BEYOND EGO PRIVILEGE: QUERYABLE MULTI-AGENT 4D OCCUPANCY WORLD MODEL

Anonymous authors
Paper under double-blind review

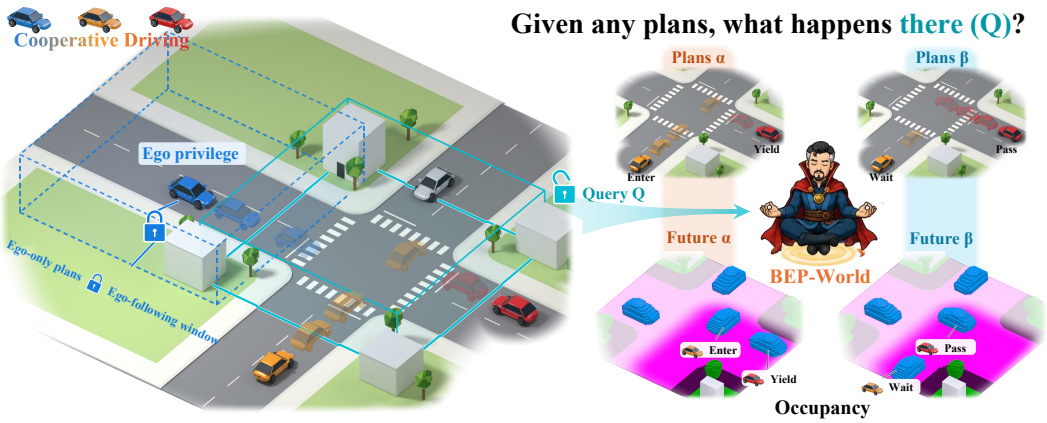

Figure 1: **Beyond Ego Privilege.** BEP-World forecasts occupancy for any of cooperative vehicle plans in the same query window Q (🔓), beyond the dashed ego-following window (🔒). This cooperative-driving schematic fixes the history and query while any plans given. For instance, given Plans-α, the orange vehicle *enters* and the red vehicle *yields* in the future, while Plans-β differ.

ABSTRACT

Ego-conditioned occupancy world models typically tie both motion conditioning and prediction windows to ego. In cooperative driving, comparing plans shared by multiple vehicles requires assessing their joint effects in the same road region, but not in a window that changes with each ego plan. We present **BEP-World** (Beyond Ego Privilege), a queryable 4D occupancy world model with a shared plan interface for ego and surrounding vehicles and separate spatial queries. Besides, we propose Query-Aligned Memory Aggregation and Grounded Plan Fusion to organize historical evidence and vehicle plans in the requested coordinates before scene dynamics, enabling direct forecasting without post-prediction warping. On OpenOccupancy, BEP-World supports forecasting, querying, multi-agent conditioning, and their combination. It predicts occupancy beyond the future ego window and benefits from surrounding-vehicle plans under non-ego queries. Furthermore, we construct BEP-Pairs, a simulation-based paired-intervention development set for diagnosing responses beyond controlled vehicles. Extensive experiments and ablations identify spatial fields as the main source of plan utilization and reveal trade-offs between forecasting accuracy, query coverage, and conditioning gains. Code at https://anonymous.4open.science/r/BEP-World.

1 INTRODUCTION

4D occupancy world models forecast scene evolution through geometry, semantics, and free space (Zheng et al., 2024; Wang et al., 2026; Yang et al., 2026). Many condition prediction on the ego trajectory and also use it to define future output frames (Gu et al., 2024; Yang et al., 2025; 2026). This design suits ego planning but gives ego a privileged role in both motion conditioning and

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.

spatial coverage. As illustrated in Figure 1, a broader interface should accept surrounding-vehicle plans and let users specify where to predict, including regions outside the ego output window.

Consider comparing plans for ego and a merging vehicle while checking occupancy in the same downstream road segment. Such plans may be shared in cooperative driving (Automation, 2020), while other road users' motions remain to be predicted. If the output window follows each candidate ego trajectory, the comparison also changes the region being assessed. Conversely, inspecting another region under fixed plans should not imply different vehicle motions. Plans specify motion conditions, whereas queries specify where predictions are reported. This motivates a joint question: ***given plans for selected vehicles, what will occupy an independently specified spatial window at each future time?***

Conditional trajectory prediction (Ngiam et al., 2021), multi-agent occupancy control (Liang et al., 2026), and spatial querying (Agro et al., 2024; Lilja et al., 2026) provide relevant capabilities. Their joint use raises a common geometric challenge: historical evidence and vehicle plans must refer to the correct physical locations as the query frame changes. Historical observations must be related to the requested region, while each plan must retain its association with a vehicle and its metric path. Moreover, resampling an already cropped ego-window forecast cannot recover predictions outside its spatial support. We address these requirements by organizing history and plans in the requested coordinates before forecasting.

We propose **BEP-World (Beyond Ego Privilege)** with two complementary modules. (1) *Query-Aligned Memory Aggregation* aligns and aggregates historical features for the requested query frame. (2) *Grounded Plan Fusion* expresses selected vehicle trajectories in that frame through transition-aligned spatial fields and agent tokens. The fields supply localized motion cues, while tokens encode trajectory-level features. Together, these modules provide query-aligned evidence and grounded plans to scene dynamics, which predicts occupancy directly in the requested frame. Ego is an optional member of the plan set and shares the same conditioning paths as surrounding vehicles, with no dedicated ego-plan branch.

On OpenOccupancy, a single trained model supports ego-conditioned forecasting, spatial querying, multi-agent conditioning, and their combination. Averaged over three training seeds, factual surrounding-vehicle plans improve vehicle-BEV mIoU by 1.12 percentage points under non-ego queries over matched ego-only conditioning. Ablations identify spatial fields as the main source of plan utilization, with trade-offs across forecasting and query accuracy. Because driving logs provide only factual outcomes, we complement these evaluations with BEP-Pairs, a small paired-intervention development set we construct in CARLA to diagnose responses beyond the directly conditioned vehicles.

Our contributions are summarized below.

- We formulate a new task, *i.e.*, *queryable multi-agent 4D occupancy world model* with a shared plan interface for not only ego but also any surrounding vehicles, separating vehicle-plan conditioning from the choice of output query window.
- We introduce BEP-World framework as an initial solution, which organizes historical spans and spatially grounded vehicle plans in the requested coordinates before scene dynamics to directly forecast occupancy in the query frame.
- We evaluate the joint interface across four settings and isolate the effects of query-conditioned prediction, ego-plan availability, and plan representation. Furthermore, we construct BEP-Pairs to complement factual evaluation with paired intervention diagnostics.

2 Related Work

Occupancy World Models. Occupancy world models forecast scene geometry and semantics on voxel grids. OccWorld (Zheng et al., 2024) jointly predicts occupancy and ego motion, while I^2-World (Liao et al., 2025) separates intra-scene and inter-scene tokenization for efficient forecasting. Diffusion-based models such as OccSora (Wang et al., 2026) and DOME (Gu et al., 2024) generate occupancy sequences with trajectory guidance. Drive-OccWorld (Yang et al., 2025) supports ego trajectories, commands, speed, and steering. GenieDrive (Yang et al., 2026) introduces mutual control attention to fuse driving controls with tri-plane scene dynamics. These ego-conditioned

interfaces center prescribed motion on the ego vehicle. We instead treat ego and surrounding vehicles as optional members of a shared plan set, specified independently of the output query window.

Multi-Agent Conditioning. Conditional trajectory models provide a broader control interface. Scene Transformer (Ngiam et al., 2021) conditions on selected trajectories through masking, while MotionLM (Seff et al., 2023) models joint multi-agent futures. CTG++ (Zhong et al., 2023) and CCDiff (Lin et al., 2025) guide traffic generation with actor-level constraints, predicting actor motion rather than dense occupancy. EOT-WM (Zhu et al., 2026) conditions video generation on ego and surrounding-vehicle trajectories, while UniScene (Li et al., 2025) and AnyScene (Zhang et al., 2026) generate occupancy from prescribed layouts. OccDirector (Liang et al., 2026) addresses multi-agent behavior and interaction in 4D occupancy through language scripts. Studies of action perturbations and causal agents motivate evaluating changed-plan responses separately from factual prediction accuracy (Schofield et al., 2025). Multi-agent control is thus already established across output representations. Our focus is the joint task, *i.e.*, selected vehicles supply identity-linked metric plans, while the dense occupancy forecast is requested in an independently specified spatial window.

Spatial Querying. Scene coordinates and query interfaces offer complementary ways to relax ego-centered prediction. COME (Shi et al., 2026) separates ego motion from scene evolution through scene-centered forecasting. SparseWorld (Dang et al., 2026) adapts its represented range using ego-state-modulated latent queries. UnO (Agro et al., 2024), GASP (Ljungbergh et al., 2026), and QueryOcc (Lilja et al., 2026) learn occupancy fields that can be evaluated at requested coordinates. Such pointwise readout supports flexible spatial sampling, but does not by itself specify how vehicle plans condition the queried future. Our model jointly organizes historical spans and multi-agent plans in the requested coordinates before forecasting occupancy directly in that frame, making spatial querying part of plan-conditioned prediction.

3 METHODOLOGY

3.1 PROBLEM FORMULATION

We define the task as *queryable multi-agent 4D occupancy forecasting*. We observe a history of semantic occupancy grids $\mathcal{H} = \{(\mathcal{O}_t, T_{t\to 0})\}_{t=-T_h+1}^{0}$, where $T_{t\to 0}$ maps each grid's coordinates to the current reference frame. A user selects a subset $\mathcal{S}$ of tracked vehicles and supplies an unordered plan set $\mathcal{A}_{\mathcal{S}} = \{a_i \mid i \in \mathcal{S}\}$. Each plan links a vehicle's current geometry to timed future waypoints. Ego and surrounding vehicles share this multi-agent plan conditioning interface and may be selected individually or together.

A spatial query $\mathcal{Q} = \{q_t\}_{t=1}^{T_f}$ specifies the pose of a fixed-size metric window Ω at each future time. Each $q_t \in \mathrm{SE}(3)$ defines a query frame and maps its coordinates to the current reference frame. The query window is the finite extent Ω in this frame, discretized as an occupancy grid. Query poses specify where the future is reported, without prescribing any vehicle's motion. For fixed history and plans, different queries refer to the same physical future. Predictions should therefore agree at corresponding points in overlapping windows, up to discretization. We restrict evaluated queries to windows supported by the dataset labels.

The task is to predict future occupancy in the requested windows, conditioned on the history and selected plans. The model computes semantic logits and voxelwise predictions as

$$\widehat{\mathcal{L}}^{\mathcal{Q}}_{1:T_f} = F_\theta(\mathcal{H}, \mathcal{A}_{\mathcal{S}}, \mathcal{Q}), \qquad \widehat{\mathcal{O}}^{\mathcal{Q}}_t = \arg\max_c \widehat{\mathcal{L}}^{\mathcal{Q}}_{t,c}. \tag{1}$$

The task requires occupancy consistent with the selected plans, prediction of unconditioned road users' motion, and reconstruction of the surrounding scene within each window. Logged-data evaluation supplies factual plans and measures conditional forecasting accuracy. Evaluating responses to changed plans requires paired outcomes, which we examine separately in simulation.

