et al., 2025a; Yang et al., 2026), and closed-loop simulation and evaluation (Yan et al., 2025a; Liu et al., 2026). More fundamentally, by jointly reasoning about the geometric and semantic evolution of dynamic 3D environments, 4D occupancy forecasting constitutes a core problem in predictive scene understanding.

Existing occupancy world models generally adopt autoregressive token prediction (Zheng et al., 2024; Wei et al., 2024; Xu et al., 2025a; Li et al., 2025b) diffusion-based synthesis (Gu et al., 2024; Li et al., 2025a; Shi et al., 2026), or end-to-end spatio-temporal modeling (Liao et al., 2025b; Chen et al., 2026; Yang et al., 2026) to forecast complete future scenes. However, despite their different architectures, these methods typically forecast the entire scene jointly, without explicitly modeling the distinct evolution mechanisms of static regions and dynamic instances. As shown in the static forecasting panel of Fig. 2, future static occupancy comprises historically observed regions (59.46%), which should be preserved, and previously unobserved regions (40.54%), which require completion. Treating both regions through homogeneous scene generation may corrupt reliable observations while failing to recover unseen structures, resulting in missing or spurious occupancy. In the dynamic forecasting panel of Fig. 2, dynamic instances exhibit a different evolution pattern: their positions change over time, whereas their identities and physical shapes should remain consistent. However, scene-level forecasting entangles motion estimation with voxel synthesis, leading to accumulated trajectory errors and non-rigid shape distortions over longer horizons.

Motivated by these observations, we propose **MotionWorld**, a training-free analytical occupancy world model that explicitly decomposes scene evolution into static propagation and instance-level motion extrapolation. For static regions, MotionWorld uses known ego-motion to propagate reliable occupancy and Boundary-aware Conservative Completion (BCC) to conservatively complete warping-induced holes. For dynamic instances, Trajectory Instance Association (TIA) associates each current instance with class-consistent historical observations to recover its sparse trajectory. Self-Validated Motion Extrapolation (SVME) then evaluates a compact set of analytical motion hypotheses through historical replay and selects the hypothesis best supported by each instance's observed trajectory. The selected motion is extrapolated into the future and used to rigidly propagate the complete current instance, thereby decoupling motion estimation from geometry synthesis and preserving instance shape. The propagated static scene and dynamic instances are finally fused

to produce future semantic occupancy. We further introduce **MotionWorld-Refine**, an optional lightweight extension that treats the analytical forecast as a structured prior and learns targeted corrections for unseen static regions and residual motion errors. Our main contributions are as follows:

- We introduce **MotionWorld**, a training-free analytical framework that decomposes 4D occupancy forecasting into ego-motion-aware static propagation and instance motion extrapolation. BCC completes boundary gaps, while TIA and SVME recover trajectories and select instance-specific motion hypotheses through historical replay.
- We further introduce **MotionWorld-Refine**, a lightweight, structure-preserving extension that reformulates learned forecasting from full-scene generation into localized residual correction over the analytical forecast. This formulation improves forecasting accuracy while retaining reliable scene structure and instance geometry.
- Extensive experiments establish complementary advantages: MotionWorld achieves competitive accuracy at **265.19** FPS on CPU, while MotionWorld-Refine sets a new state of the art with only **13%** of GAST's parameters, improving mIoU and IoU by **1.26%** and **5.14%**. Across the 8-second rollout, these gains reach **13.29%** and **11.81%** on average.

2 Related Work

Occupancy world models. Early approaches establish a token-based autoregressive paradigm that compresses occupancy into discrete latents (Kingma & Welling, 2013; Van Den Oord et al., 2017) and predicts future scene tokens (Zheng et al., 2024; Wei et al., 2024). Subsequent work improves the efficiency and temporal expressiveness of this paradigm through structured latent representations and intra–inter scene tokenization (Xu et al., 2025a; Liao et al., 2025b). A complementary line adopts continuous latent spaces and diffusion models to improve generation fidelity, temporal coherence, and controllability (Gu et al., 2024; Li et al., 2025a). More recent structured predictors incorporate decoupled voxel flow or geometry-aware spatio-temporal context to better capture scene evolution (Zhang et al., 2024; Chen et al., 2026; Du et al., 2026). Despite advances in scene representation and geometric propagation, these methods largely rely on learned dynamics rather than explicitly selecting motion states for individual occupancy instances. In contrast, MotionWorld separates static background and dynamic instances directly in semantic occupancy space, propagating the former with ego-motion and extrapolating the latter under instance-specific motion states selected from historical observations.

Motion-aware forecasting. In occupancy space, flow-based methods propagate observed voxels using ego-motion and learned dynamic flow (Zhang et al., 2024), while related approaches jointly predict occupancy and scene flow to improve temporal reasoning (Liao et al., 2025a). At the instance level, trajectory forecasters predict multimodal futures from tracked agent histories and map context (Gao et al., 2020; Shi et al., 2022). Anchor-based methods further score or refine predefined trajectory hypotheses (Chai et al., 2020; Phan-Minh et al., 2020). These methods either learn dense motion in occupancy space or rely on pre-associated tracks and learned trajectory predictors. In contrast, TIA recovers sparse instance histories directly from semantic occupancy, and SVME selects an instance-specific analytical motion hypothesis by testing how well it predicts held-out observations within that history. The selected motion then propagates the complete instance without learning a motion selector or synthesizing its geometry anew.

3 Methodology

3.1 Overview

We introduce **MotionWorld**, a simple yet effective training-free analytical framework that disentangles static-scene persistence from dynamic-instance motion for future semantic occupancy forecasting, as illustrated in Fig. 3. Let T denote the current time, N_h the number of input frames, and N_f the number of future prediction frames. Given an occupancy history $\mathcal{O}_{T-N_h+1:T}$, we extract dynamic-instance voxels $\mathcal{O}^{dyn}_{T-N_h+1:T}$ from each input frame and retain the remaining occupied voxels in the current frame as the static scene $\mathcal{O}^{st}_{T}$. In the dynamic branch, we present *Trajectory Instance Association* (TIA, **§3.2**) to construct instance representations $\mathcal{T}$ from $\mathcal{O}^{dyn}_{T-N_h+1:T}$ using

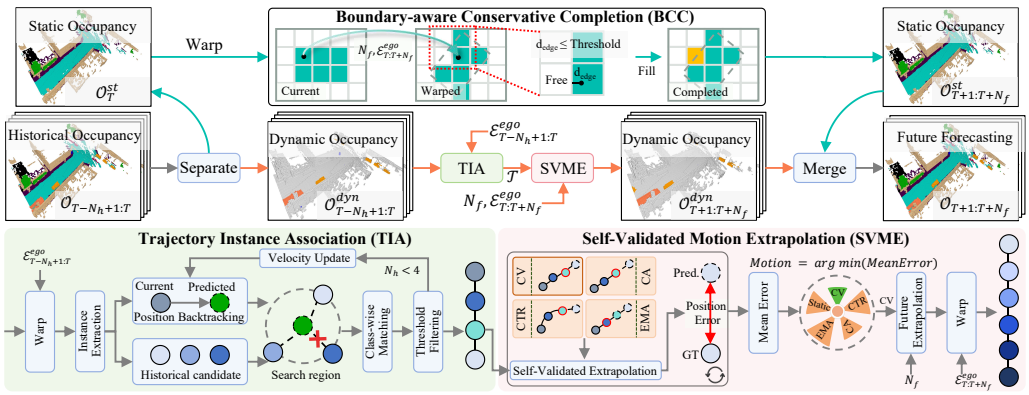


Figure 3: **MotionWorld** consists of two branches for future semantic occupancy: a dynamic branch that associates instances across ego-aligned historical occupancies via TIA, selects per-instance motion hypotheses through historical replay using SVME, and rigidly propagates each instance; and a static branch that warps the current static occupancy and fills warping-induced holes using BCC.

known historical ego-motion $\mathcal{E}^{ego}_{T-N_h+1:T}$. *Self-Validated Motion Extrapolation* (SVME, **§3.3**) then selects instance-specific motion hypotheses through historical replay and uses N_f and future ego-motion $\mathcal{E}^{ego}_{T:T+N_f}$ to produce the final dynamic forecast $\mathcal{O}^{dyn}_{T+1:T+N_f}$. In parallel, the static branch propagates $\mathcal{O}^{st}_{T}$ into the next N_f ego frames using $\mathcal{E}^{ego}_{T:T+N_f}$ and applies *Boundary-aware Conservative Completion* (BCC, **§3.4**) to fill warping-induced boundary gaps, producing $\mathcal{O}^{st}_{T+1:T+N_f}$. The propagated static and dynamic components are then merged voxel-wise to obtain the final forecast $\mathcal{O}_{T+1:T+N_f}$, with dynamic labels taking precedence in voxel conflicts. Furthermore, **§3.5** introduces **MotionWorld-Refine**, a module that learns lightweight corrections to refine the merged predictions.

