ROADOCC LEARNS WHEN TO PERSIST, TRANSPORT, OR REFRESH MEMORY FOR ROADSIDE OCCUPANCY PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Fixed roadside cameras repeatedly observe a stable scene overlaid by sparse moving traffic. Temporal memory can recover weak observations, but reusing moving evidence at stale locations can corrupt occupancy predictions. Motion compensation addresses displacement, while reliance on the resulting history remains a separate learning problem. We introduce RoadOcc, which learns soft routing among fixed-coordinate history (*Persist*), velocity-addressed history (*Transport*), and current evidence (*Refresh*). Motion state and class-consistent historical support supervise these source choices. Dynamic-aware cross-attention (DCA) updates candidate locations, multi-scale voxel velocity estimation (VVE) constructs transport addresses from multi-scale current–history correspondence, and velocity-guided dynamic sparse fusion (VDSF) combines routed evidence under fixed sparse-token budgets. On InfraOcc, RoadOcc reaches 65.29 mIoU and 32.37 dynamic mIoU, gains of 4.44 and 4.71 over STCOcc. Controlled address experiments show that VVE raises dynamic mIoU by 0.87 over fixed-coordinate reading. Across three seeds, supervised P/T/R adds 1.40 dynamic points over motion-corrected retrieval, while removing Refresh costs 0.32 points. Results from two transfer models, Occ3D-nuScenes, and longer intervals provide additional support. Code will be released.

1 INTRODUCTION

Semantic occupancy prediction reconstructs free space, static layout, and traffic participants in a common voxel volume. Camera-only methods have made this representation practical through sparse voxel queries, BEV lifting, tri-plane features, and unified panoptic volumes (Cao & de Charette, 2022; Li et al., 2023a; Wang et al., 2023; Tian et al., 2023; Li et al., 2025b; Huang et al., 2023; Wei et al., 2023; Tang et al., 2024; Huang et al., 2024b). However, these methods are predominantly developed for moving platforms, where each observation is interpreted in an ego-centric frame. A roadside sensor instead repeatedly observes a persistent scaffold overlaid by sparse, short-lived dynamics. History is most valuable for weak or occluded actors, yet stale footprints can overwrite exactly these targets. The fixed world frame therefore shifts the central temporal problem to local evidence management. Static history is already aligned, whereas dynamic history must be selectively relocated or rejected.

The same history can provide either valid support or stale evidence. Fixed-coordinate memory preserves static layout and can recover weak or occluded vehicle boundaries, but it does not relocate an object that moved within the fixed coordinate system. Reading such memory from the former footprint can blur a dynamic boundary, duplicate an object, or attach it to nearby static layout. Discarding history avoids these errors but also removes useful support. The challenge is therefore not simply whether to use history, but to decide where temporal evidence is needed, where supporting history should be read, and how strongly to rely on each source. These cases can coexist within one scene or along a single object boundary, making a uniform temporal fusion rule insufficient.

Recent temporal methods demonstrate the value of multi-frame context for occupancy, flow, and 3D perception (Ma et al., 2024a; Zheng et al., 2024; Wang et al., 2025; Chen et al., 2025a; Liao et al., 2025; Liu et al., 2025; Huang & Huang, 2022; Chang et al., 2024; Leng et al., 2025). Existing

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.

designs already incorporate temporal memory, motion alignment, and adaptive fusion. Motion alignment addresses displacement, but a flow-proposed address may still lack class-consistent support for the current voxel. A stationary object may reuse fixed-coordinate memory, a moving object may retrieve memory near $\mathbf{x}-\Delta t\,\mathbf{v}_t(\mathbf{x})$, and a newly appearing object may have to rely on current evidence. Adaptive weighting can represent these alternatives. Our focus is to supervise the source choice explicitly using motion state and semantic support, giving temporal fusion a structured learning target. We therefore formulate temporal occupancy as explicit evidence-source routing. *Should each selected voxel Persist, Transport, or Refresh?* Persist reuses the fixed-coordinate source, Transport applies backwarp sampling to velocity-addressed history, and Refresh uses no history, only current evidence. Flow proposes where history should be read, while P/T/R learns preferences over the three sources. The supervision favors supported fixed history for stationary targets and supported transported history for moving targets, using Refresh when the corresponding candidate lacks support.