3.2 BEP-WORLD ARCHITECTURE

Figure 2 is an overview of BEP-World. We build on the tri-plane tokenizer of GenieDrive (Yang et al., 2026). To represent ego as a physical vehicle independently of the query origin, we fill its body with a fixed car-class template in both training inputs and targets. Appendix E motivates this

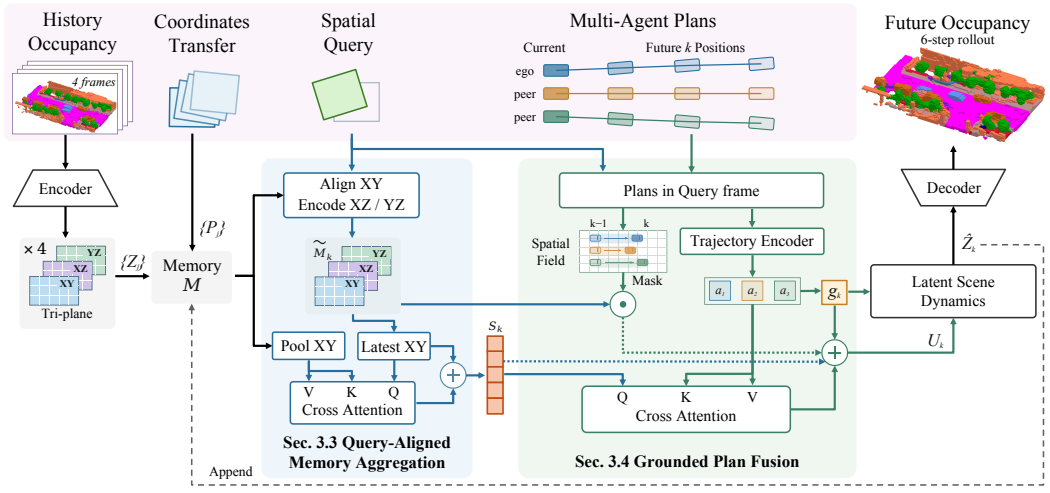

Figure 2: **BEP-World architecture.** Query-Aligned Memory Aggregation (Sec. 3.3) aligns and aggregates tri-plane history in the requested query frame. Grounded Plan Fusion (Sec. 3.4) injects selected vehicle plans in that frame through spatial fields, agent-token attention, and joint plan broadcast before scene dynamics. Predicted features are decoded directly in the query frame and appended to memory for autoregressive rollout in the future occupancy.

representation choice with empty-center and forecast-retention diagnostics, without treating them as causal evidence that filling improves BEP-World.

An encoder E_ϕ maps each observed grid to tri-plane features $Z_j = (Z_j^{xy}, Z_j^{xz}, Z_j^{yz})$. We store them as memory $\mathcal{M} = \{(Z_j, P_j)\}$, where j indexes memory entries and $P_j \in \mathrm{SE}(3)$ maps entry j's coordinates to the reference frame at $t = 0$. At step k, Query-Aligned Memory Aggregation constructs an initial state S_k and aligned memory $\widetilde{\mathcal{M}}_k$. Grounded Plan Fusion produces U_k, and the backbone predicts $\widehat{Z}_k^{\mathcal{Q}} = \Phi_\theta(U_k, g_k, \widetilde{\mathcal{M}}_k)$, where g_k summarizes the selected plans. The decoder returns logits $D_\phi(\widehat{Z}_k^{\mathcal{Q}})$ directly in the requested frame, without a decoded occupancy warp. Appending $(\widehat{Z}_k^{\mathcal{Q}}, q_k)$ to memory yields an autoregressive rollout. Query geometry and plans thus enter through separate modules before scene dynamics.

3.3 Query-Aligned Memory Aggregation

At step k, the query module aligns each memory entry's XY features Z_j^{xy} with the requested frame q_k through

$$R_{k \leftarrow j} = q_k^{-1} P_j, \qquad \widetilde{Z}^{xy}_{j \to k}(\mathbf{u}) = \mathcal{B}\Big(Z_j^{xy}, \pi_{xy}(R^{-1}_{k \leftarrow j} \bar{\mathbf{u}})\Big) + (1 - c_{j,k}(\mathbf{u}))\, e_{\mathrm{unk}}. \tag{2}$$

Here $\bar{\mathbf{u}} = (u_x, u_y, 0, 1)^\top$, π_{xy} selects horizontal coordinates, and $\mathcal{B}$ denotes zero-padded bilinear sampling. Coverage $c_{j,k}$ is obtained by sampling an all-ones grid. The learned feature e_{unk} distinguishes missing historical support from observed free space.

The aligned latest plane then attends to pooled XY features from all memory entries. Keys encode cell positions in q_k and relative time, while values carry scene features. Multi-scale Gaussian distance biases favor nearby evidence while allowing historical information beyond the latest window to contribute. The attention residual augments the initial XY state.

XZ and YZ features aggregate along horizontal axes and do not admit the same pointwise warp. We add an embedding of the relative horizontal rotation and translation to these planes, leaving the backbone to construct their target-frame features. The module returns S_k and $\widetilde{\mathcal{M}}_k$. This coordinate-aware construction does not guarantee cross-query consistency.

3.4 GROUNDED PLAN FUSION

Each plan links a vehicle's current box and class to a future trajectory, yielding boxes $\{b_i^t\}_{t=0}^{T_f}$. Identity is grounded in current geometry without a learned slot index. At step k, all boxes are expressed in q_k.

Agent tokens and joint plan token. A shared encoder embeds transition geometry, displacement, heading change, class, time, and validity. A per-agent temporal transformer combines the current step with the complete supplied trajectory to produce an agent token $a_{i,k}$. Ego shares the encoder and role embedding with surrounding vehicles. A joint plan token g_k projects the validity-masked mean of agent tokens together with the normalized valid-agent count. It is permutation-invariant and zero for an empty set, while individual tokens preserve plan-specific geometry.

Transition-aligned spatial field. For each $k-1 \to k$ transition, we rasterize source and target footprints, displacement, heading change, swept support, and class into a field F_k. Both endpoints share q_k, so displacement describes vehicle motion within one coordinate frame. Oriented Gaussian footprints scale with vehicle dimensions. Max pooling and normalized sums combine vehicles independently of ordering. A convolutional encoder ϕ_F produces a residual restricted to the swept plan-support region M_k. This mask limits direct field injection only, leaving subsequent scene dynamics unrestricted.

Local grounding and global communication. Scene tokens attend to valid agent tokens and a learned null token. XY attention adds a relative-geometry bias that favors nearby paths while retaining access to distant plans. Broadcasting g_k separately makes joint plan information available throughout the scene. Before the first intra block, we fuse

$$U_k = S_k + \alpha_F \, \mathcal{P}_{xy}(M_k \odot \phi_F(F_k)) + \chi_k \left[\tanh(\alpha_A)\, \Delta_k^{\text{agent}} + \tanh(\alpha_J)\, \mathbf{1}\phi_G(g_k)\right], \tag{3}$$

where $\mathcal{P}_{xy}$ inserts the field into XY tokens, Δ_k^{agent} is the attention residual, and the α terms are learned scalar gains. The indicator χ_k marks a nonempty valid plan set. With no plans, all explicit plan-conditioning residuals vanish while the query remains available. The joint plan token also replaces the ego-specific input to the backbone's scene-reading plan stream, and we disable the original ego-pose branch. Spatial and temporal dynamics can therefore propagate plan information beyond the plan-support region.

3.5 LEARNING OBJECTIVES

Training samples empty, ego-only, and random agent subsets from factual sequences, with randomized ordering. Queries include ego windows, translations, and yaw changes within label-supported bounds. Targets are rasterized directly in each requested frame. Training uses native factual positions and headings. Shared XY-plan evaluation retains current dimensions and height but derives headings from waypoint displacements without retraining.

For target $Y_k^{\mathcal{Q}}$, we optimize

$$\mathcal{L}_{\text{train}} = \sum_{k=1}^{T_f} \left[\mathcal{L}_{\text{CE}}(\widehat{\mathcal{L}}_k^{\mathcal{Q}}, Y_k^{\mathcal{Q}}) + \mathcal{L}_{\text{Lovasz}}(\text{softmax}\, \widehat{\mathcal{L}}_k^{\mathcal{Q}}, Y_k^{\mathcal{Q}}) + \lambda_{\text{ctrl}} \mathcal{L}_{\text{CE}}^{M_k^{\uparrow}}\right], \tag{4}$$

with $\lambda_{\text{ctrl}} = 0.5$. The first two terms supervise the full valid volume. The final term is plan-support cross-entropy (Support CE), averaged within $M_k^{\uparrow}$, the plan-support region upsampled and extended through height. It includes background labels, vanishes for empty support, and leaves responses elsewhere unconstrained by any unchanged-scene penalty.

We train with a three-stage strategy. Stage 1: Tokenizer training. We train the tokenizer with occupancy reconstruction supervision to adapt its representation to our data. Stage 2: Latent forecasting. We freeze the tokenizer and train the forecasting model to match encoded future latents. Stage 3: End-to-end optimization. We unfreeze the tokenizer and jointly optimize all modules using the occupancy objective above and scheduled sampling. Inference uses predicted latents throughout rollout. At inference, the model rolls out predicted latents autoregressively. Training budgets and optimization settings are provided in the Appendix A.2.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Data and training. We use OpenOccupancy (Wang et al., 2023) on the nuScenes training and validation splits. Its $102.4 \times 102.4 \times 8$ m label volume supports $64 \times 64 \times 8$ m query windows at 0.4 m resolution. Targets are constructed in the query frames, with a shared physical ego-body template. BEP-World uses 4 occupancy observations and predicts 6 future time steps at 0.5 s intervals.

Evaluation settings. We evaluate spatial querying and surrounding-vehicle plan conditioning separately and jointly. S_0**: Ego-only forecasting** uses ego plans and future ego windows. S_1**: Spatial querying** varies query windows under ego-only conditioning. S_2**: Multi-agent conditioning** adds surrounding-vehicle plans while retaining future ego windows. S_3**: BEP conditioning** combines multi-agent plans with non-ego query windows. For S_2 and S_3, we equally average the scores obtained with up to 1, 3, or 7 surrounding-vehicle plans. All four settings measure factual conditional prediction accuracy. Appendix C.1 reports every seed and sample standard deviations.