3.2 Trajectory Instance Association

TIA constructs representations for each current dynamic instance by associating instances across the dynamic occupancy sequence $\mathcal{O}^{dyn}_{T-N_h+1:T}$ using known historical ego-motion $\mathcal{E}^{ego}_{T-N_h+1:T}$. Its overall process is defined by

$$\mathcal{T} = \{(I_i^T, \tau_i, \mathbf{v}_i)\}_{i=1}^{N_T} = \mathrm{TIA}\left(\mathcal{O}^{dyn}_{T-N_h+1:T}, \mathcal{E}^{ego}_{T-N_h+1:T}\right), \tag{1}$$

where N_T is the number of current dynamic instances. For the i-th instance, I_i^T, τ_i, and $\mathbf{v}_i$ denote its complete current instance, sparse historical BEV trajectory, and estimated planar velocity, respectively. TIA first uses $\mathcal{E}^{ego}_{T-N_h+1:T}$ to warp the historical dynamic occupancies $\mathcal{O}^{dyn}_{T-N_h+1:T-1}$ into the ego coordinate frame at time T, yielding the aligned sequence $\bar{\mathcal{O}}^{dyn}_{T-N_h+1:T-1}$.

Complete current instance I_i^T**.** We obtain current instances by identifying connected components of same-class occupied voxels in the BEV projection of $\mathcal{O}^{dyn}_{T}$. Each instance I_i^T is represented by its semantic class, occupied voxel set, and BEV center (Li et al., 2024) $\boldsymbol{\mu}^T_{i,xy}$. These instances form the current instance set $\mathcal{I}_T = \{I_i^T\}_{i=1}^{N_T}$. Applying the same extraction to $\bar{\mathcal{O}}^{dyn}_{T-N_h+1:T-1}$ yields the candidate sets $\{\bar{\mathcal{I}}^T_{T-a}\}_{a=1}^{N_h-1}$.

Velocity-guided trajectory recovery $(\tau_i, \mathbf{v}_i)$**.** Using $\mathcal{I}_T$ and $\{\bar{\mathcal{I}}^T_{T-a}\}_{a=1}^{N_h-1}$, we initialize each trajectory at its current BEV center, $\tau_i(0) = \boldsymbol{\mu}^T_{i,xy}$, with zero velocity $\mathbf{v}_i = \mathbf{0}$. The historical frames are then processed from recent to distant. Temporal offsets are measured in frames, and velocities are expressed as planar displacement per frame. At offset a, the current center is backtracked as

$$\mathbf{q}_i^{(a)} = \boldsymbol{\mu}^T_{i,xy} - \mathbb{I}[\|\mathbf{v}_i\|_2 > \theta_v]\,\sqrt{a}\,\mathbf{v}_i, \tag{2}$$

where $\mathbb{I}[\cdot]$ is the indicator function and θ_v is a speed threshold that suppresses backtracking for nearly static instances. Instead of the linear displacement $a\mathbf{v}_i$, we use $\sqrt{a}\mathbf{v}_i$ so that the backtracking distance grows sublinearly with the temporal offset, reducing the influence of velocity errors on distant

historical frames. Within each semantic class, we perform one-to-one Hungarian matching (Kuhn, 1955) between the backtracked centers and the candidate centers in $\bar{\mathcal{I}}_{T-a}^{T}$. Matches are then filtered by distance and voxel-count ratio.

For an accepted match with center $\bar{\boldsymbol{\mu}}_{j,xy}^{T-a}$, we set $\tau_i(-a) = \bar{\boldsymbol{\mu}}_{j,xy}^{T-a}$ and compute a velocity observation from its displacement to the current center. The running velocity is updated as

$$\mathbf{v}_i \leftarrow \mathbf{v}_i + \mathbf{K}_i^{(a)}\big(\mathbf{z}_i^{(a)} - \mathbf{v}_i\big), \qquad \mathbf{z}_i^{(a)} = \frac{\boldsymbol{\mu}_{i,xy}^{T} - \bar{\boldsymbol{\mu}}_{j,xy}^{T-a}}{a}, \tag{3}$$

where $\mathbf{K}_i^{(a)}$ is an age-aware Kalman gain (Basar, 2001). The updated velocity guides association in the next historical frame, jointly producing the sparse trajectory τ_i and the final velocity estimate $\mathbf{v}_i$.

Collecting the current geometry and recovered motion of all instances gives $\mathcal{T} = \{(I_i^T, \tau_i, \mathbf{v}_i)\}_{i=1}^{N_T}$, which is passed to SVME for future extrapolation.

3.3 Self-Validated Motion Extrapolation

Given the representations $\mathcal{T}$ from TIA, the prediction horizon N_f, and future ego-motion $\mathcal{E}_{T:T+N_f}^{ego}$, SVME selects an instance-specific motion hypothesis and uses it to forecast dynamic occupancy:

$$\mathcal{O}_{T+1:T+N_f}^{dyn} = \bigcup_{i=1}^{N_T} O_i^{T+1:T+N_f}(m_i^\star) = \mathrm{SVME}\Big(\mathcal{T}, N_f, \mathcal{E}_{T:T+N_f}^{ego}\Big). \tag{4}$$

Here, $m_i^\star$ denotes the motion hypothesis selected for instance i, and $O_i^{T+1:T+N_f}(m_i^\star)$ denotes its future occupancy under $m_i^\star$, expressed in the corresponding future ego frames. Aggregating these instance predictions yields the dynamic forecast $\mathcal{O}_{T+1:T+N_f}^{dyn}$.

Selected motion hypothesis $m_i^\star$**.** SVME evaluates a compact motion bank $\mathcal{M} = \{\mathrm{Static}, \mathrm{CV}, \mathrm{CA}, \mathrm{CTR}, \mathrm{EMA}\}$, comprising stationary, constant-velocity, constant-acceleration, constant-turn-rate, and exponentially smoothed motion hypotheses.

For a current instance I_i^T, we evaluate each candidate by replaying its observed trajectory τ_i. Its valid historical anchor–endpoint pairs are

$$\mathcal{R}_i = \{(s, \ell) \mid s < 0,\ \ell > 0,\ s + \ell \le 0,\ s, s+\ell \in \mathrm{dom}(\tau_i)\}, \tag{5}$$

where $\mathrm{dom}(\tau_i)$ contains the offsets of observed positions relative to the current frame T. Offset 0 corresponds to T, and negative offsets indicate historical frames. At each anchor s, candidate m uses only observations at or before s to predict a displacement $\boldsymbol{\Delta}_{i,m}^{(s)}(\ell)$. Its prediction is compared with the observed endpoint to obtain the mean replay error

$$E_i(m) = \frac{1}{|\mathcal{R}_i|} \sum_{(s,\ell)\in\mathcal{R}_i} \Big\| \tau_i(s) + \boldsymbol{\Delta}_{i,m}^{(s)}(\ell) - \tau_i(s+\ell) \Big\|_2. \tag{6}$$

Because the endpoint is excluded when estimating the candidate's motion, the replay tests its ability to predict an observed but withheld position. For instance i, we select the hypothesis with the lowest valid replay error:

$$m_i^\star = \arg\min_{m\in\mathcal{M}} E_i(m), \tag{7}$$

If no valid pair exists, SVME falls back to the velocity estimated by TIA, or to stationary motion when that estimate is unavailable.