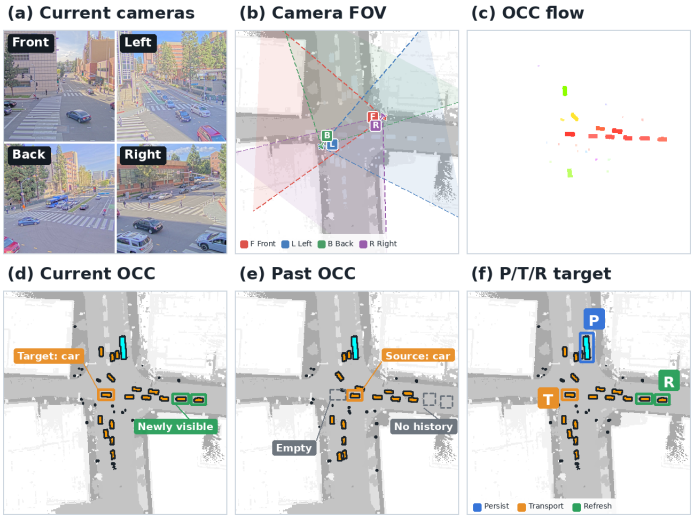


Figure 1: **Real-scene evidence for explicit source routing.** The top row shows current views, camera FOVs, and current-aligned occupancy flow (a–c). The bottom row shows current/past occupancy and P/T/R targets (d–f), illustrating a persistent bus, transported car, and newly visible cars assigned Refresh.

RoadOcc realizes this principle through a DCA–VVE–VDSF chain. DCA localizes the dynamic candidate and refines the current observation. VVE then combines current and temporal flow into dynamic flow, after which VDSF selects fixed-coordinate, backwarped, or current evidence. The stages therefore resolve where temporal reasoning is needed, where history should be read, and which source should enter fusion. Their intermediate outputs allow separate diagnostics of dynamic localization, historical addressing, and source selection. Figure 1 grounds the three evidence-source alternatives. Our contributions are summarized below.

- We propose RoadOcc, which formulates roadside temporal occupancy as a Persist–Transport–Refresh source-selection problem. Motion state and class-consistent support supervise the source preferences, complementing motion-based historical addressing.
- We implement this formulation in a sparse, multi-scale pipeline. DCA allocates image updates, VVE estimates transport addresses, and VDSF combines routed sources while retaining a short background memory path.
- We report joint occupancy and motion improvements on InfraOcc, three-seed routing controls, and module transfer to two temporal models. Occ3D-nuScenes and temporal-gap evaluations further test the design beyond the primary roadside setting and training cadence.

2 Related Work

2.1 Camera Occupancy and Roadside Perception

Camera occupancy methods lift multi-view image evidence into volumetric representations. MonoScene Cao & de Charette (2022) and VoxFormer Li et al. (2023a) establish image-based semantic scene completion, while BEVDet, BEVDepth, and BEVFormer improve multi-view lifting and BEV reasoning (Huang et al., 2021; Li et al., 2023b; 2025b). Occupancy-specific designs span tri-plane, dense, sparse, Gaussian, and forward–backward representations (Huang et al., 2023; Wei et al., 2023; Zhang et al., 2023; Tang et al., 2024;

Huang et al., 2024b; Li et al., 2023c), as well as panoptic and instance-aware prediction (Wang et al., 2024; Li et al., 2025a). Unlike MC-BEVRO's roadside BEV occupancy monitoring (Vaghela et al., 2025), RoadOcc routes evidence for class-aware 3D occupancy. InfraOcc provides a roadside occupancy benchmark emphasizing persistent infrastructure (Yang et al., 2026). Figure 2 reveals its temporal asymmetry. Fixed roadway structure provides broad, stable support, while dynamic participants form sparse, short-lived trails. This asymmetry favors retaining static context while selectively relocating moving evidence.

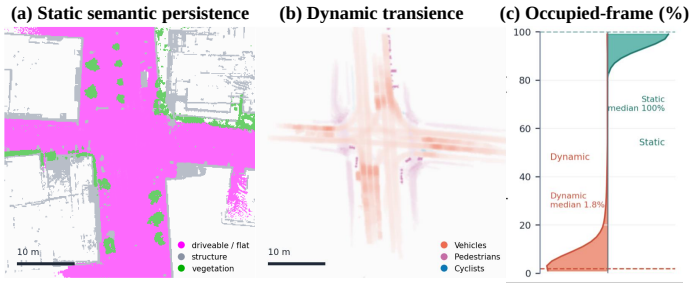


Figure 2: **Static-to-dynamic persistence in InfraOcc.**

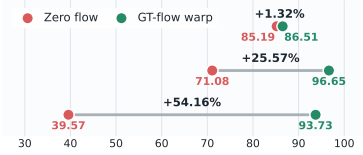
2.2 Temporal Occupancy, Motion, and Alignment