Metrics. Occupied IoU measures occupancy geometry, while semantic mIoU measures class-wise prediction accuracy. Full-window (Q) scores cover the entire query window. Novel-region (N) scores cover its portion outside the contemporaneous future ego window, which may still have been observed in the history. Vehicle-BEV scores evaluate height-projected vehicle occupancy across the full window, including unconditioned vehicles and background false positives but excluding the physical ego body. Δ and Δ_A denote the V-mIoU gain, in percentage points, from adding surrounding-vehicle plans to the same model's ego-only prediction under matched inputs, queries, and scoring. Results average the 1/2/3-second horizons, except mIoU@3s, which scores the final prediction horizon. Detailed scoring and input protocols appear in Appendix B.

Baselines. The baseline models include: **OccWorld** (Zheng et al., 2024), **DOME** (Gu et al., 2024), **COME** (Shi et al., 2026), **GenieDrive** (Yang et al., 2026), and **I^2-World** (Liao et al., 2025) assess S_0 forecasting. **UnO** (Agro et al., 2024), **GASP-DINOv3** (Ljungbergh et al., 2026), and **Query-Occ** (Lilja et al., 2026) provide S_1 spatial-query references. **CTG++** (Zhong et al., 2023) and **CCD-iff** (Lin et al., 2025) assess S_2 multi-agent conditioning. **I^2-World + Warp** tests whether transforming an ego-window forecast suffices for spatial querying. **Copy-Paste** shows how far persistence and plan-driven transport can support S_0–S_3 with observed current ground-truth occupancy.

4.2 MAIN RESULT

S_0 **and** S_1**: Forecasting and spatial querying.** In Table 1, BEP-World exceeds I^2-World in ego-only forecasting and slightly surpasses Copy-Paste in average IoU and mIoU, but remains lower in mIoU@3s. Persistence therefore remains a strong long-horizon reference. The distinctive query benefit is spatial coverage: BEP-World achieves 12.38% N-mIoU, whereas I^2-World + Warp scores zero because the novel region lacks source predictions. Direct querying predicts beyond the future ego window rather than merely re-expressing a cropped forecast. Higher N-mIoU but lower Q-mIoU than Copy-Paste shows that this coverage benefit does not eliminate full-window semantic errors.

S_2**: Multi-agent conditioning.** Adding surrounding-vehicle plans improves V-mIoU by 1.07 percentage points over matched ego-only conditioning, averaged across seeds. Each comparison uses the same checkpoint, isolating plan utilization from cross-model differences. Stronger vehicle scores but lower full-scene 3D mIoU than Copy-Paste show that vehicle-level benefits do not imply uniformly better scene prediction. Responses to changed plans are tested in Sec. 4.3.

S_3**: BEP conditioning.** Under non-ego queries, additional plans improve V-mIoU by 1.12 percentage points over matched ego-only conditioning, averaged across seeds. Together with higher N-mIoU than Copy-Paste, this demonstrates plan utilization and novel-region prediction in the same queried outputs, rather than either capability only in isolation. The lower Q-mIoU still indicates a full-scene accuracy trade-off. Sec. 5.1 separates the contributions of query-conditioned prediction and plan conditioning through same-checkpoint inference controls.

Table 1. **Main results across four settings.** BEP-World reports three-seed means, with one checkpoint per seed across settings. Scores (%) average 1/2/3-second horizons unless marked @3s. **Best**/second-best scores are marked within each panel. Copy-Paste denotes copying the current ground-truth occupancy with multi-agent editing as future observations. † denotes adaptation, R denotes shared vehicle renderer.

(a) S_0: `Ego-only forecasting`

Method	IoU ↑	mIoU@3s ↑	mIoU ↑
OccWorld	38.54	11.20	16.93
DOME	40.83	15.50	20.56
COME	50.71	29.06	33.47
GenieDrive	55.77	30.33	37.34
I^2-World	58.55	33.22	39.56
Copy-Paste	59.46	**37.26**	42.85
BEP-World	**59.98**	36.70	**43.13**

(b) S_1: `Spatial querying`

	Full (Q) ↑		Novel (N) ↑	
Method	IoU	mIoU	IoU	mIoU
UnO	37.70	–	21.02	–
GASP-DINOv3†	36.60	–	20.44	–
QueryOcc†	30.94	16.00	21.72	8.45
I^2-World + Warp	54.27	36.06	0.00	0.00
Copy-Paste	55.29	**39.79**	13.07	8.55
BEP-World	**55.49**	38.90	**26.02**	**12.38**

(c) S_2: `Multi-agent conditioning`

	3D ↑	Vehicle BEV ↑		
Method	mIoU	IoU	mIoU	Δ (pp)
CTG++^{R}	–	42.49	45.58	+0.16
CCDiff^R	–	42.24	45.59	+0.06
Copy-Paste	**42.58**	50.17	51.44	+0.50
BEP-World	42.00	**54.06**	**55.77**	**+1.07**

(d) S_3: `BEP conditioning`

	Q ↑	N ↑	Vehicle BEV ↑
Method	mIoU	mIoU	mIoU
Copy-Paste	**39.12**	8.76	46.64
BEP-World	37.75	**11.67**	**49.77**

Q/N: full / outside the future ego window.
Δ: V-mIoU gain over matched ego-only (pp).
S_2, S_3: mean over $k \in \{1,3,7\}$.

4.3 Result on BEP-Pairs

We construct **BEP-Pairs** (Beyond Ego Privilege Pairs), a CARLA (Dosovitskiy et al., 2017) development set pairing reference and plan-intervention rollouts from matched initial states with a shared query. It contains 120 pairs across 57 initial layouts on two road segments in Town10HD_Opt, varying following gaps, speeds, and acceleration/deceleration plans. Half supply only the ego plan ($K = 1$), and half also supply one surrounding-vehicle plan ($K = 2$). Each pair has a repeated-reference control. We evaluate the three frozen model seeds with pilot-selected preprocessing and no CARLA training. Appendix D details construction and scoring.

Table 2 compares each seed with two matched inference controls. Ego-only hides surrounding-vehicle plans while retaining history, ego plan, and query. Reference + local edit transports that seed's predicted vehicle voxels according to changed plans and leaves the remaining forecast unchanged. Signed-change F1 and recall score vehicle additions and removals against paired simulator outcomes. The stable false-change rate, sFCR, scores semantic changes where paired GT is unchanged, equally weighting vehicle, other-occupied, and free-space strata. Low sFCR alone cannot distinguish accurate dynamics from a plan-insensitive forecast.

Table 2. **BEP-Pairs development diagnostic on 120 pairs.** Three-seed means (%) use common-valid 3D surface annotations, equally averaging 57 layouts and 1/2/3-second horizons within each seed. Ctrl. denotes supplied-plan vehicles, Resp. followers of intervened vehicles, and Rec. signed-change recall. **Best**/second-best scores are marked per metric.

	Occupancy prediction ↑			Signed vehicle change ↑				GT-stable
Method	Scene mIoU	Ctrl. IoU	Resp. IoU	All F1	Ctrl. F1	Resp. F1	Resp. Rec.	sFCR ↓
BEP-World (ego-only)	45.72	55.16	47.17	30.75	40.69	10.76	9.70	0.89
Reference + local edit	**45.91**	**61.01**	46.17	**39.85**	**51.09**	0.00	0.00	**0.25**
BEP-World	45.87	61.00	**47.76**	36.78	48.17	**12.01**	**11.05**	0.96

Local editing leads in whole-scene and controlled-change F1 but produces no responder changes, separating geometric transport from interaction prediction. The full model reaches 12.01 responder-change F1, yet its 11.05% recall indicates that most response changes remain missed. Full and ego-only inputs coincide at $K = 1$, so $K = 2$ isolates the additional plan: controlled IoU rises from 48.90 to 60.39 and responder-change F1 from 8.65 to 11.12, with positive gains in all three seeds. However, sFCR increases from 0.76 to 0.90. The evidence supports plan utilization and partial follower responses, not reliable interaction recovery or uniform superiority to local editing.

5 Analysis and Ablation

5.1 Separating Spatial Queries from Motion Conditioning

Under the S_3 protocol, Table 3 compares query-conditioned prediction with post-prediction warp under matched plans. Warp scores unsupported cells as free, so its zero N-mIoU reflects missing spatial support, not insensitivity to motion. Query-conditioned prediction extends coverage while retaining a plan-conditioning gain, but its Q-mIoU is 0.81 percentage points below warp with multi-agent plans. Unlike resampling one forecast, it re-aligns memory and reruns dynamics, allowing frame-dependent errors. The comparison supports separately controllable query and plan inputs, not guaranteed coordinate-invariant predictions.

Table 3. **Prediction path × plans.** Three-seed means of matched-checkpoint controls. Query denotes query-conditioned prediction. Δ_A is the within-path V-mIoU gain. **Best**/second-best scores are marked per metric.

	Ego only		Ego + k	
Metric ↑	Warp	Query	Warp	Query
Q-mIoU	38.04	37.16	**38.56**	37.75
N-mIoU	0.00	11.23	0.00	**11.67**
V-mIoU	49.39	48.65	**50.26**	49.77
Δ_A (pp)	–	–	+0.87	**+1.12**

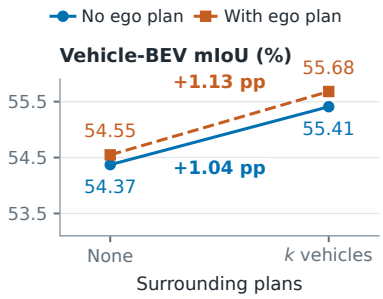

Figure 3: **Ego-plan availability.** Q = ego, and $k \in \{1,3,7\}$ averaged.

5.2 Ego as an Optional Plan-Set Member

Figure 3 shows V-mIoU gains of 1.04 percentage points without an ego plan and 1.13 with it. Surrounding-plan benefits therefore do not require an explicit ego entry. This establishes optional ego-plan conditioning, not removal of ego information: Q = ego still supplies future coordinates, and V-mIoU excludes the physical ego body. The small ego-plan benefit does not imply that ego motion is unimportant. Full scores appear in Appendix C.4.

Table 4. **Plan representation and support supervision.** Matched-budget retraining with seed 17. Δ_A is the V-mIoU gain over matched ego-only conditioning. **Best**/second-best scores are marked.

	Components			S_0	S_1	S_2: Multi-agent	
Variant	Field	Tokens	Support CE	mIoU ↑	N-mIoU ↑	V-mIoU ↑	Δ_A (pp) ↑
Token-Only	✗	✓	✓	42.46	12.47	55.05	+0.13
Field-Only	✓	✗	✓	**42.63**	11.62	**56.06**	+1.11
No Extra Support CE	✓	✓	✗	39.68	**12.52**	55.85	**+1.45**
BEP-World (Full)	✓	✓	✓	42.35	12.46	55.71	+1.27

5.3 Spatial Grounding of Multi-Agent Plans

Table 4 separates forecasting accuracy from plan utilization. Token-only retains strong S_0 accuracy but gains only 0.13 percentage points from surrounding plans, versus 1.11 for Field-only, supporting

spatial grounding as the main plan-utilization path. Adding tokens to Field-only raises N-mIoU from 11.62 to 12.46, but slightly lowers S_0 and vehicle-BEV mIoU. Removing Support CE raises vehicle scores and conditioning gain but reduces full-scene accuracy. These single-seed results support complementary roles and a capability trade-off, not uniform superiority or interaction fidelity.