Future instance occupancy $O_i^{T+1:T+N_f}$**.** With $m_i^\star$ selected, SVME uses the full observed trajectory τ_i to predict the future planar displacements of instance i:

$$\boldsymbol{\Delta}_i^{1:N_f} = \boldsymbol{\Delta}_{i,m_i^\star}(1:N_f). \tag{8}$$

Let O_i^T denote the occupied voxels of its current instance I_i^T. Applying the same displacement to every voxel at each future step gives

$$\bar{O}_i^{T+1:T+N_f} = \mathrm{FutureExtrapolate}\Big(O_i^T, \boldsymbol{\Delta}_i^{1:N_f}\Big). \tag{9}$$

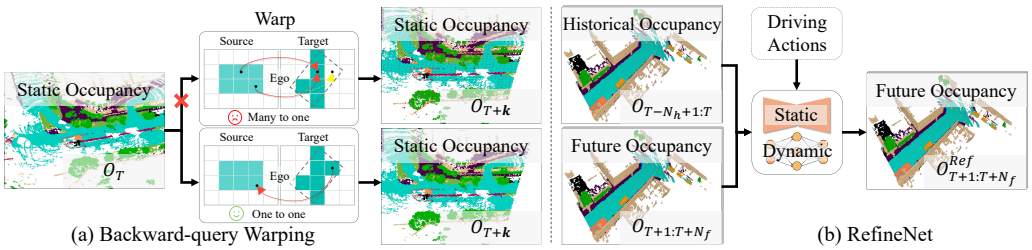


Figure 4: **Static propagation and learned refinement.** (a) Backward warping with boundary-aware conservative completion. (b) **MotionWorld-Refine** corrects static occupancy and dynamic motion.

This rigid translation preserves the observed instance geometry, including its vertical structure. We combine the extrapolated instances into $\bar{\mathcal{O}}^{dyn}_{T+1:T+N_f}$ in the current ego frame, then transform them into the corresponding future ego frames:

$$\mathcal{O}^{dyn}_{T+1:T+N_f} = \mathrm{Warp}\Big(\bar{\mathcal{O}}^{dyn}_{T+1:T+N_f}, \mathcal{E}^{ego}_{T:T+N_f}\Big). \tag{10}$$

The resulting dynamic forecast is subsequently combined with the static forecast produced by BCC.

3.4 Boundary-Aware Conservative Completion

Given the current static occupancy $\mathcal{O}^{st}_T$ and future ego-motion $\mathcal{E}^{ego}_{T:T+N_f}$, we first propagate the static scene into future ego frames and then complete gaps along its transformed boundaries:

$$\mathcal{O}^{st}_{T+1:T+N_f} = \mathrm{BCC}\Big(\mathrm{Warp}\Big(\mathcal{O}^{st}_T, \mathcal{E}^{ego}_{T:T+N_f}\Big), \tau_{\mathrm{edge}}, N_f\Big) \tag{11}$$

where τ_{edge} controls the extent of boundary completion.

Warped static occupancy. Static structures remain fixed in the world, so their apparent displacement arises from changes in the ego coordinate frame. For each future step $k \in \{1, \ldots, N_f\}$, the known ego-motion $\mathcal{E}^{ego}_{T:T+N_f}$ determines the relative transform $\mathbf{T}^{ego}_{T\to T+k}$ by composing the pose changes from T to $T+k$. For a target voxel centered at $\mathbf{y}$ in frame $T+k$, we apply the inverse transform to locate its corresponding position in the current frame:

$$\bar{\mathcal{O}}^{st}_{T+k}(\mathbf{y}) = \mathcal{O}^{st}_T\Big(\big(\mathbf{T}^{ego}_{T\to T+k}\big)^{-1}\mathbf{y}\Big). \tag{12}$$

Sampling the current static occupancy at these source positions avoids the many-to-one collisions of forward scattering and yields the warped static occupancy $\bar{\mathcal{O}}^{st}_{T+1:T+N_f}$, as illustrated in Fig. 4(a).

Completed static occupancy. Voxel discretization can still leave thin free-space gaps along the warped semantic boundaries. For a transformed occupied voxel, BCC considers the neighboring target cell across a boundary only when its distance to that boundary satisfies $d_{\mathrm{edge}} \le \tau_{\mathrm{edge}}$. The source semantic label is copied only if the neighboring cell remains free after warping. Thus, BCC fills local boundary gaps without overwriting existing occupancy, producing $\mathcal{O}^{st}_{T+1:T+N_f}$.

Finally, the completed static occupancy is combined with $\mathcal{O}^{dyn}_{T+1:T+N_f}$ from SVME. Dynamic labels take precedence where the two components overlap, yielding the MotionWorld forecast $\mathcal{O}_{T+1:T+N_f}$.

3.5 MotionWorld-Refine

MotionWorld-Refine uses MotionWorld's analytical forecast as a structured prior and learns residual corrections for static occupancy and dynamic instance motion. As shown in Fig. 4(b), its overall mapping is

$$\mathcal{O}^{ref}_{T+1:T+N_f} = \mathrm{Refine}\Big(\mathcal{O}_{T-N_h+1:T}, \mathcal{O}_{T+1:T+N_f}, \mathcal{E}^{ego}_{T-N_h+1:T}, \mathcal{E}^{ego}_{T:T+N_f}\Big), \tag{13}$$

where $\mathcal{O}^{ref}_{T+1:T+N_f}$ is the refined future occupancy sequence.

Table 1: Short-horizon occupancy forecasting on Occ3D-nuScenes with ground-truth and camera-predicted inputs, evaluated by semantic mIoU and geometric IoU at 1/2/3s and their averages (no camera-visible mask). Dark/light purple mark the best/second-best results per input setting.

		Semantic mIoU (↑)				Geometric IoU (↑)					
Method	Input	1 s	2 s	3 s	Avg.	1 s	2 s	3 s	Avg.	Params	Venue
I^2-World	Occ GT	47.62	38.58	32.98	39.73	54.29	49.43	45.69	49.80	22.71M	ICCV 2025
Occ-LLM	Occ GT	24.02	21.65	17.29	20.99	36.65	32.14	28.77	35.52	7B	ICRA 2025
RenderWorld	Occ GT	28.69	18.89	14.83	20.89	37.74	28.41	24.08	30.08	N/A	ICRA 2025
OccLLaMA-O	Occ GT	25.05	19.49	15.26	19.93	34.56	28.53	24.41	29.17	7B	ICRA 2026
GenieDrive	Occ GT	50.47	41.47	35.83	42.59	56.87	51.46	47.08	51.80	3.47M	CVPR 2026
GAST	Occ GT	55.54	46.33	40.18	47.40	60.77	55.97	51.96	56.24	12.09M	ECCV 2026
MotionWorld (Ours)	Occ GT	54.48	43.47	37.40	45.12	61.27	54.88	50.88	55.68	0	–
MotionWorld-Refine (Ours)	Occ GT	**57.09**	**47.40**	**41.49**	**48.66**	**65.86**	**61.12**	**57.10**	**61.38**	1.58M	–
I^2-World	Camera	21.67	18.78	16.47	18.97	30.55	28.76	26.99	28.77	22.71M	ICCV 2025
Occ-LLM	Camera	11.28	10.21	9.13	10.21	27.11	24.07	20.19	23.79	7B	ICRA 2025
RenderWorld	Camera	2.83	2.55	2.37	2.58	14.61	13.61	12.98	13.73	N/A	ICRA 2025
OccTENS-F	Camera	17.17	10.38	7.82	11.79	27.60	25.14	20.33	24.35	N/A	RAL 2026
OccLLaMA	Camera	10.34	8.66	6.98	8.66	25.81	23.19	19.97	22.99	7B	ICRA 2026
SparseWorld	Camera	14.93	13.15	11.51	13.20	22.96	22.10	21.05	22.03	N/A	AAAI 2026
SparseWorld-TC-Small	Camera	16.47	15.08	13.94	15.16	22.80	22.35	21.70	22.28	79.43M	CVPR 2026
GAST	Camera	22.90	19.93	17.17	20.16	**31.62**	**29.92**	**28.00**	**29.87**	12.09M	ECCV 2026
MotionWorld (Ours)	Camera	22.47	19.29	17.04	19.62	29.95	27.69	25.75	27.82	0	–
MotionWorld-Refine (Ours)	Camera	**22.93**	**20.26**	**18.11**	**20.53**	31.22	29.49	27.66	29.49	1.58M	–

Static residual actions. Historical occupancies are aligned with the future ego frames and compared with the analytical static forecast. The static branch predicts a voxel-wise action: KEEP the forecast label, ERASE an unsupported voxel, or WRITE a static class. Applying these actions corrects missing, spurious, and mislabeled voxels while retaining static content that needs no change.