Temporal occupancy methods improve memory through recurrent aggregation, motion-aware alignment, and explicit association. Recurrent and video-based models maintain or forecast scene state (Huang & Huang, 2022; Chang et al., 2024; Huang et al., 2024a; Mohan et al., 2026), while occupancy–flow methods couple volumetric semantics with motion (Ma et al., 2024a; Zheng et al., 2024; Wang et al., 2025; Chen et al., 2025a; Liao et al., 2025; Liu et al., 2025; Murhij & Yudin, 2024; Zheng et al., 2026). Other approaches organize temporal cues, learn semantic associations, or retrieve traversal priors (Leng et al., 2025; Chen et al., 2025b; Ouyang et al., 2025; Yuan et al., 2025). RoadOcc builds on temporal alignment and adaptive fusion by supervising a choice among fixed-coordinate history, transported history, and current evidence. Its emphasis is the motion- and support-based learning target for source selection, rather than the weighted-sum operation alone.

Table 1: **Generated GT-flow target at $t + 0.5$ s.** Zero-flow and GT-flow transport share the inter-frame transform and differ only by object displacement. Same/Dyn./Occ. denote unchanged semantics, dynamic class, and occupied class. Values are percentages and N counts valid queries in millions.

Speed	N (M)	Zero-flow transport			GT-flow warp			Δ Same
		Same	Dyn.	Occ.	Same	Dyn.	Occ.	
All	11.83	72.93	78.86	90.92	**89.21**	**95.79**	**99.38**	+16.28
Slow ($<$ 2 m/s)	7.87	85.19	94.08	98.80	86.51	**96.25**	**99.70**	+1.32
Med. (2–5 m/s)	1.16	71.08	71.38	87.50	**96.65**	**97.64**	**99.53**	+25.57
Fast ($\geq$ 5 m/s)	2.80	39.57	39.59	70.41	**93.73**	**93.75**	**98.43**	+54.16

Figure 3: **Motion compensation restores temporal consistency.** GT flow preserves dynamic support across speed ranges, whereas fixed-coordinate retrieval becomes increasingly misaligned as speed grows.



Flow-based alignment is closest to this problem. AsyncBEV aligns asynchronous cross-modal observations with BEV feature flow (Wang et al., 2026). In the fixed roadside setting, Table 1 and Figure 3 show that GT-flow warping improves same-class agreement by 1.32 points for slow objects, but by 25.57 and 54.16 points for medium and fast objects, motivating selective motion compensation. This is a diagnostic of the constructed targets. Low-speed target repair also uses future semantic support (Appendix B.2).

3 Method

3.1 Preliminary Analysis

Let $\mathcal{I}_t = \{I_t^n\}_{n=1}^4$ denote the synchronized roadside images at time t. RoadOcc predicts semantic occupancy $\mathbf{y}_t$ over a voxel domain Ω and a planar velocity field $\mathbf{v}_t : \Omega \to \mathbb{R}^2$. Among the 18 occupancy states, the dynamic set is $\mathcal{C}_{\text{dyn}} = \{\text{bicycle, bus, car, motorcycle, pedestrian, truck}\}$. Velocity is expressed in the common roadside horizontal frame.

Because fixed roadside cameras share a persistent physical frame, cached history already lies on the current voxel grid at inference. We only reconcile relative BEV augmentations during training.