5.4 Qualitative Analysis

The Visualizations shown in Figure 4. The additional plans improve local car IoU for both a selected vehicle and an unconditioned one under a shifted query. Local transport leaves the follower forecast unchanged, whereas BEP-World shifts it toward the braking ground truth. This illustrates plan influence beyond controlled occupancy, although residual geometry errors indicate only partial interaction recovery.

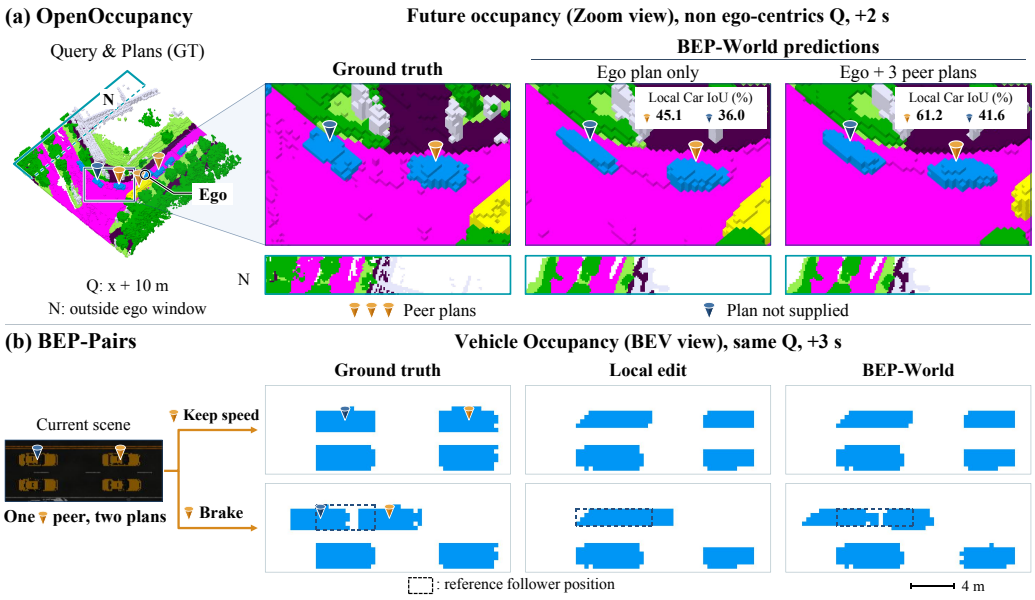

Figure 4: **Visualizations.** (a) OpenOccupancy at $+2$ s: ego-only versus three additional factual plans with a 10 m forward query. Orange / blue markers denote selected / unconditioned vehicles, N the region outside the future ego window. Car IoUs are local. (b) BEP-Pairs at $+3$ s: keep-speed versus braking plans under a fixed query. Dashed boxes mark reference follower positions.

6 Conclusion

We formulated *queryable multi-agent 4D occupancy world model*, a task that asks what will occupy independently specified spatial windows under plans supplied for selected vehicles. As an initial solution, **BEP-World** shares a plan interface between ego and surrounding vehicles and organizes historical evidence and plans in query coordinates before scene dynamics. Experiments demonstrate forecasting, querying, multi-agent conditioning, and their combination within one model. Ablations identify spatial fields as the main source of plan utilization and reveal trade-offs among forecasting accuracy, query coverage, and conditioning gains. BEP-Pairs further shows partial responses of unconditioned vehicles to changed plans. Together, the task and framework provide a basis for assessing multi-vehicle plans in common spatial windows, while cross-query consistency and reliable interaction prediction remain open challenges.

Limitation and future work. Current supervision focuses on factual futures. Query-specific occupancy targets also lack explicit constraints enforcing consistent predictions of the same event across different spatial windows. Future work will develop richer supervision covering diverse plausible futures and paired plan interventions, together with cross-query consistency constraints at shared physical locations. These signals could better distinguish changes caused by vehicle plans from changes in the observation window, supporting more reliable joint prediction.

AI USE STATEMENT

We used generative AI tools, including OpenAI Codex, to assist with drafting and polishing the manuscript, identifying and exploring related work, and research ideation and execution. Research assistance included assisting coding, planning experiments and debugging code. AI tools also assisted with implementing the CARLA data-generation and evaluation pipelines. The authors take full responsibility for the final manuscript, implementation, and reported results.

ETHICS STATEMENT

This work uses public driving datasets and CARLA simulations to study conditional occupancy forecasting. Predicted responses to changed plans may be inaccurate and should not be interpreted as safety guarantees.

REPRODUCIBILITY STATEMENT

Section 3 describes the model and objectives. Appendix A specifies initialization and training, Appendix B fixes the scoring protocol, and Appendix C reports complete per-seed results and matched ablations. Appendix D documents the paired-intervention audits. Code has been released to support reproduction at: https://anonymous.4open.science/r/BEP-World.

REFERENCES

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

Cooperative Driving Automation. *Taxonomy and definitions for terms related to cooperative driving automation for on-road motor vehicles*. SAE International, 2020.

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*, volume 40, pp. 3497–3505, 2026.

Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pp. 1–16. PMLR, 2017.

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.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11971–11981. IEEE, 2025.

Zhuding Liang, Tianyi Yan, Dubing Chen, Jiasen Zheng, Huan Zheng, Cheng-zhong Xu, Yida Wang, Kun Zhan, and Jianbing Shen. Occdirector: Language-guided behavior and interaction generation in 4d occupancy space. In *European Conference on Computer Vision*, pp. 37–54. Springer, 2026.

Zhimin Liao, Ping Wei, Ruijie Zhang, Shuaijia Chen, Haoxuan Wang, and Ziyang Ren. I^2-world: Intra-inter tokenization for efficient dynamic 4d scene forecasting. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 25810–25819. IEEE, 2025.

Adam Lilja, Ji Lan, Junsheng Fu, and Lars Hammarstrand. Queryocc: Query-based self-supervision for 3d semantic occupancy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21399–21408, 2026.

Haohong Lin, Xin Huang, Tung Phan, David Hayden, Huan Zhang, Ding Zhao, Siddhartha Srinivasa, Eric Wolff, and Hongge Chen. Causal composition diffusion model for closed-loop traffic generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27542–27552. IEEE, 2025.

William Ljungbergh, Adam Lilja, Adam Tonderski, Arvid Laveno Ling, Carl Lindström, Willem Verbeke, Junsheng Fu, Christoffer Petersson, Lars Hammarstrand, and Michael Felsberg. Gasp: Unifying geometric and semantic self-supervised pre-training for autonomous driving. In *WACV*, pp. 3077–3087. IEEE, 2026.

Jiquan Ngiam, Benjamin Caine, Vijay Vasudevan, Zhengdong Zhang, Hao-Tien Lewis Chiang, Jeffrey Ling, Rebecca Roelofs, Alex Bewley, Chenxi Liu, Ashish Venugopal, et al. Scene transformer: A unified architecture for predicting multiple agent trajectories. *arXiv preprint arXiv:2106.08417*, 2021.

Hunter Schofield, Mohammed Elmahgiubi, Kasra Rezaee, and Jinjun Shan. Beyond simulation: Benchmarking world models for planning and causality in autonomous driving. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1308–1314. IEEE, 2025.

Ari Seff, Brian Cera, Dian Chen, Mason Ng, Aurick Zhou, Nigamaa Nayakanti, Khaled S Refaat, Rami Al-Rfou, and Benjamin Sapp. Motionlm: Multi-agent motion forecasting as language modeling. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8545–8556. IEEE, 2023.

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*, 38:70–96, 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*, 36:64318–64330, 2023.

Lening Wang, Wenzhao Zheng, Yilong Ren, Han Jiang, Zhiyong Cui, Haiyang Yu, and Jiwen Lu. Occsora: 4d occupancy generation models as world simulators for autonomous driving. *IEEE Transactions on Image Processing*, 2026.

Xiaofeng Wang, Zheng Zhu, Wenbo Xu, Yunpeng Zhang, Yi Wei, Xu Chi, Yun Ye, Dalong Du, Jiwen Lu, and Xingang Wang. Openoccupancy: A large scale benchmark for surrounding semantic occupancy perception. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17804–17813. IEEE, 2023.

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*, volume 39, 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 35680–35690, 2026.

Haiming Zhang, Junfei Zhou, Feng Jiang, Jingzhong Li, Zhenglong Guo, Penglin Dai, Jifeng Dai, Yan Xie, and Benjin Zhu. Anyscene: Towards highly controllable driving scene generation at anywhere and beyond. *arXiv preprint arXiv:2605.26113*, 2026.

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

Ziyuan Zhong, Davis Rempe, Yuxiao Chen, Boris Ivanovic, Yulong Cao, Danfei Xu, Marco Pavone, and Baishakhi Ray. Language-guided traffic simulation via scene-level diffusion. In *Conference on robot learning*, pp. 144–177. PMLR, 2023.

Jian Zhu, Zhengyu Jia, Tian Gao, Jiaxin Deng, Shidi Li, Lang Zhang, Fu Liu, Peng Jia, and Xianpeng Lang. Other vehicle trajectories are also needed: A driving world model unifies ego-other vehicle trajectories in video latent space. In *AAAI*, volume 40, pp. 13934–13942, 2026.

A Implementation and Training

This appendix specifies the implementation behind Section 3, the scope of the repeated runs, and the provenance of the reported checkpoints. Appendix B fixes the evaluation contract; Appendix C expands the quantitative evidence; Appendices D and E document the intervention and representation diagnostics, respectively.

A.1 Representation and Conditioning

Occupancy and memory. The $64 \times 64 \times 8$ m window is discretized into $160 \times 160 \times 20$ voxels at 0.4 m resolution. The tokenizer produces 40×40 XY and 40×5 XZ/YZ planes with 64 channels, giving 2,000 scene tokens. Four observations initialize a nine-entry memory, padded by repeating the earliest observation. Each autoregressive step appends a latent and its query pose and discards the oldest entry. Five intra blocks perform spatial and plan-aware processing; five inter blocks incorporate temporal memory. The output is decoded directly in the requested coordinates.

Query-Aligned Memory Aggregation. Each memory XY plane is pooled to 10×10 keys. Their positions are transformed into the query frame, with relative-time embeddings and four Gaussian distance-bias scales of 3/6/12/24 m, clipped below at -20. Bilinear XY sampling includes a coverage mask and a learned unknown feature. XZ/YZ planes instead receive an embedding of relative horizontal rotation and translation; this is not an exact rigid warp of the full 3D representation.