Dynamic displacement corrections. For each dynamic instance, the refinement branch predicts a bounded planar position-correction sequence $\delta\mathbf{p}_i^{1:N_f}$ from the aligned history and analytical forecast. It updates the displacement produced by SVME as

$$\mathbf{\Delta}_i^{\mathrm{ref},\,1:N_f} = \mathbf{\Delta}_i^{1:N_f} + \delta\mathbf{p}_i^{1:N_f}. \tag{14}$$

The corrected displacement is applied to the complete current instance using the same rigid extrapolation and future ego warp as SVME. It adjusts the predicted trajectory without changing the instance's observed geometry. The refined static and dynamic occupancies are then combined, with dynamic labels taking precedence in voxel conflicts. Training strategy is provided in Appendix A.1.

4 EXPERIMENTS

We evaluate MotionWorld and MotionWorld-Refine against the previous methods (Zhang et al., 2024; Gu et al., 2024; Zheng et al., 2024; Liao et al., 2025b; Xu et al., 2025b; Yang et al., 2026; Yan et al., 2025b; Wei et al., 2024; Du et al., 2026; Jin et al., 2026; Chen et al., 2026; Dang et al., 2026) on Occ3D-nuScenes (Tian et al., 2023) under both ground-truth and camera-predicted occupancy inputs. Additional experimental details are provided in Appendix A.1 and Appendix A.2.

4.1 MAIN RESULTS

Forecasting accuracy. As shown in Table 1, the parameter-free MotionWorld achieves 44.84 average mIoU and 55.62 average IoU with Occ GT input, remaining competitive with GAST. MotionWorld-Refine further establishes the best results of 48.66 mIoU and 61.38 IoU, outperforming GAST by 1.26 and 5.14 points with only 1.58M parameters. Under the Camera setting, MotionWorld-Refine achieves the best average mIoU of 20.53, while its average IoU of 29.49 remains 0.38 points below GAST. One possible reason is that our structure-preserving formulation inherits geometric errors from predicted inputs that localized refinement may not fully recover.

Long-horizon stability. We further evaluate long-horizon forecasting up to eight seconds, as reported in Table 2. Averaged across all forecasting horizons, MotionWorld outperforms GAST by 9.85 mIoU and 6.55 IoU points, while MotionWorld-Refine raises the scores to 13.29 and 11.81 points. The persistent advantage at increasing horizons confirms that propagating static content and complete instances limits the error accumulation caused by repeated scene generation.

Table 2: Long-horizon occupancy forecasting on Occ3D-nuScenes from 1 to 8 seconds. Dark and light purple denote the best and second-best result at each horizon, respectively.

Method	Semantic mIoU (%) ↑								
	1s	2s	3s	4s	5s	6s	7s	8s	Avg.
COME (Shi et al., 2026)	33.78	24.57	21.35	18.25	15.84	13.85	12.99	11.96	19.07
GAST (Chen et al., 2026)	36.23	30.61	27.13	24.36	21.98	19.81	17.83	15.96	24.12
MotionWorld (Ours)	54.84	43.55	37.20	32.91	29.65	26.84	24.41	22.22	33.97
MotionWorld-Refine (Ours)	**57.47**	**47.53**	**41.28**	**36.77**	**33.09**	**29.78**	**26.81**	**24.00**	**37.41**
	Geometric IoU (%) ↑								
Method	1s	2s	3s	4s	5s	6s	7s	8s	Avg.
COME (Shi et al., 2026)	44.20	36.25	32.86	30.03	26.93	24.70	23.30	21.44	29.96
GAST (Chen et al., 2026)	49.09	45.11	42.11	39.52	37.19	34.99	32.87	30.87	39.06
MotionWorld (Ours)	61.30	54.49	50.07	46.31	42.72	39.26	35.93	32.85	45.61
MotionWorld-Refine (Ours)	**65.90**	**60.81**	**56.33**	**52.33**	**48.37**	**44.43**	**40.49**	**36.59**	**50.87**

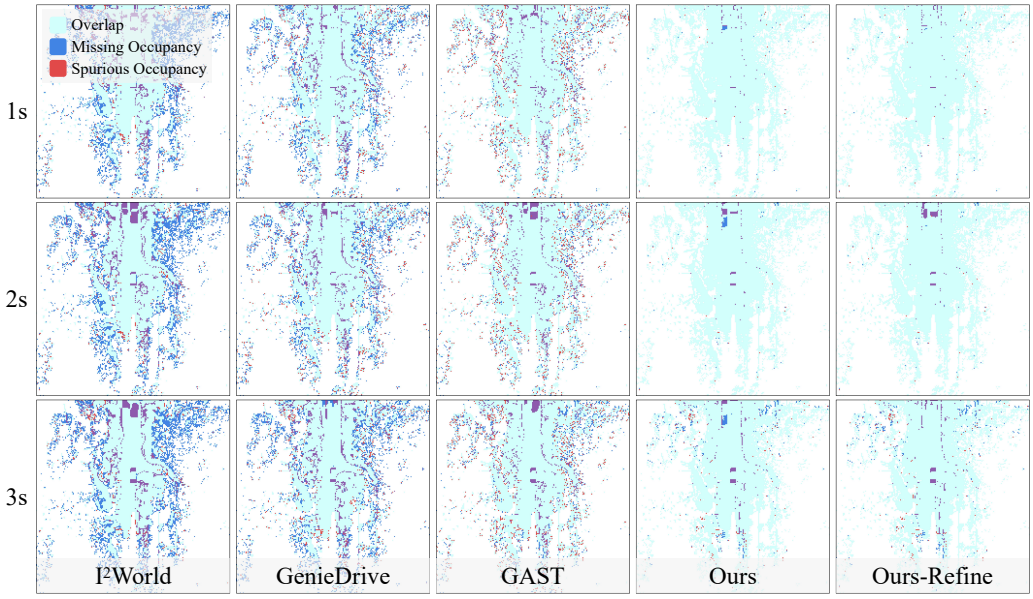


Figure 5: Qualitative occupancy forecasting on Occ3D-nuScenes at 1s/2s/3s. Cyan denotes overlap with the ground truth, whereas blue and red indicate missing and spurious occupancy, respectively.

Accuracy–efficiency trade-off. The complexity–accuracy comparison in Table 3 shows that MotionWorld requires no learned parameters and reaches 265.19 FPS on two CPU cores, 3.5× the FPS reported by GAST. MotionWorld-Refine maintains 82.48 FPS while achieving the best accuracy. The two variants therefore offer complementary operating points: maximum efficiency without training, or lightweight refinement for higher forecasting accuracy.

Visual Comparison. Figure 5 visualizes how forecasting errors evolve from one to three seconds. The baselines exhibit increasingly dense missing occupancy and scattered false positives, indicating that generation errors accumulate with the horizon. In contrast, MotionWorld preserves coherent regions by propagating observed static content and complete dynamic instances. MotionWorld-Refine retains this structural continuity while removing residual missing and spurious voxels.

4.2 Ablation Studies

Training-free analytical core. Table 4 examines how explicit scene evolution contributes to MotionWorld. Persistence performs poorly because it ignores both ego-motion and object motion, whereas whole-scene warping raises mIoU from 12.73 to 41.58 and IoU from 21.89 to 54.98. Separating static and dynamic content in the structured core preserves this geometric accuracy but, by itself, does not improve semantic alignment. Associating instances with TIA and extrapolating them under a fixed constant-velocity model increases mIoU to 44.21. Replacing backward-query warping with forward scattering reduces mIoU from 44.84 to 44.74, indicating that backward querying

Table 3: Accuracy–efficiency comparison on Occ3D-nuScenes under ground-truth occupancy input. MotionWorld is evaluated on CPU. FLOPs account only for neural network operations.