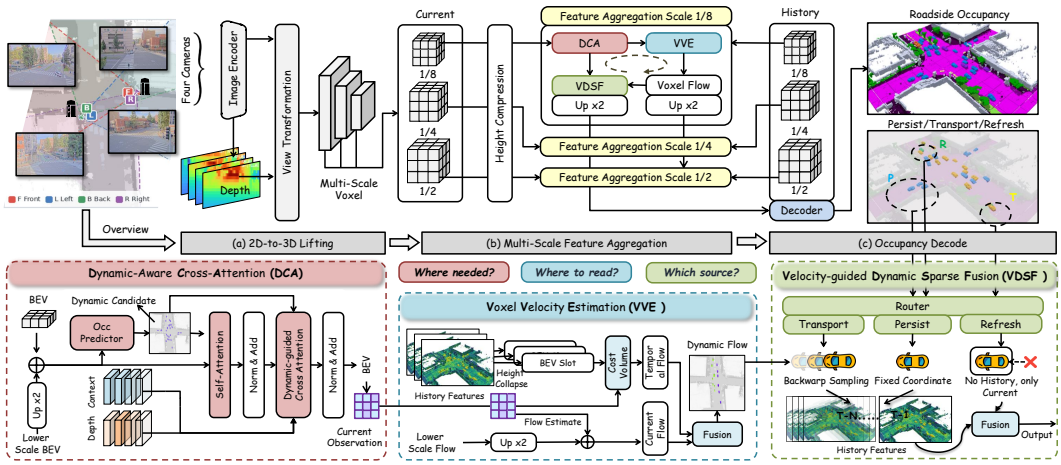


Figure 4: **RoadOcc framework.** The image encoder and view transformation construct multi-scale voxel features. At each feature aggregation scale, DCA produces the current observation, VVE fuses current, temporal, and lower-scale flow into dynamic flow, and VDSF routes fixed-coordinate, backwarped, or current evidence to fusion. Coarse-to-fine outputs feed occupancy decoding for semantic occupancy and voxel flow.

Under a locally constant velocity approximation, history supporting a current voxel $\mathbf{x}$ is read from

$$\widetilde{\mathbf{h}}_{t-\Delta t}(\mathbf{x}) = \mathrm{Sample}\big(\overline{\mathbf{h}}_{t-\Delta t}, \mathbf{x} - \Delta t\, \mathbf{v}_t(\mathbf{x})\big)\,, \tag{1}$$

where $\overline{\mathbf{h}}$ is same-grid history and Sample is trilinear interpolation. The three candidates are fixed-coordinate history at $\mathbf{x}$ (Persist), flow-addressed history at $\mathbf{x} - \Delta t\, \mathbf{v}_t(\mathbf{x})$ (Transport), and current evidence (Refresh). Equation 1 specifies an address, while VDSF learns how to combine these sources under motion- and support-based supervision.

3.2 Overview

RoadOcc decodes $\{\mathbf{V}_t^s\}_{s\in\{1/8,1/4,1/2\}}$ from coarse to fine. As shown in Figure 4, the encoder lifts four camera views into multi-scale voxel features. At each aggregation stage, DCA determines where evidence is needed, VVE determines where history should be read, and VDSF determines which source should enter fusion. The output feature and dynamic flow seed the next finer stage before occupancy decoding. Supplementary Figure A4 visualizes this progression.

3.3 Dynamic-aware Cross-attention (DCA)

Dynamic objects occupy few voxels, so uniformly repeating image cross-attention spends most computation on stable structure. After the first nonempty-guided image update, DCA combines the current representation with an upsampled lower-scale BEV to derive a dynamic candidate. Let $\boldsymbol{\pi}^s_{t,\mathrm{cur}}$ and $\boldsymbol{\pi}^s_{t,\mathrm{sta}}$ denote the provisional distributions produced by the current and static prediction heads after the first image update, and let $\mathcal{C}^+_{\mathrm{sta}}$ contain the static classes and free space. The mean semantic discrepancy $\delta^s_{t,\mathrm{sta}}$ over $\mathcal{C}^+_{\mathrm{sta}}$ and current dynamic probability $p^s_{t,\mathrm{dyn}}$ form the candidate map

$$c^s_t(\mathbf{x}) = \mathrm{Norm}\big(\max\{\delta^s_{t,\mathrm{sta}}(\mathbf{x}), p^s_{t,\mathrm{dyn}}(\mathbf{x})\}\big)\,. \tag{2}$$

Here Norm divides by the per-sample spatial maximum, clamped to 10^{-6}. Discrepancy recovers departures from the static hypothesis, while dynamic probability retains visible actors. Their maximum also preserves early semantic misses detected through change evidence.

The candidate map controls a second image cross-attention rather than directly reweighting the voxel feature. Following projected deformable attention (Li et al., 2025b) and occupancy-aware reference sampling (Liao et al., 2025), let $\mathbf{q}^s_{t,\mathbf{x}}$ be the BEV query at $\mathbf{x}$, $\mathcal{R}_n(\mathbf{x})$ its visible height anchors in camera n, and $\mathbf{F}^n_t$ the corresponding image feature. DCA retains $\mathcal{R}^s_{n,c}(\mathbf{x}) = \{\mathbf{r} \in \mathcal{R}_n(\mathbf{x}) : u_{\mathbf{r}} <$