Plan interface. We accept at most eight vehicles, including ego. Each token encodes current geometry, trajectory geometry, class, time, and validity through a shared per-agent encoder and temporal transformer. Ego and surrounding vehicles use the same role embedding. Invalid tokens are masked. The joint token projects the validity-masked mean together with the valid-agent count divided by eight, and is zero for an empty set. XY scene-to-agent attention uses relative geometry with distance scales of 6/12/20/40 m, clipped below at -4; vertical-plane attention is content-based. The original ego-pose branch is disabled. In shared-XY evaluation, current dimensions and height are retained and headings are derived from waypoint displacement. Native factual training also supplies headings.

Table 5. **Transition-field channels.** Source and target boxes are both expressed in the current requested frame. The field has 17 channels on the 40×40 XY grid.

Channels	Quantity	Construction
1–4	Source/target footprints	Binary and oriented Gaussian footprints; max across agents.
5–8	Forward/backward displacement	XY displacement divided by the 64 m span; splatted at source / target.
9	Displacement magnitude	Metric length clipped at 80 m and divided by 80.
10–11	Heading change	Sine and cosine, splatted at the source.
12	Swept support	Maximum Gaussian footprint over both endpoints and three interpolated boxes.
13–17	Vehicle class	Five class indicators, splatted at the source.

Spatial field and missing conditions. Oriented Gaussian scales are 0.35 times the box length and width. Attribute splats sum Gaussian-weighted agent values and divide by the summed weights, clamped below at one. A 3×3 convolution followed by a 1×1 convolution maps the field to 64 channels. Thresholding swept support at 0.05 and dilating by one latent cell defines the direct-injection mask. Outside this mask, the field residual is zero, but the scene is neither masked out nor forced unchanged: token attention and subsequent dynamics can propagate effects there. An absent plan means unspecified behavior, not zero motion or free space. An empty plan set removes all explicit plan residuals while retaining the history and query inputs.

A.2 Three-Stage Optimization

The additional optimization budgets are listed in Table 6.

Table 6. **Training recipe for the reported model.** Batch sizes are global across eight A100 GPUs. A new optimizer is initialized at each stage boundary; interrupted runs within a stage restore full training state.

Stage	Updates	Batch	Optimizer / initial LR	Objective
Tokenizer continuation	92,400	64	Adam / 10^{-4}	CE + Lovasz + 0.005 normalized KL
Frozen-tokenizer predictor	47,976	64	AdamW / 10^{-4}	Latent feature matching
End-to-end forecasting	128,088	24	AdamW / 5×10^{-5}	Full-volume CE + Lovasz + 0.5 Support CE

Tokenizer continuation uses dropout 0.5, a fixed learning rate, and mixed precision with an FP32 classification head. Predictor and end-to-end training use weight decay 0.01, gradient-norm clipping at 5, and a step learning-rate schedule with the decay boundary at five-sixths of the stage budget (39,980 and 106,740 updates). The frozen stage fits the predictor to the continued tokenizer's latent representation; the final stage unfreezes the tokenizer and optimizes the decoded occupancy objective in Section 3.5. Support CE is inactive in the frozen stage. Scheduled sampling mixes encoded factual and predicted latents during training; inference always rolls out predicted latents. Cached memory is detached, while the current encoded state can retain encoder gradients.

Training distribution. Tokenizer crops and observed dynamics histories are ego-centered in the reported model. Half the sampling slots draw from all training frames, and half from frames with a supported surrounding-vehicle anchor, but the latter are also rendered in ego coordinates. Thus, the sampler oversamples anchor-eligible scenes; it does not add anchor-centered historical inputs or double the update budget. There are 28,130 training frames, of which 19,320 are anchor-eligible. Anchor eligibility uses current annotations and observed-history support with a 2 m margin, without selecting by future motion. Future queries and plan subsets are randomized as described in Section 3.5.

Checkpoint and repeat scope. Seeds 17, 29, and 43 share tokenizer continuation and the predictor initialization, then repeat the frozen and end-to-end stages. We use each final 128,088-update EMA checkpoint for all evaluation settings, with no validation-best selection. Standard deviations therefore describe downstream training variation conditional on common pretraining, not independent end-to-end pretraining or test-set confidence intervals. Module ablations use matched seed-17 budgets and initialization. Optimizer, EMA, random-number state, and consumed-data position are restored after interruption; migrations do not restart a stage's update budget.

A.3 What the Objective Does and Does Not Enforce

Full-volume CE and Lovasz supervise factual future occupancy in sampled query frames. Additional Support CE emphasizes the selected-plan support, including its background labels, without imposing an unchanged-scene penalty outside it. Training does not use paired simulator interventions, a causal-effect loss, or a cross-query overlap-consistency loss. Separating the query and plan interfaces permits independent input changes, but does not mathematically enforce identical predictions at overlapping physical locations across separately evaluated queries. Query accuracy and direct-versus-warp comparisons test useful readout capability, not that stronger consistency property.

B Evaluation Protocol

B.1 Prediction Origins, Queries, and Plans

Prediction origins. A prediction origin is the current frame from which a future rollout starts. We use 4,519 common origins for S_0/S_2 and 4,117 for S_1/S_3, keeping the scored origins fixed across methods within each setting. The two groups therefore do not use identical sample sets.

Query set. Queries are defined relative to the contemporaneous future ego frame and remain within label support. The eleven queries comprise the ego window, forward offsets of 5/10/15 m, left/right offsets of 10 m, yaw offsets of $\pm30^\circ$ and $\pm45^\circ$, and a time-varying forward offset of 0/3/6/9/12/15 m

over the six future steps. The last query changes its offset over time, unlike the constant-offset queries. S_1 averages full-window scores over all eleven queries and novel-region scores over the ten non-ego queries. S_3 uses the ten non-ego queries for all reported metrics.

Plan selection and averaging. For S_2/S_3, we retain the ego plan and add up to $k \in \{1,3,7\}$ eligible surrounding-vehicle plans. Vehicles are ranked by current distance, and the three limits use nested subsets of the same ordered list. These lists remain fixed across queries. Scenes with fewer than k eligible vehicles use the available plans and remain in the evaluation. For each query, horizon, and plan limit, confusion counts are pooled across origins before computing scores. Reported results then equally average scores over the 1/2/3-second horizons, the relevant queries, and the three plan limits. Counts are not pooled across plan limits, and results are not weighted by the number of available vehicles.

B.2 Metrics and Aggregation

Metric computation. For semantic class c, $\mathrm{IoU}_c = \mathrm{TP}_c/(\mathrm{TP}_c + \mathrm{FP}_c + \mathrm{FN}_c)$. Semantic mIoU averages non-free classes whose accumulated union is nonzero; a class with nonzero union and zero intersection retains a score of zero. Occupied IoU first merges all non-free classes. Q metrics use the full query window. N metrics restrict scoring to its part outside the contemporaneous future ego window, not necessarily outside the observed history. The masks are geometric and shared across methods, rather than derived from their predicted support.

Vehicle-BEV scores height-project car, bus, truck, trailer, and construction-vehicle labels. V-IoU merges these five classes; V-mIoU averages their class-wise IoUs with nonzero union. These are not vehicle/background two-class mIoUs. The full query window is scored, including unconditioned vehicles and background false positives, with the physical ego region excluded.

Order of aggregation. For query q, plan limit k, and horizon h, let C_{qkh} be the confusion counts summed over common prediction origins, and let $f(C)$ be the relevant score. We report

$$\overline{m} = \frac{1}{|\mathcal{Q}|\,|\mathcal{K}|\,3} \sum_{q\in\mathcal{Q}} \sum_{k\in\mathcal{K}} \sum_{h\in\{1,2,3\}\,\mathrm{s}} f(C_{qkh}). \tag{5}$$

For S_0/S_1, the plan-limit set has one element; for S_0/S_2, the query set has one element. S_1 uses different query sets for Q and N, as specified above. Multi-seed reporting then averages per-seed scores, not their confusion counts. This order prevents frequent queries or plan limits from receiving additional weight.

Input protocols and conditioning gains. Our S_0/S_1 evaluations retain native ego payloads, whereas S_2/S_3 standardize plans to XY waypoints and current geometry. To compute Δ or Δ_A, we compare multi-agent and ego-only predictions from the same checkpoint, holding the ego input format, query, prediction origins, targets, and scoring masks fixed. The gain is the difference in vehicle-BEV mIoU, reported in percentage points and averaged over the matched query, horizon, and plan-limit settings. It is not obtained by subtracting the S_0/S_1 scores in the main table from S_2/S_3 scores.

Baseline inputs. UnO and GASP-DINOv3 use LiDAR, QueryOcc uses images, and CTG++/CCDiff use actor histories and maps. CTG++/CCDiff predictions pass through a shared vehicle renderer and are evaluated only on vehicle-BEV metrics, not full-scene 3D occupancy.

Warp and Copy-Paste. I^2-World + Warp resamples an ego-window forecast into the requested query frame. Unsupported cells are predicted free and remain scored. Copy-Paste rigidly moves current vehicle voxels along supplied plans, clears vacated voxels to free, and preserves other content in world coordinates. It reads out the requested window without an intermediate future-ego crop. No future occupancy labels are used, and unsupported predictions remain free and scored.

Comparison scope. Shared targets, origins, and scoring do not imply equal input information, training data, or optimization lineage across all baselines. Comparisons within a checkpoint isolate plan availability or readout; matched retrained ablations isolate the specified component under the common recipe. Neither should be conflated with a cross-method causal comparison. Unsupported predictions remain scored: a model is not rewarded for abstaining in difficult spatial regions.

C Complete Results and Matched Ablations

C.1 Main Model: All Seeds and Horizons

Tables 7–9 report each final checkpoint behind the main-paper averages. They retain all main-table metrics, expand S_0 by horizon, and include additional S_3 IoUs and matched conditioning gains. All scores average 1/2/3-second horizons unless a particular horizon is shown. Sample standard deviations use $n - 1 = 2$ in the denominator and are computed from unrounded scores. They are not scene-bootstrap confidence intervals, and we do not infer significance from three runs sharing pretraining.

Table 7. S_0 **per-seed results (%).** Mean and sample SD use unrounded scores.

	Occupied IoU				Semantic mIoU			
Seed	1 s	2 s	3 s	Mean	1 s	2 s	3 s	Mean
17	64.27	59.06	54.71	59.34	49.67	41.45	35.92	42.35
29	64.84	59.71	55.29	59.95	50.38	42.17	36.70	43.08
43	65.62	60.40	55.91	60.64	51.34	43.05	37.48	43.96
Mean	64.91	59.72	55.30	59.98	50.46	42.22	36.70	43.13
SD	0.68	0.67	0.60	0.65	0.84	0.80	0.78	0.81

Table 8. S_1 **and** S_2 **per-seed results (%).** Q/N scores use the S_1 query protocol; S_2 keeps the future ego window. Δ_A is in percentage points.