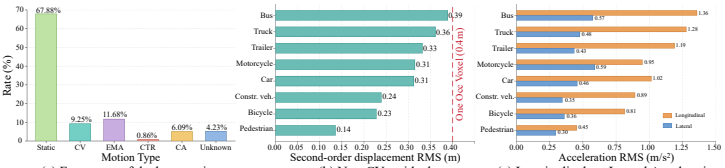
Method	#Params ↓	FLOPs ↓	FPS ↑	mIoU ↑	IoU ↑
DOME (Gu et al., 2024)	444.07M	2928.66G	6.54	27.10	36.36
COME (Shi et al., 2026)	692.97M	4461.46G	0.30	34.23	44.13
I^2-World Liao et al. (2025b)	22.67M	494.58G	37.04	39.73	49.80
GAST (Chen et al., 2026)	12.09M	416.27G	74.97	47.40	56.24
MotionWorld	**0**	**0**	**265.19**	44.84	55.62
MotionWorld-Refine	1.58M	161.93G	82.48	**48.66**	**61.38**

Table 4: MotionWorld's core ablations.

Variant	mIoU ↑	IoU ↑
Persistence	12.73	21.89
Whole-scene Warp	41.58	54.98
Structured Core	41.41	55.00
TIA + Fixed CV	44.21	55.48
w/o Backward Warp	44.74	55.56
w/o BCC	44.60	55.00
MotionWorld	44.84	55.62
+ Static Refine	46.96	61.10
+ Dynamic Refine	46.15	55.81
MotionWorld-Refine	**48.66**	**61.38**

Table 5: Motion ablation.

Motion bank	Δ mIoU	Δ IoU
Ours	+0.0000	+0.0000
w/o Static	−1.9612	−0.4503
w/o CV	+0.0091	−0.0018
w/o EMA	−0.2607	−0.0262
w/o CTR	+0.0023	−0.0024
w/o CA	−0.0023	−0.0085
Ours	+0.0000	+0.0000
Ours + VD	−0.0956	−0.0073



(a) Frequency of the best motion (b) Non-CV residual (c) Longitudinal vs. Lateral Acceleration

Figure 6: Motion analysis: (a) best-fitting hypotheses, (b) non-CV residuals, and (c) longitudinal and lateral accelerations.

better preserves transformed occupancy under voxel discretization. Removing BCC further lowers mIoU to 44.60 and IoU to 55.00, confirming its role in recovering static boundary occupancy. Replacing the shared motion assumption with the self-validated motion models of SVME ultimately achieves 44.84 mIoU and 55.62 IoU, supporting instance-specific motion selection together with geometry-preserving warping and completion.

Static and dynamic refinement. Table 4 evaluates both refinement branches on top of MotionWorld. Static refinement improves mIoU from 44.84 to 46.96 and IoU from 55.62 to 61.10, consistent with repairing missing and spurious occupancy. Dynamic refinement reaches 46.15 mIoU and 55.81 IoU by correcting instance trajectories while preserving their geometry. Combining both branches achieves the best results of 48.66 mIoU and 61.38 IoU, supporting their complementary roles in correcting the analytical forecast.

Motion-bank composition. Figure 6(a) shows that stationary motion accounts for 67.88% of the best-fitting hypotheses. Class-wise non-CV displacement residuals remain below the 0.4 m voxel size in Fig. 6(b), suggesting that quantization can mask fine-grained departures from constant velocity. This helps explain why TIA with fixed CV already achieves 44.21 mIoU. Nevertheless, the predominantly longitudinal acceleration in Fig. 6(c) motivates the inclusion of acceleration-aware motion candidates. Table 5 shows that removing the stationary candidate or EMA reduces mIoU by 1.9612 or 0.2607, respectively. Omitting CV, CTR, or CA yields marginal changes, suggesting overlapping coverage, while adding velocity-decay (VD) motion slightly reduces both metrics. We retain the five-candidate bank to cover the observed motion diversity, aiming for robustness across trajectories despite the small aggregate gains of some candidates.

Additional ablations, limitations, and visualizations are provided in Appendices B, A.4, and C, respectively.

5 Conclusion

We presented MotionWorld, a training-free analytical framework that formulates 4D occupancy forecasting as explicit scene evolution rather than repeated full-scene generation. It disentangles static persistence from dynamic instance motion through ego-motion-aware propagation, instance association, and self-validated motion extrapolation, preserving scene structure while reducing long-horizon error accumulation. MotionWorld-Refine further learns targeted corrections on top of this structured analytical prior. Experiments on Occ3D-nuScenes show that MotionWorld achieves competitive accuracy at 265.19 FPS on CPU, while MotionWorld-Refine delivers state-of-the-art performance under GT occupancy. These results highlight explicit scene evolution as an efficient and interpretable foundation for occupancy world models.

AI USE STATEMENT

In this work, we used OpenAI's generative AI tools to assist with English translation and language editing. All AI-assisted suggestions and text were manually reviewed and verified by the authors. The authors retained full control over the method, implementation, experimental results, and final scientific claims. We take responsibility for the final content of this work, including all text and artifacts produced with the assistance of generative AI.

REFERENCES

Tamer Basar. *A New Approach to Linear Filtering and Prediction Problems*, pp. 167–179. 2001.

Bert Besser and Matthias Poloczek. Greedy matching: Guarantees and limitations. *Algorithmica*, pp. 201–234, 2017.

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 *CVPR*, pp. 11621–11631, 2020.

Yuning Chai, Benjamin Sapp, Mayank Bansal, and Dragomir Anguelov. Multipath: Multiple probabilistic anchor trajectory hypotheses for behavior prediction. In *CoRL*, pp. 86–99, 2020.

Dubing Chen, Jin Fang, Wencheng Han, Xinjing Cheng, Junbo Yin, Chengzhong Xu, Fahad Shahbaz Khan, and Jianbing Shen. Alocc: Adaptive lifting-based 3d semantic occupancy and cost volume-based flow predictions. In *ICCV*, pp. 4156–4166, 2025.

Sitao Chen, Zhuangwei Zhuang, Hui Luo, Qingyao Wu, and Mingkui Tan. Geometry-aware spatio-temporal context modeling for 4d occupancy forecasting. *arXiv preprint arXiv:2608.15279*, 2026.

Chenxu Dang, Haiyan Liu, Jason Bao, Pei An, Xinyue Tang, An Pan, Jie Ma, Bingchuan Sun, and Yan Wang. Sparseworld: A flexible, adaptive, and efficient 4d occupancy world model powered by sparse and dynamic queries. In *AAAI*, pp. 3497–3505, 2026.

Jiayuan Du, Yiming Zhao, Zhenglong Guo, Yong Pan, Wenbo Hou, Zhihui Hao, Kun Zhan, and Qijun Chen. Sparseworld-tc: Trajectory-conditioned sparse occupancy world model. In *CVPR*, pp. 7425–7434, 2026.

Jiyang Gao, Chen Sun, Hang Zhao, Yi Shen, Dragomir Anguelov, Congcong Li, and Cordelia Schmid. Vectornet: Encoding hd maps and agent dynamics from vectorized representation. In *CVPR*, pp. 11522–11530, 2020.

Songen Gu, Wei Yin, Bu Jin, Xiaoyang Guo, Junming Wang, Haodong Li, Qian Zhang, and Xiaoxiao Long. Dome: Taming diffusion model into high-fidelity controllable occupancy world model. *arXiv preprint arXiv:2410.10429*, 2024.

Junjie Huang, Guan Huang, Zheng Zhu, Yun Ye, and Dalong Du. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*, 2021.

Bu Jin, Songen Gu, Xiaotao Hu, Yupeng Zheng, Xiaoyang Guo, Qian Zhang, Xiaoxiao Long, and Wei Yin. Occtens: 3d occupancy world model via temporal next-scale prediction. *IEEE Robotics and Automation Letters*, pp. 3566–3573, 2026.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Harold W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, pp. 83–97, 1955.