$c_t^s(\mathbf{r})\}$ and updates the query as

$$\mathbf{q}_{t,\mathbf{x}}^{s,+} = \mathbf{q}_{t,\mathbf{x}}^{s} + \mathbf{W}_o \frac{1}{|\mathcal{V}_\mathbf{x}|} \sum_{n \in \mathcal{V}_\mathbf{x}} \sum_{\mathbf{r} \in \mathcal{R}_{n,c}^s(\mathbf{x})} \beta_{n,\mathbf{r}}^s \, \mathcal{A}_{\text{def}}\big(\mathbf{q}_{t,\mathbf{x}}^s, \mathcal{P}_n(\mathbf{r}), \mathbf{F}_t^n\big)\,, \tag{3}$$

where $\mathbf{r} = (\mathbf{x}, z_\mathbf{r})$ indexes a voxel-height score shared across cameras, $\mathcal{V}_\mathbf{x}$ contains valid cameras, and $\mathcal{P}_n$ projects anchors. Depth consistency $\beta_{n,\mathbf{r}}^s$ multiplies normalized deformable weights before camera averaging. During training, $u_\mathbf{r}$ is sampled from a normal distribution with mean 0.5 and standard deviation 1, truncated to $[0, 1]$. At inference, it is fixed to 0.5. The stochastic rule exposes uncertain boundary voxels during training, while deterministic thresholding keeps inference stable. Queries without retained support follow the residual path unchanged.

The candidate gates a second cross-attention over image depth context and also allocates VDSF tokens. DCA therefore localizes where additional image and temporal evidence is needed without repeatedly updating the stable scene.

3.4 Multi-scale Voxel Velocity Estimation (VVE)

DCA refines the current observation but does not locate displaced history. VVE estimates its address from the DCA feature and nearest valid history. From the current occupancy distribution $\boldsymbol{\pi}_t^s(\mathbf{x})$, it computes dynamic support $d_t^s = \sum_{c \in \mathcal{C}_{\text{dyn}}} \pi_{t,c}^s$ and nonempty support $e_t^s = 1 - \pi_{t,\text{free}}^s$. Here d_t^s guides height compression into the current BEV slot $\mathbf{B}_t^s$, while e_t^s guides temporal token selection. Applying the same compression to history gives $\overline{\mathbf{B}}_{t-\Delta t}^s$.

VVE next compares each current BEV slot with a small neighborhood in the historical slot, producing the local cost volume

$$\mathbf{K}_t^s(\mathbf{x}, \delta) = \left\langle \frac{\mathbf{B}_t^s(\mathbf{x})}{\|\mathbf{B}_t^s(\mathbf{x})\|_2}, \frac{\overline{\mathbf{B}}_{t-\Delta t}^s(\mathbf{x}+\delta)}{\|\overline{\mathbf{B}}_{t-\Delta t}^s(\mathbf{x}+\delta)\|_2} \right\rangle, \quad \delta \in [-r, r]^2, \tag{4}$$

where $r = 2$ yields a 5×5 neighborhood and both norms use an $\epsilon = 10^{-6}$ floor. The cost-volume decoder converts this local correspondence evidence into temporal flow.

VVE estimates flow recursively rather than solving each resolution in isolation. Let s^- denote the preceding coarser stage and set the upsampled lower-scale flow $\mathbf{v}_{t,\uparrow}^s$ to zero at $1/8$ and to $\text{Up}_s(\mathbf{v}_{t,\text{cur}}^{s^-})$ otherwise. This prior initializes the current flow. Current appearance and its semantic embedding $\mathbf{S}_t^s$ predict a current-only correction, while the history feature and cost volume predict the temporal flow. Their fusion is

$$\begin{aligned} \mathbf{v}_{t,\text{cur}}^s = \mathbf{v}_{t,\uparrow}^s + \Delta\mathbf{v}_{t,\text{cur}}^s, \quad \mathbf{v}_{t,\text{tmp}}^s = f_{\text{tmp}}(\mathbf{B}_t^s, \overline{\mathbf{B}}_{t-\Delta t}^s, \mathbf{K}_t^s), \\ \widehat{\mathbf{v}}_t^s = (1-\alpha_t^s)\mathbf{v}_{t,\text{cur}}^s + \alpha_t^s \mathbf{v}_{t,\text{tmp}}^s + \Delta\mathbf{v}_{t,\text{ref}}^s. \end{aligned} \tag{5}$$

Here f_{tmp} is the cost-volume decoder and $\alpha_t^s = 0.5\,\mathbf{1}_{\text{hist}} \max_z d_t^s$ gates history, becoming zero at sequence starts. The residual $\Delta\mathbf{v}_{t,\text{ref}}^s$ produces the dynamic flow, which is lifted across height as the VDSF address. Coarse stages capture motion structure and fine stages refine boundaries.