	S_1: Query				S_2: Multi-agent			
Seed	Q-IoU	Q-mIoU	N-IoU	N-mIoU	3D mIoU	V-IoU	V-mIoU	Δ_A
17	55.20	38.50	26.11	12.46	41.28	53.71	55.71	1.27
29	55.44	38.84	25.91	12.33	41.86	53.92	55.55	0.85
43	55.83	39.36	26.05	12.34	42.85	54.56	56.06	1.09
Mean	55.49	38.90	26.02	12.38	42.00	54.06	55.77	1.07
SD	0.32	0.44	0.10	0.07	0.79	0.44	0.26	0.21

Table 9. S_3 **per-seed results (%).** Mean and sample SD use unrounded scores.

Seed	Q-IoU	Q-mIoU	N-IoU	N-mIoU	V-IoU	V-mIoU	Δ_A (pp)
17	52.77	37.39	23.27	11.66	48.43	49.88	1.27
29	53.08	37.62	23.36	11.65	48.46	49.54	0.95
43	53.55	38.23	23.60	11.70	48.82	49.88	1.14
Mean	53.13	37.75	23.41	11.67	48.57	49.77	1.12
SD	0.39	0.43	0.17	0.03	0.22	0.20	0.16

Both S_2 and S_3 have positive matched V-mIoU gains in every seed. However, the mean S_0 advantage over Copy-Paste is smaller than the downstream seed variation: the seed-17 score is below that reference, while seeds 29 and 43 are above it. The evidence supports useful plan conditioning and querying across runs, not uniform forecasting dominance. Novel-region prediction is also distinct from prediction of previously unseen space, since historical observations can cover that region.

C.2 Retrained Module Ablations

The variants in Table 4 share tokenizer pretraining and predictor initialization, followed by 47,976 frozen-tokenizer updates and 128,088 end-to-end updates with matched data/task sampling, optimizer settings, and final EMA selection. Architectural ablations apply in both stages. No Extra Support CE reuses Full's frozen stage because that loss is inactive there. Query geometry, history attention, and scene dynamics are unchanged. Tokens comprise both the agent-token and joint-plan-token paths. Removing them leaves the spatial field and global scene dynamics intact. Support CE is the extra plan-support cross-entropy with $\lambda_{\text{ctrl}} = 0.5$. Removing it retains full-volume CE and Lovasz.

Evaluation follows Table 1, averaging 1/2/3-second horizons. S_2/S_3 equally average surrounding-vehicle plan limits $k \in \{1,3,7\}$, and each conditioning gain uses the same variant's matched ego-only run. All variants, including Full, use seed 17 here, whereas the main results average three seeds. Small ablation differences are not accompanied by a multi-seed significance estimate.

Table 10. **Module-ablation summary for S_0–S_2.** Components follow Table 4. Δ_A is the V-mIoU gain over matched ego-only conditioning. **Best**/second-best scores are marked per metric.

	S_0	S_1: Query		S_2: Multi-agent		
Variant	mIoU ↑	Q-mIoU ↑	N-mIoU ↑	3D mIoU ↑	V-mIoU ↑	Δ_A (pp) ↑
Token-Only	42.46	38.48	12.47	41.23	55.05	+0.13
Field-Only	**42.63**	38.42	11.62	**42.77**	**56.06**	+1.11
No Extra Support CE	39.68	36.10	**12.52**	38.64	55.85	**+1.45**
BEP-World (Full)	42.35	**38.50**	12.46	41.28	55.71	+1.27

Table 11. **Additional S_0–S_2 ablation metrics.** All models use seed 17. Together with Table 10, this reports the main metrics and all three S_0 semantic horizons.

	S_0				S_1		S_2
Variant	mIoU 1 s	2 s	3 s	IoU	Q-IoU	N-IoU	V-IoU
Token-Only	49.63	41.62	36.14	59.33	55.08	26.14	51.41
Field-Only	49.94	41.76	36.18	58.14	53.79	23.52	54.54
No Extra Support CE	46.61	38.80	33.63	59.20	55.10	26.22	53.74
Full	49.67	41.45	35.92	59.34	55.20	26.11	53.71

Table 12. **Complete S_3 ablation results.** Same seed-17 models, ten non-ego queries, and equal weighting of surrounding-plan limits 1/3/7. Each gain uses that variant's matched ego-only prediction.

Variant	Q-IoU	Q-mIoU	N-IoU	N-mIoU	V-IoU	V-mIoU	Δ_A (pp)
Token-Only	52.36	37.12	23.06	11.45	45.61	48.75	0.07
Field-Only	53.34	38.13	23.44	11.73	48.77	49.79	1.23
No Extra Support CE	52.63	35.05	23.32	11.73	48.42	49.99	1.46
Full	52.77	37.39	23.27	11.66	48.43	49.88	1.27

What the complete ablation supports. Removing the field nearly eliminates matched plan gains in both S_2 and S_3, whereas Field-Only retains most of Full's gains. Token fusion improves S_1 novel-region mIoU over Field-Only, but does not improve every forecasting or query metric: Field-Only has higher S_0 mIoU and S_3 Q-mIoU. Support CE improves full-scene forecasting relative to its removal, while the model without it has slightly larger conditioning gains and S_3 V-mIoU. The components therefore produce measurable trade-offs rather than monotonic improvements in all columns. These controlled removals do not separately identify the agent-token and joint-token contributions.

C.3 Same-Checkpoint Query and Plan Controls

Table 13 expands the query/plan comparison in the main text. Direct predicts in the requested frame; Warp first predicts in the future ego window and resamples its output. Both use the same model weights. Positive N-mIoU under Direct demonstrates output beyond the future ego crop, while Warp's zero reflects missing source support. Direct's lower full-window Q-mIoU remains visible for each seed. This is an inference/readout comparison, not a separately trained query-module removal and not a test of overlap consistency across independently queried predictions.

Table 13. **Query/plan inference controls, by seed.** Warp and Direct use the same checkpoint, history, and plans. Multi denotes the average over surrounding-plan limits 1/3/7. Δ_A compares Multi against Ego within each readout and seed; SD is computed on these paired gains.

Seed	Readout	Plans	Q-mIoU	N-mIoU	V-mIoU	Δ_A (pp)
17	Warp	Ego	37.35	0.00	49.27	–
17	Direct	Ego	36.78	11.30	48.61	–
17	Warp	Multi	37.98	0.00	50.30	1.03
17	Direct	Multi	37.39	11.66	49.88	1.27
29	Warp	Ego	38.04	0.00	49.39	–
29	Direct	Ego	37.10	11.20	48.59	–
29	Warp	Multi	38.48	0.00	50.07	0.68
29	Direct	Multi	37.62	11.65	49.54	0.95
43	Warp	Ego	38.73	0.00	49.53	–
43	Direct	Ego	37.61	11.20	48.73	–
43	Warp	Multi	39.22	0.00	50.42	0.89
43	Direct	Multi	38.23	11.70	49.88	1.14
Mean	Warp	Ego	38.04	0.00	49.39	–
Mean	Direct	Ego	37.16	11.23	48.65	–
Mean	Warp	Multi	38.56	0.00	50.26	0.87
Mean	Direct	Multi	37.75	11.67	49.77	1.12
SD	Warp	Ego	0.69	0.00	0.13	–
SD	Direct	Ego	0.42	0.06	0.08	–
SD	Warp	Multi	0.63	0.00	0.18	0.18
SD	Direct	Multi	0.43	0.03	0.20	0.16

C.4 Ego-Plan Availability

Figure 3 uses the final seed-17 checkpoint on 4,117 anchors with the shared XY-plan interface and fixed $Q = \text{ego}$. Scores average 1/2/3-second horizons, with surrounding-vehicle plan limits $k \in \{1,3,7\}$ equally weighted. Dropping ego does not replace it with another vehicle. These are matched input-availability controls, not separately trained models.

Table 14. **Full ego-plan availability scores.** Δ_A compares surrounding plans against no surrounding plans at fixed ego-plan availability. **Best**/second-best scores are marked per metric.

Supplied plans	3D mIoU ↑	V-mIoU ↑	Δ_A (pp) ↑
None	41.07	54.37	–
Ego only	40.53	54.55	–
Surrounding only	**41.36**	55.41	+1.04
Ego + Surrounding	41.28	**55.68**	**+1.13**

D BEP-PAIRS: PROTOCOL, AUDITS, AND FULL RESULTS

D.1 CONSTRUCTION AND FIXED INFERENCE PROTOCOL

Sixty initial configurations on two road segments in Town10HD_Opt each receive two plan interventions, yielding 120 reference/intervention pairs. Following gaps range from 6 to 20 m and initial desired speeds from 4 to 8 m/s, with acceleration and deceleration interventions. There are 60 pairs each with $K = 1$ and $K = 2$ supplied plans, where K includes ego. Each pair has matched pre-intervention history and a repeated-reference rollout. Physical matching groups the configurations into 57 distinct layouts. Samples are not selected by collision outcomes, response magnitude, or model scores. Plan count co-varies with actor density, so comparisons between $K = 1$ and $K = 2$ are not controlled plan-count ablations. The separate pilot and additional four-plan cases are excluded from the reported 120 pairs.

All checkpoints remain frozen; there is no CARLA training or per-seed input retuning. Preprocessing is fixed from the pilot. Unknown input voxels use the free-class embedding, and input class-12 voxels below -0.4 m are masked before this imputation. This is a flat-town diagnostic rule, not observed-free evidence or a modification to scoring GT. The shared world-fixed query can extend beyond training-time relative query offsets. Consequently, this out-of-domain diagnostic does not isolate interaction prediction from query extrapolation.

D.2 SCORING AND MATCHED CONTROLS

Metrics use native surface annotations on pair-common valid support, excluding unknown voxels. Controlled and responder regions use 1.2 m dilations of paired actor footprints, with overlaps between role regions removed. Scene mIoU scores non-free semantics; controlled/responder IoUs merge vehicle classes within their respective regions. Unlike OpenOccupancy's vehicle-BEV metric, these IoUs score valid 3D surface voxels within the extruded actor regions.

Signed occupancy changes. Let V_A,V_B denote reference and intervention vehicle occupancy. Added and removed voxels are $D^+ = V_B \setminus V_A$ and $D^- = V_A \setminus V_B$. We compare predicted and ground-truth D^+ and D^- separately, then sum their TP/FP/FN counts to obtain $\mathrm{F1} = 2\mathrm{TP}/(2\mathrm{TP}+\mathrm{FP}+\mathrm{FN})$ and recall $\mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$. All/controlled/responder change scores use the corresponding valid regions. A correct final vehicle location alone does not demonstrate a correct predicted change between plans.

Stable false-change rate. On valid voxels whose GT semantic label is unchanged between A and B, sFCR measures the fraction with changed predicted semantic labels. We compute this rate separately for vehicle, other occupied, and free strata, then equally average nonempty strata. This prevents the large free-space region from dominating the score. Lower sFCR is better, but freezing the whole scene can trivially suppress changes; it must be read with responder-change recall.