Bohan Li, Jiazhe Guo, Hongsi Liu, Yingshuang Zou, Yikang Ding, Xiwu Chen, Hu Zhu, Feiyang Tan, Chi Zhang, Tiancai Wang, et al. Uniscene: Unified occupancy-centric driving scene generation. In *CVPR*, pp. 11971–11981, 2025a.

Xiang Li, Pengfei Li, Yupeng Zheng, Wei Sun, Yan Wang, et al. Semi-supervised vision-centric 3d occupancy world model for autonomous driving. In *ICLR*, pp. 62563–62580, 2025b.

Zhiqi Li, Zhiding Yu, David Austin, Mingsheng Fang, Shiyi Lan, Jan Kautz, and Jose M Alvarez. Fb-occ: 3d occupancy prediction based on forward-backward view transformation. *arXiv preprint arXiv:2307.01492*, 2023.

Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Qiao Yu, and Jifeng Dai. Bevformer: learning bird's-eye-view representation from lidar-camera via spatiotemporal transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 2020–2036, 2024.

Zhimin Liao, Ping Wei, Shuaijia Chen, Haoxuan Wang, and Ziyang Ren. Stcocc: Sparse spatial-temporal cascade renovation for 3d occupancy and scene flow prediction. In *CVPR*, pp. 1516–1526, 2025a.

Zhimin Liao, Ping Wei, Ruijie Zhang, Shuaijia Chen, Haoxuan Wang, and Ziyang Ren. I2-World: Intra-inter tokenization for efficient dynamic 4d scene forecasting. In *ICCV*, pp. 25810–25819, 2025b.

Haisong Liu, Yang Chen, Haiguang Wang, Zetong Yang, Tianyu Li, Jia Zeng, Li Chen, Hongyang Li, and Limin Wang. Fully sparse 3d occupancy prediction. In *ECCV*, pp. 54–71, 2024.

Tianran Liu, Shengwen Zhao, Mozhgan Pourkeshavarz, Weican Li, and Nicholas Rhinehart. Occ-sim: Multi-kilometer simulation with long-horizon occupancy world models. *arXiv preprint arXiv:2603.28887*, 2026.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Jinhyung Park, Chensheng Peng, Wenzhao Zheng, Kris Kitani, Wei Zhan, et al. S2go: Streaming sparse gaussian occupancy. In *ICLR*, pp. 62911–62932, 2026.

Tung Phan-Minh, Elena Corina Grigore, Freddy A Boulton, Oscar Beijbom, and Eric M Wolff. Covernet: Multimodal behavior prediction using trajectory sets. In *CVPR*, pp. 14062–14071, 2020.

Shaoshuai Shi, Li Jiang, Dengxin Dai, and Bernt Schiele. Motion transformer with global intention localization and local movement refinement. *NeurIPS*, pp. 6531–6543, 2022.

Yining Shi, Kun Jiang, Qiang Meng, Ke Wang, Jiabao Wang, Wenchao Sun, Tuopu Wen, et al. Come: Adding scene-centric forecasting control to occupancy world model. *NeurIPS*, pp. 70–96, 2026.

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. *NeurIPS*, pp. 64318–64330, 2023.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *NeurIPS*, 2017.

Julong Wei, Shanshuai Yuan, Pengfei Li, Qingda Hu, Zhongxue Gan, and Wenchao Ding. Occllama: An occupancy-language-action generative world model for autonomous driving. *arXiv preprint arXiv:2409.03272*, 2024.

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

Tianshuo Xu, Hao Lu, Xu Yan, Yingjie Cai, Bingbing Liu, and Yingcong Chen. Occ-llm: Enhancing autonomous driving with occupancy-based large language models. In *ICRA*, pp. 8434–8441, 2025b.

Tianyi Yan, Dongming Wu, Wencheng Han, Junpeng Jiang, Xia Zhou, Kun Zhan, Cheng-zhong Xu, and Jianbing Shen. Drivingsphere: Building a high-fidelity 4d world for closed-loop simulation. In *CVPR*, pp. 27531–27541, 2025a.

Ziyang Yan, Wenzhen Dong, Yihua Shao, Yuhang Lu, Haiyang Liu, Jingwen Liu, Haozhe Wang, Zhe Wang, Yan Wang, Fabio Remondino, and Yuexin Ma. Renderworld: World model with self-supervised 3d label. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6063–6070, 2025b.

Yu Yang, Jianbiao Mei, Yukai Ma, Siliang Du, Wenqing Chen, Yijie Qian, Yuxiang Feng, and Yong Liu. Driving in the occupancy world: Vision-centric 4d occupancy forecasting and planning via world models for autonomous driving. In *AAAI*, pp. 9327–9335, 2025.

Zhenya Yang, Zhe Liu, Yuxiang Lu, Liping Hou, Chenxuan Miao, Siyi Peng, Bailan Feng, Xiang Bai, and Hengshuang Zhao. Geniedrive: Towards physics-aware driving world model with 4d occupancy guided video generation. In *CVPR*, pp. 35680–35690, 2026.

Haiming Zhang, Ying Xue, Xu Yan, Jiacheng Zhang, Weichao Qiu, Dongfeng Bai, Bingbing Liu, Shuguang Cui, and Zhen Li. An efficient occupancy world model via decoupled dynamic flow and image-assisted training. *arXiv preprint arXiv:2412.13772*, 2024.

Wenzhao Zheng, Weiliang Chen, Yuanhui Huang, Borui Zhang, Yueqi Duan, and Jiwen Lu. Occworld: Learning a 3d occupancy world model for autonomous driving. In *ECCV*, pp. 55–72, 2024.

A Additional Experiments

A.1 Training Strategy

We supervise static residual actions with a class-balanced focal loss, soft static-IoU loss, and occupancy Dice loss:

$$\mathcal{L}_{\mathrm{st}} = \mathcal{L}_{\mathrm{act}} + \lambda_{\mathrm{sIoU}}\mathcal{L}_{\mathrm{sIoU}} + \lambda_{\mathrm{occ}}\mathcal{L}_{\mathrm{occ}}. \tag{15}$$

Dynamic trajectory corrections are supervised with position, soft instance-alignment, and confidence-gating losses. The full objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{st}} + \lambda_{\mathrm{pos}}\mathcal{L}_{\mathrm{pos}} + \lambda_{\mathrm{align}}\mathcal{L}_{\mathrm{align}} + \lambda_{\mathrm{gate}}\mathcal{L}_{\mathrm{gate}}, \tag{16}$$

where longer forecasting horizons receive larger weights during dynamic supervision.

A.2 Experimental Setup

Dataset. We conduct experiments on Occ3D-nuScenes (Tian et al., 2023), a widely used semantic occupancy benchmark built upon the large-scale nuScenes dataset (Caesar et al., 2020). It comprises 600 training scenes and 150 validation scenes, with keyframes sampled at 2 Hz. Each sample covers a spatial range of $[-40, 40]$ m along the x- and y-axes and $[-1, 5.4]$ m along the z-axis. With an isotropic voxel size of 0.4 m, the space is discretized into a $200 \times 200 \times 16$ voxel grid. Each voxel is assigned one of 18 labels, comprising 17 semantic categories and one free-space category.

Evaluation metrics. Following prior works (Zheng et al., 2024; Liao et al., 2025b; Du et al., 2026; Chen et al., 2026), we adopt geometric intersection over union (IoU) and semantic mean intersection over union (mIoU) as the evaluation metrics. Geometric IoU evaluates binary occupancy by distinguishing occupied voxels from free space, whereas semantic mIoU measures the class-wise segmentation quality over all non-free semantic categories. We report both metrics at each future timestamp and their average over all timestamps on the validation set.

Implementation details. Following previous work (Liao et al., 2025b), MotionWorld takes five consecutive occupancy frames as input, comprising four historical frames and the current frame, and analytically extrapolates the next six frames. MotionWorld-Refine uses four historical occupancy frames as temporal evidence to refine the six future frames predicted by MotionWorld. Specifically, we employ a lightweight 2D U-Net to predict voxel-wise corrections for static occupancy and an MLP to estimate residual motion corrections for dynamic instances. These refinement modules are trained for 30 epochs using the AdamW optimizer (Loshchilov & Hutter, 2017) with a weight decay of 0.01. The learning rate is initialized to 0.001 and decayed to zero following a cosine annealing schedule. All experiments are conducted with a total batch size of 8 on eight NVIDIA A100 GPUs.