3.5 Velocity-guided dynamic sparse fusion (VDSF)

Velocity-guided sparse fusion. Applying the full temporal memory after DCA and VVE would mostly repeat static context. VDSF makes gather–fuse–scatter (Liao et al., 2025) motion aware by combining DCA guidance with nonempty support under a fixed token budget

$$\mathbf{g}_t^s = \text{clip}(c_t^s + \eta\, e_t^s, 0, 1), \qquad \mathcal{X}_t^s = \text{TopK}(\mathbf{g}_t^s), \tag{6}$$

where η balances change evidence and occupied support. The selected set $\mathcal{X}_t^s$ receives full stage-local history. Its complement uses a short fixed-coordinate path, preserving background context at lower cost.

Each scale receives multi-frame history features and their VVE motion priors. For $\mathbf{x} \in \mathcal{X}_t^s$, elapsed time converts dynamic flow into a voxel offset. Backwarp sampling produces the Transport candidate $\mathbf{h}^T$, while the fixed-coordinate read produces the Persist candidate $\mathbf{h}^P$.

Table 2: **Main InfraOcc occupancy results.**

Method	gIoU	mIoU			Bike	Bus	Car	Motorcycle	Pedestrian	Truck	Other	Barrier	Cone	Driveable	Sidewalk	Terrain	Manmade	Vegetation	Free
		all	dyn	sta															
BEVDet '21	68.62	45.14	21.19	63.11	9.23	42.75	38.04	9.68	15.01	12.45	42.88	70.36	53.77	72.04	68.07	73.09	64.28	60.38	94.87
BEVFormer '22	81.72	47.76	10.73	75.53	7.97	20.39	17.65	3.47	8.62	6.30	42.47	76.26	72.83	85.19	81.95	88.95	79.52	77.05	97.56
BEVDepth '23	71.89	46.81	19.64	67.19	14.84	33.87	38.74	3.42	16.86	10.14	48.73	71.05	62.37	73.31	69.32	76.07	70.51	66.15	95.54
TPVFormer '23	73.80	43.09	4.36	72.13	0.10	3.69	13.80	0.41	6.86	1.33	42.86	75.29	69.69	80.61	77.62	86.63	72.61	71.70	96.12
SurroundOcc '23	83.78	50.77	11.37	80.32	12.66	21.90	18.54	2.06	9.50	3.55	65.09	**79.04**	72.70	86.38	83.99	91.10	83.16	81.09	97.80
CONet '23	78.56	49.25	18.93	72.00	3.06	37.24	39.47	4.42	17.64	11.72	45.01	75.31	64.94	84.29	79.45	84.27	72.96	69.79	97.08
ALOcc '25	75.66	43.31	5.86	71.40	8.41	4.04	8.52	0.00	4.48	9.71	63.90	68.97	63.40	90.09	77.77	78.16	65.07	63.85	97.12
Let Occ Flow '24	87.18	51.38	8.78	83.34	11.27	11.53	14.07	4.34	8.12	3.34	62.06	75.06	85.56	92.55	85.63	92.98	87.10	85.74	98.56
SparseOcc '24	71.61	45.42	9.41	72.42	1.50	20.94	24.87	0.00	4.60	4.57	67.68	76.37	61.51	74.67	76.59	88.54	68.79	65.19	96.31
GaussianFormer '24	77.89	43.63	10.45	68.51	3.13	22.11	23.17	1.29	8.21	4.80	36.70	68.54	55.60	82.59	79.37	74.20	77.85	73.21	96.71
CRT-Fusion-S '24	88.73	55.94	21.98	81.40	13.02	45.16	38.48	7.34	18.99	8.90	63.53	71.99	74.82	93.38	84.40	91.19	86.39	85.53	98.73
STCOcc-S '25	78.54	50.71	23.48	71.14	13.43	47.83	39.62	10.21	17.83	11.93	50.18	71.54	58.33	90.30	78.78	84.90	67.72	67.32	97.47
CRT-Fusion '24	80.35	53.83	27.02	73.94	32.58	46.63	39.49	12.22	20.81	10.41	53.08	72.27	68.15	89.06	79.09	84.48	72.06	73.30	97.68
STCOcc '25	92.54	60.85	27.66	85.74	27.13	48.78	44.93	10.23	**22.85**	12.03	68.84	70.93	81.60	95.88	89.41	95.36	91.85	92.04	99.18
RoadOcc (ours)	**94.03**	**65.29**	**32.37**	**89.98**	**37.60**	**55.46**	**49.36**	**13.01**	22.66	**16.13**	**87.66**	72.05	**86.03**	**97.85**	**92.81**	**96.90**	**92.91**	**93.66**	**99.35**