Counts are pooled within each initial-layout group before equally averaging layouts and the 1/2/3-second horizons, then seeds. Reference/intervention branches contribute to absolute occupancy counts; signed effects compare the two branches. Local editing starts from each seed's own Full reference prediction, uses the supplied plans, and leaves content outside its declared support unchanged. It reads neither future occupancy targets nor the model's intervention prediction. Thus it tests how much performance can come from moving controlled content without predicting a response.

Table 15. **BEP-Pairs validity checks.** All checks cover the three reported seeds and the fixed 120-pair set. They verify implementation invariants, not the accuracy of physical responses.

Check	Verified outcome
Coverage and loading	Six neural evaluations complete 120/120 pairs with strict checkpoint loading.
Repeated reference	Identical A/A-repeat predictions; only four historical observations are encoded.
Future-target independence	Future occupancy targets do not enter neural inference or local editing.
$K = 1$ identity control	Full and ego-only outputs are bitwise identical for all 60 cases in every seed.
Unavailable intervention	All 20 non-ego-only interventions produce unchanged outputs under ego-only input in each seed.
Matched scoring	GT counts, ROI support, stable-stratum denominators, case sets, and layout weights agree across methods and seeds.

D.3 Per-Seed Results and the Two-Plan Subset

Table 16 expands the main-paper three-seed means without selecting the strongest checkpoint separately for simulation. Table 17 isolates the 60 two-plan pairs, where Full and ego-only differ in surrounding-plan availability. All 60 one-plan cases are exact identity controls between those two inputs.

Table 16. **BEP-Pairs full results.** All 120 admitted pairs; every seed and matched control. All scores are percentages. Scene: semantic mIoU; Ctrl./Resp.: controlled/responder vehicle IoU. Higher is better except sFCR.

		Occupancy			Change F1			Resp.	
Method	Seed	Scene	Ctrl.	Resp.	All	Ctrl.	Resp.	recall	sFCR
Ego-only	17	49.52	55.85	49.59	28.75	41.09	9.81	8.38	1.50
Ego-only	29	43.43	52.97	44.24	29.46	38.46	8.41	7.24	0.64
Ego-only	43	44.22	56.65	47.69	34.05	42.51	14.06	13.47	0.52
Ego-only	Mean	45.72	55.16	47.17	30.75	40.69	10.76	9.70	0.89
Ego-only	SD	3.31	1.93	2.72	2.88	2.06	2.94	3.31	0.54
Local edit	17	49.92	61.36	47.88	39.63	51.16	0.00	0.00	0.21
Local edit	29	43.58	59.55	44.38	39.35	50.12	0.00	0.00	0.27
Local edit	43	44.23	62.11	46.25	40.59	51.98	0.00	0.00	0.28
Local edit	Mean	45.91	61.01	46.17	39.85	51.09	0.00	0.00	0.25
Local edit	SD	3.49	1.31	1.75	0.65	0.93	0.00	0.00	0.04
Full	17	49.71	61.51	49.53	35.17	49.01	10.62	9.27	1.53
Full	29	43.63	59.15	45.42	35.76	46.21	10.34	8.92	0.73
Full	43	44.28	62.35	48.32	39.41	49.28	15.06	14.96	0.62
Full	Mean	45.87	61.00	47.76	36.78	48.17	12.01	11.05	0.96
Full	SD	3.34	1.66	2.11	2.30	1.70	2.65	3.39	0.50

Table 17. **BEP-Pairs $K = 2$ subset.** The 60 two-plan pairs; means and sample SD over three seeds. All scores are percentages. Scene: semantic mIoU; Ctrl./Resp.: controlled/responder vehicle IoU. Higher is better except sFCR.

		Occupancy			Change F1			Resp.	
Method	Seed	Scene	Ctrl.	Resp.	All	Ctrl.	Resp.	recall	sFCR
Ego-only	Mean	45.63	48.90	46.71	23.75	31.64	8.65	7.38	0.76
Ego-only	SD	3.57	2.10	3.24	2.34	1.73	3.17	3.06	0.41
Local edit	Mean	45.94	60.50	46.28	38.96	49.99	0.00	0.00	0.26
Local edit	SD	3.72	1.50	1.87	0.76	1.10	0.00	0.00	0.03
Full	Mean	45.93	60.39	47.86	35.61	46.36	11.12	10.03	0.90
Full	SD	3.63	1.62	2.01	1.25	1.61	2.66	3.27	0.33

For two-plan pairs, additional plans improve mean controlled IoU from 48.90 to 60.39 and responder-change F1 from 8.65 to 11.12, with positive paired differences in each seed. Responder recall remains only 10.03, and sFCR rises from 0.76 to 0.90. Local editing achieves stronger controlled-change F1 but zero responder-change F1 and recall. These results support partial response prediction beyond local transport, not reliable interaction recovery. The 3.34-point sample standard deviation of Full's overall scene mIoU also shows substantial downstream seed variation on this development diagnostic. A single map, two road segments, and following interactions do not establish generalization to broad urban counterfactual behavior.

E Physical Ego Occupancy

E.1 Representation Rationale and Construction

Spatial querying exposes a distinction that ego-centered forecasting can leave implicit: the coordinate origin is a reporting choice, whereas ego is a physical occupied object. If missing ego-body annotations are treated as free space, a model can learn a regularity tied to where the vehicle usually appears in the training grid. Re-centering on another vehicle then moves a genuinely occupied object into that familiar location. This motivates checking both the occupancy labels and the temporal behavior of the learned representation, rather than assuming that a coordinate-aware architecture alone removes such biases.

In our OpenOccupancy pipeline, non-free source voxel centers are transformed into each requested frame and rasterized at 0.4 m resolution, using majority occupied-class assignment where multiple source voxels map to one cell. A fixed ego-body mask is converted to metric points in the physical ego frame, transformed into the query frame, and rasterized as car. It follows the actual ego pose in both observations and targets; no vehicle is inserted at a translated query origin. This template is a geometric approximation, not recovered sensor ground truth. The vehicle-BEV evaluation excludes the physical ego region as specified in Appendix B.

E.2 Two Diagnostic Observations

Anchor-centered diagnostic. Without retraining, we evaluate the original I^2-World in windows centered and yaw-aligned to a surrounding vehicle. In scene-0796 (record 1775), this anchor is present in all four observations and future labels, yet its box contains no predicted car voxels across six future steps and is predicted 78.65–100% free. The center vehicle comes from recorded occupancy, not an inserted template. Original training labels are 97.12% free within a 231-voxel ego-body mask over 3,000 sampled frames. This case is consistent with an empty-center prior, but the accompanying coordinate and motion distribution shifts prevent isolating its cause.

Ego-fill diagnostic. We revisit a zero-shot test of the original I^2-World (Liao et al., 2025) tokenizer and forecaster on the first 512 Occ3D-nuScenes (Tian et al., 2023) validation records. A compact, fixed car-class template of 261 voxels is inserted into observed grids and evaluation targets before token export, without retraining either model. This diagnostic uses ego-centered output frames, with current reconstruction at 0 s and six future predictions. It is separate from our OpenOccupancy

experiments and is not a BEP-World ablation. Let M denote the template and R its bounding box expanded by three voxels and clipped to the grid. Panel (a) reports the fractions of M predicted as car or free. Panel (b) pools intersections and unions across samples for $|P_{\text{car}} \cap M|/|(P_{\text{car}} \cap R) \cup M|$. This local template IoU penalizes car predictions outside M within R, including any nearby cars, and is not full-scene semantic IoU.

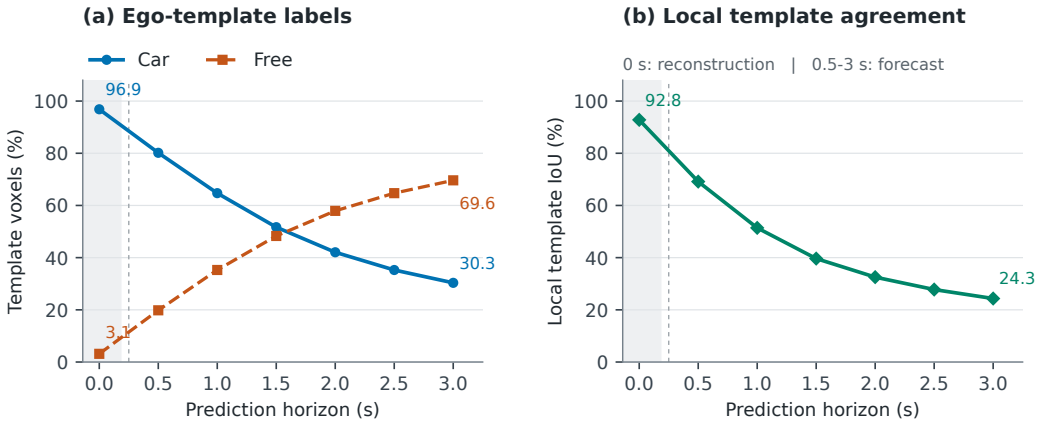

Figure 5: **Ego-body retention in a pretrained forecaster.** Original I^2-World on 512 Occ3D validation samples with inference-time ego filling only. (a) Predicted labels within the ego template. (b) Local template IoU. The shaded 0 s column denotes current reconstruction, not a future prediction.

E.3 Interpretation, Insight

In the ego-fill diagnostic, car retention falls from 96.85% at reconstruction to 30.35% at 3 s, while the predicted-free fraction rises from 3.14% to 69.63%. Local template IoU falls from 92.80% to 24.33%. The high initial reconstruction score and subsequent decay separate two questions: can the tokenizer represent an occupied center, and does the temporal predictor retain it? The first does not ensure the second. Consequently, tokenizer reconstruction alone is insufficient evidence that a world model has removed a spatial prior or learned persistent object dynamics.

The anchor-centered case complements this observation because its center vehicle comes from recorded occupancy rather than an inserted template. Together, the diagnostics motivate representing physical ego occupancy consistently and checking forecasting separately from reconstruction.

F Effect of the Training Recipe

We compare direct end-to-end training with three-stage training, using the same BEP-World architecture. Both use 128,088 end-to-end updates with the same future-query and plan-subset sampling rules.

Table 18 compares final EMA checkpoints with seed 17 under the same evaluation targets, anchors, and scoring protocol. The three-stage results here are a single run, not the three-seed means in the main table.

Table 18. **Direct versus three-stage adaptation.** Scores average 1/2/3-second horizons. S_2/S_3 equally average plan limits $k \in \{1,3,7\}$. Δ_A is the within-checkpoint V-mIoU gain from surrounding plans (pp). Difference is three-stage minus direct (pp). **Best**/second-best scores are marked per row.