A.3 Performance and Efficiency

Fig. 7 presents scatter plots comparing the performance and inference efficiency of different methods across various settings. Our method achieves a favorable trade-off between forecasting performance and computational efficiency.

A.4 Limitations

Despite its strong structural preservation, MotionWorld remains constrained by its propagation-based formulation. First, the predefined motion primitives cannot fully capture nonlinear behaviors such as turning and interaction-induced motion, resulting in instance-level displacement errors. Second, static propagation can only preserve previously observed scene content. Regions that are occluded or outside the current field of view cannot be recovered when they become visible in future ego frames, leading to large coherent missing regions as the prediction horizon increases.

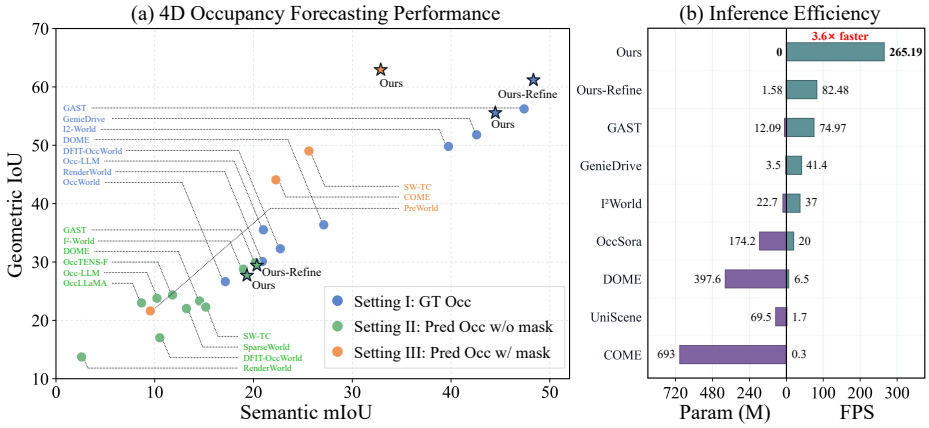


Figure 7: **Performance and FPS.**

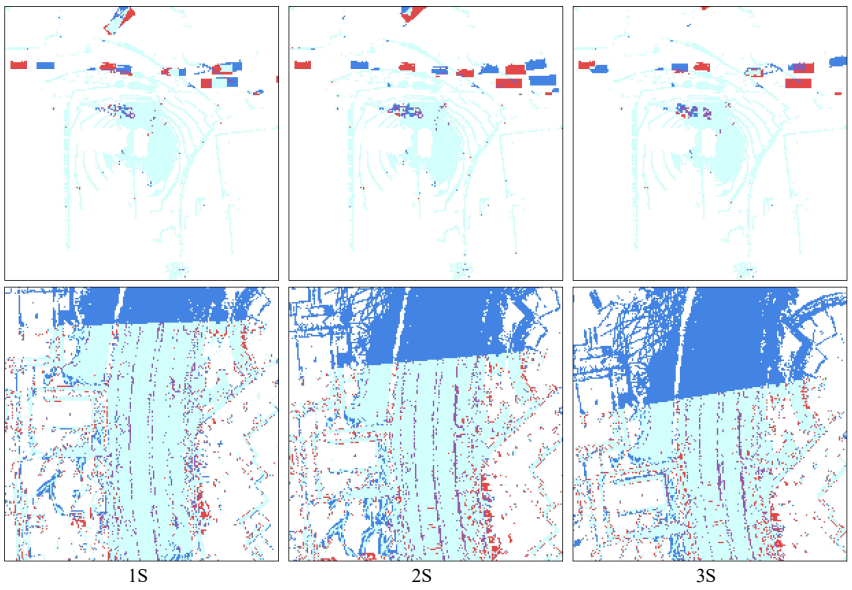


Figure 8: **Limitations of MotionWorld.** Representative errors in dynamic motion extrapolation (top) and static scene propagation (bottom) over increasing prediction horizons.

B Additional Ablations

B.1 Matching type selection.

Table 6 shows that Greedy (Besser & Poloczek, 2017) and Hungarian matching yield comparable forecasting accuracy, indicating that most class-wise association cases are relatively unambiguous. Nevertheless, Greedy provides no measurable efficiency advantage in our setting because the assignment matrices are typically small. We therefore adopt Hungarian matching, which returns an order-independent, globally minimum-cost one-to-one assignment at essentially the same computational cost.

B.2 Effect of History Length.

Table 7 evaluates the sensitivity of MotionWorld to the occupancy history length, where the total number of frames includes the current frame. Increasing the history length from three to five frames improves the overall mIoU from 44.18% to 44.66% and the dynamic mIoU from 47.86% to 49.38%. Extending the history to six frames yields only marginal improvements and slightly decreases IoU. We therefore use five frames in total, including four previous frames and the current frame.

Table 6: Ablation of instance association strategies on Occ3D-nuScenes.

Matcher	mIoU at Horizon (%)			Overall	IoU
	1 s	2 s	3 s	mIoU (%)	(%)
Nearest Neighbor	54.3074	43.3683	37.3408	44.7369	55.5793
Greedy One-to-One	**54.4876**	**43.4899**	**37.4095**	**44.8557**	55.6156
Hungarian (Ours)	54.4801	43.4704	37.3986	44.8427	**55.6163**

Table 7: Sensitivity to occupancy history length.

Total frames	Previous frames	mIoU↑	Dyn. mIoU↑	IoU↑
3	2	44.18	47.86	55.53
4	3	44.45	48.78	55.56
5	4	44.66	49.38	55.59
6	5	44.67	49.45	55.57

Table 8: Generalization of MotionWorld across different occupancy inputs on Occ3D-nuScenes.

Occupancy Input	mIoU (%)				IoU (%)			
	1 s	2 s	3 s	Overall	1 s	2 s	3 s	Overall
BEVDet (Huang et al., 2021)	16.53	14.29	12.48	14.43	25.43	23.26	21.16	23.29
FBOcc (Li et al., 2023)	19.56	16.73	14.76	17.03	26.60	24.47	22.68	24.61
STCOcc (Liao et al., 2025a)	21.51	18.31	16.09	18.64	29.75	27.36	25.36	27.52
ALOcc (Chen et al., 2025)	**22.47**	**19.29**	**17.04**	**19.62**	**29.95**	**27.69**	**25.75**	**27.82**

Table 9: Sensitivity to the main parameters. Bold values are the defaults.

Parameter	Tested values	mIoU range	Dyn. mIoU range	IoU range
d_{max} (m)	10 / **15** / 20	44.43–44.53	48.69–48.97	55.53–55.57
r_{mass}	0.30 / **0.50** / 0.70	44.41–44.56	48.63–49.07	55.54–55.57
τ_{static}	0.30 / **0.40** / 0.50	44.42–44.48	48.66–48.84	55.54–55.55
λ_{comp}	0 / **0.03** / 0.06	44.47–44.48	48.82–48.83	55.55–55.55

B.3 Effect of different occupancy inputs.

We further apply MotionWorld to occupancy predictions from four different perception models without retraining. As reported in Table 8, the same forecasting pipeline remains effective across all input sources. ALOcc (Chen et al., 2025) provides the strongest input and achieves 19.62% mIoU and 27.82% IoU, while the consistent results on BEVDet (Huang et al., 2021), STCOcc (Liao et al., 2025a), and FBOcc (Li et al., 2023) demonstrate the plug-and-play compatibility of MotionWorld.

B.4 Sensitivity to Main Parameters.

Table 9 evaluates the sensitivity of MotionWorld to its main parameters. Performance remains stable across all tested settings: the maximum variations are only 0.15% in mIoU, 0.44% in dynamic mIoU, and 0.04% in IoU. In particular, the completion weight λ_{comp} has a negligible influence within the tested range. These results demonstrate that MotionWorld is robust to parameter variations and does not rely on careful hyperparameter tuning. We use the bold values as the default configuration in all experiments.