Persist–Transport–Refresh routing. Dynamic flow proposes where to read history, but a corrected address may still lack semantic support. VDSF therefore decides which source should enter fusion. It retains $\mathbf{h}^P$ and $\mathbf{h}^T$ and defines the current-only Refresh candidate as $\mathbf{h}^R = \mathbf{q}_t$. During training, the route target is constructed from the current and nearest historical semantic occupancy labels together with the motion target, and assigns each current dynamic voxel according to its motion state and class-consistent historical support as follows.

$$z_t(\mathbf{x}) = \begin{cases} P, & \|\mathbf{v}_t(\mathbf{x})\| \le \epsilon \ \wedge \ \mathrm{Supp}(\mathbf{x}), \\ T, & \|\mathbf{v}_t(\mathbf{x})\| > \epsilon \ \wedge \ \mathrm{Supp}(\mathbf{x} - \Delta t\, \mathbf{v}_t(\mathbf{x})), \\ R, & \text{otherwise}, \end{cases} \tag{7}$$

where Supp tests same-class occupancy at the same height within the local correspondence radius. We use $\epsilon = 10^{-3}$ m/s and a one-voxel radius to absorb discretization error. The construction follows the semantic occupancy label space used by InfraOcc and is evaluated at every execution scale using its physical cell size, with a native-grid endpoint for full-resolution supervision. The motion target selects which historical candidate is tested according to motion state. The fixed-coordinate candidate is used for stationary voxels and the transported candidate for moving voxels. If that candidate lacks support, the target is Refresh, even if the alternative historical address has same-class support. This is a motion-conditioned source preference. Targets use GT motion, while inference uses predicted VVE motion and learned route probabilities. Future occupancy is used only for offline target construction, never as model input.

At each feature aggregation scale, the router combines the current observation, nearest fixed-coordinate history feature, normalized dynamic flow, and upsampled coarser P/T/R distribution. It predicts stage-local logits. A lightweight native head upsamples the finest context with full-resolution flow. Softmax gives $\boldsymbol{\alpha}_t^s = (\alpha_P, \alpha_T, \alpha_R)$ and the routed candidate

$$\mathbf{h}_t^{\mathrm{PTR}} = \alpha_P \mathbf{h}^P + \alpha_T \mathbf{h}^T + \alpha_R \mathbf{h}^R, \qquad (\alpha_P, \alpha_T, \alpha_R) = \mathrm{softmax}(\psi(\cdot)). \tag{8}$$

P/T/R therefore compete as one route distribution rather than independent weights. Transport uses backwarp sampling, Persist uses the fixed-coordinate source, and Refresh uses no history, only current evidence. Fusion uses soft probabilities and reserves argmax for diagnostics.

Let $\mathbf{H}_t^{\mathrm{PTR,s}}(\mathbf{x})$ concatenate the routed multi-frame history at scale s. VDSF combines this temporal evidence with the current observation and its semantic embedding through the fusion MLP

$$\mathbf{q}_{t,\mathrm{fuse}}^s(\mathbf{x}) = \mathbf{q}_t^s(\mathbf{x}) + \Phi_s\Big([\mathbf{q}_t^s(\mathbf{x}), \mathbf{H}_t^{\mathrm{PTR,s}}(\mathbf{x}), \mathrm{Emb}(\boldsymbol{\pi}_t^s(\mathbf{x}))]\Big). \tag{9}$$

The output residual is scattered to $\mathcal{X}_t^s$ and joined with the background path. The fused volume and dynamic flow feed the next scale and are retained for subsequent frames. Detailed memory, token, source-matching, and routing settings appear in Appendices E and G.