Setting	Metric ↑	Direct end-to-end adaptation	Three-stage adaptation	Difference (pp)
S_0	Occupied IoU	**59.37**	59.34	−0.03
	mIoU	41.12	**42.35**	+1.23
S_1	Q-mIoU	37.46	**38.50**	+1.04
	N-mIoU	12.30	**12.46**	+0.16
S_2	3D mIoU	40.23	**41.28**	+1.05
	V-mIoU	55.37	**55.71**	+0.34
	Δ_A (pp)	**1.59**	1.27	−0.33
S_3	Q-mIoU	36.53	**37.39**	+0.86
	N-mIoU	11.63	**11.66**	+0.03
	V-mIoU	**49.96**	49.88	−0.08
	Δ_A (pp)	**1.55**	1.27	−0.28

Three-stage adaptation improves full-window semantic mIoU across all four settings, with gains of 1.23 points in S_0 and 1.04 points in S_1. Its benefit is primarily semantic forecasting accuracy: S_0 occupied IoU is nearly unchanged, and novel-region mIoU gains are small. Although absolute S_2 vehicle mIoU increases, the incremental plan gains in both S_2 and S_3 decrease. The results support the adopted recipe as a way to improve factual semantic prediction.

G Additional Temporal Visualizations

Figures 6–8 show three additional OpenOccupancy validation examples. The top row contains the four observed input frames; $t-3, \ldots, t$ denote frame indices, not elapsed seconds. The remaining rows compare future ground truth with BEP-World predictions using the ego plan only or the ego plan plus three peer plans, with the checkpoint, history, and query sequence held fixed within each scene. Historical grids use their native ego frames; future queries use the indicated translation relative to each corresponding future ego pose, rather than a fixed world-space window. All panels show the full $64 \times 64 \times 8$ m volume at 0.4 m resolution, with Occ3D colors and the same rendering view. White omits free or invalid cells. These selected examples retain prediction errors and do not imply uniform improvement from additional plans.

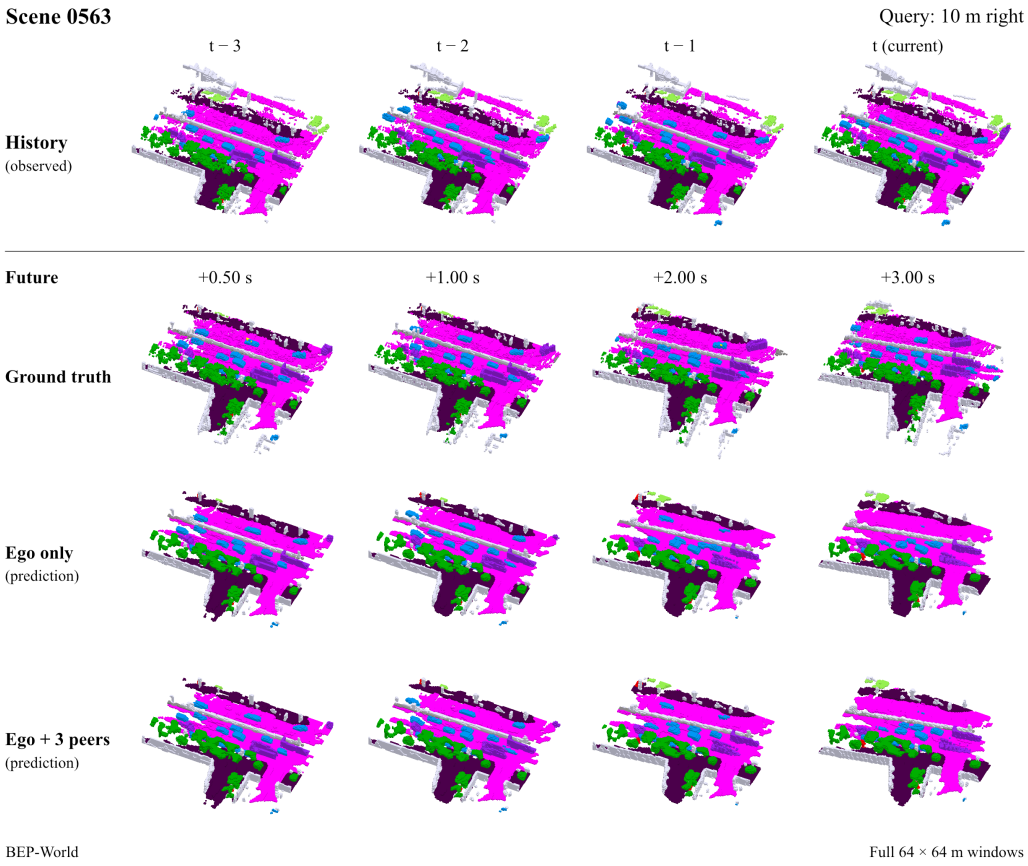

Figure 6: **Temporal occupancy forecasts in scene-0563.** Four observed frames and four future horizons under a query shifted 10 m to the right. Future columns show actual elapsed times rounded to two decimals; rows compare ground truth, ego-only conditioning, and conditioning on ego plus three peer plans.

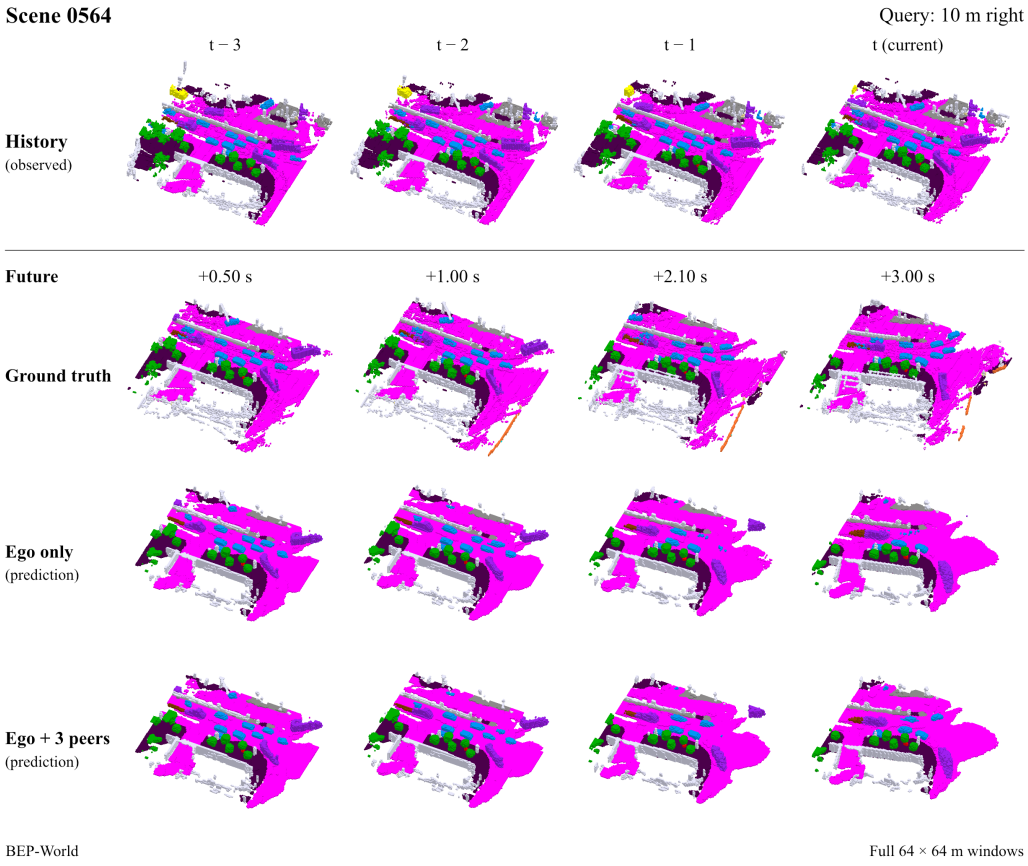

Figure 7: **Temporal occupancy forecasts in scene-0564.** The same history and prediction layout as Figure 6, with a query shifted 10 m to the right. The third displayed forecast is at +2.10 s, reflecting its recorded timestamp.

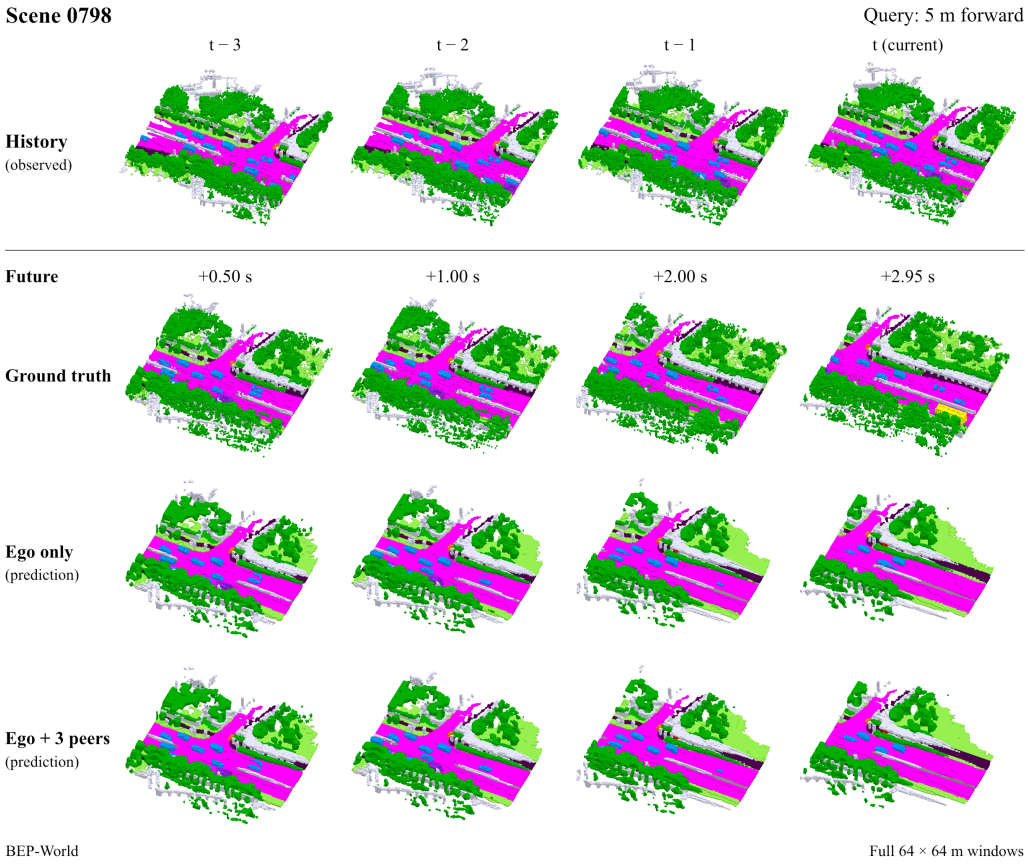

Figure 8: **Temporal occupancy forecasts in scene-0798.** The same history and prediction layout as Figure 6, with a query shifted 5 m forward. The final displayed forecast is at $+2.95$ s, reflecting its recorded timestamp.