C Additional Qualitative Results

We present additional qualitative results in Fig. 9 and Fig. 10.

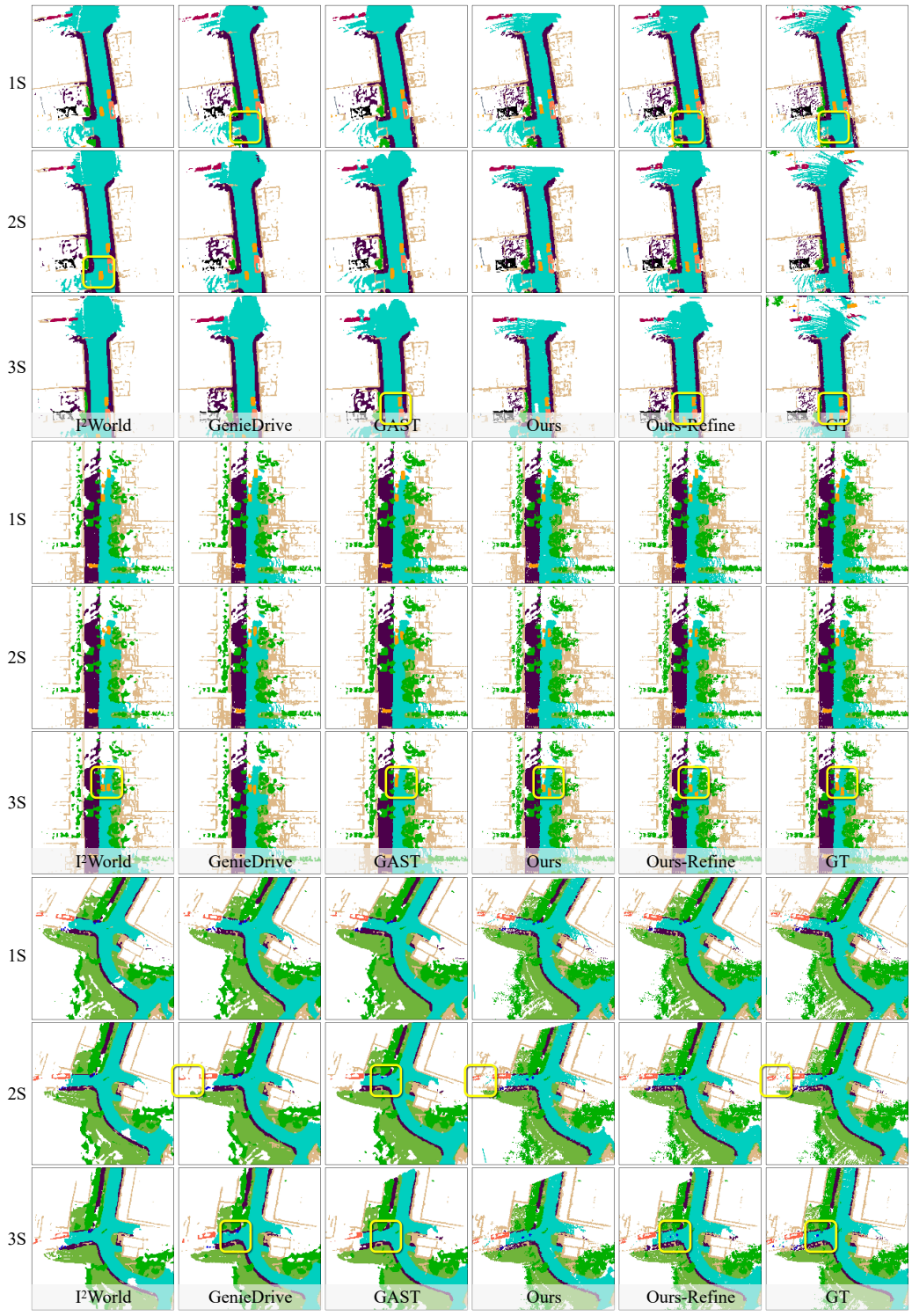


Figure 9: **Qualitative results on Occ3D-nuScenes.** The highlighted boxes show that our method maintains reliable motion estimates.

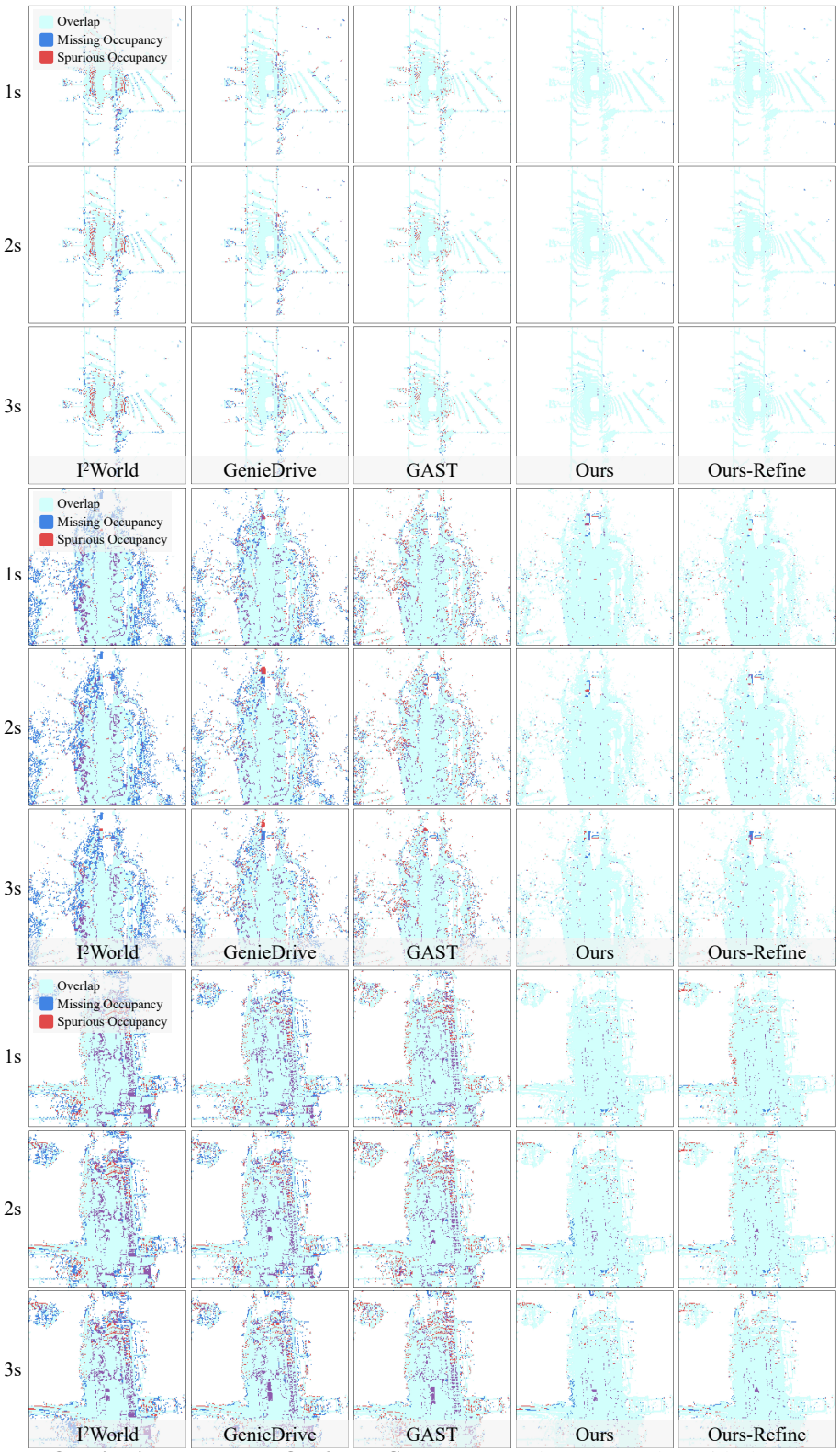


Figure 10: **Qualitative results on Occ3D-nuScenes.** Cyan denotes overlap with the ground truth, whereas blue and red indicate missing and spurious occupancy, respectively.