3.6 Training Objective

Training jointly optimizes four functional objectives.

$$\mathcal{L} = \lambda_{\mathrm{depth}}\mathcal{L}_{\mathrm{depth}} + \mathcal{L}_{\mathrm{sem}} + \lambda_{\mathrm{mot}}\mathcal{L}_{\mathrm{mot}} + \lambda_{\mathrm{route}}\mathcal{L}_{\mathrm{route}}. \tag{10}$$

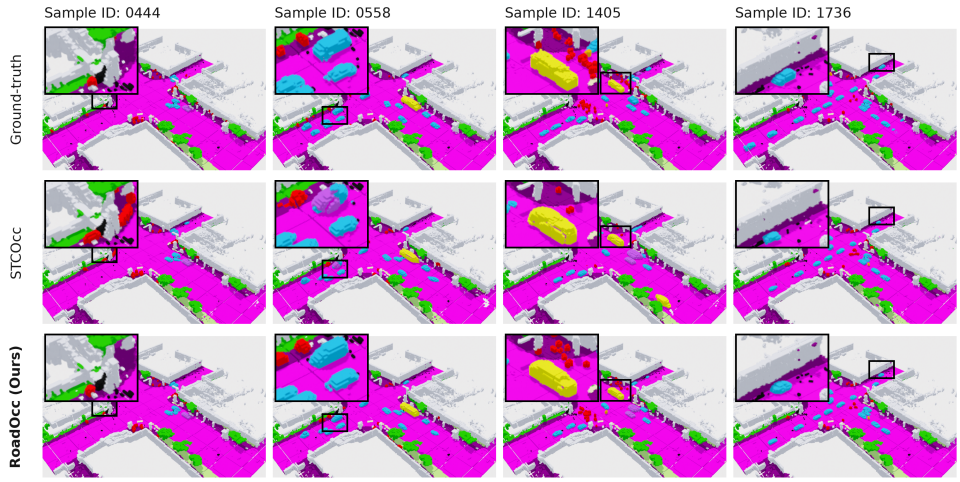


Figure 5: **Qualitative comparison with STCOcc.** RoadOcc more closely recovers dynamic-object presence, semantic identity, and geometry, including spurious pedestrian occupancy, vehicle-class confusion, bus-shape distortion, and an incomplete car footprint.

Table 3: **Joint occupancy–velocity comparison.**

Method	Occupancy					Velocity	Diagnostics	
	mIoU	Dyn.	Sta.	gIoU	RayIoU	Direct MAVE	TP-MAVE	DSR
CRT-Fusion (NIPS 2024)	53.83	27.02	73.94	80.35	48.08	2.227	1.839	57.56
+ VDSF P/T/R	57.15	28.30	78.79	83.46	52.34	1.990	1.958	59.45
Δ	+3.32	+1.28	+4.85	+3.11	+4.26	−0.237	+0.119	+1.89
STCOcc (CVPR 2025)	60.85	27.66	85.74	92.54	55.94	1.946	0.703	57.12
+ VDSF P/T/R	61.81	28.80	86.56	93.59	58.07	1.770	**0.648**	58.78
Δ	+0.96	+1.14	+0.82	+1.05	+2.13	−0.176	−0.055	+1.66
RoadOcc (ours)	**65.29**	**32.37**	**89.98**	**94.03**	**65.44**	**1.669**	0.661	**61.11**

$\mathcal{L}_{\text{depth}}$ supervises camera depth for view transformation, while $\mathcal{L}_{\text{sem}}$ covers full- and multi-scale occupancy together with the static, dynamic, and DCA auxiliaries. $\mathcal{L}_{\text{mot}}$ supervises the dynamic flow used for backwarp sampling, whereas $\mathcal{L}_{\text{route}}$ supervises the P/T/R router on valid dynamic voxels. Together, they separate where history features are read from which source enters fusion. Full target construction, weights, and schedules are given in supplementary Sec. E.

4 Experiments

Dataset and metrics. We evaluate camera-only roadside occupancy on InfraOcc (Xiang et al., 2024; Yang et al., 2026). Four synchronized views are mapped to 18 occupancy states. Under the InfraOcc protocol, semantic mIoU excludes construction vehicle, trailer, other flat, and Free. Free IoU is reported separately in the class-wise table. We report semantic and geometric IoU, next-frame dynamic IoU, and motion coverage. Direct MAVE covers all GT dynamic voxels, while TP-MAVE conditions on semantic matches and is paired with recall (DSR). InfraOcc is the primary fixed-roadside protocol. We additionally evaluate module transfer, Occ3D-nuScenes, and temporal-gap changes. Full protocols and metric definitions appear in the supplement.

Implementation details. RoadOcc uses a ResNet-50 backbone, 256×704 images, and three aggregation scales at 3.2, 1.6, and 0.8 m, followed by 0.4 m output voxels over $[-64, 64] \times [-64, 64] \times [-4.8, 1.6]$ m. Frames are 0.5 s apart. AdamW training runs for 24 epochs from 5×10^{-4} with 10^{-2} weight decay and cosine annealing. Comparisons share the voxel range, ignored classes, and evaluator. Controlled variants keep the backbone, data split, optimizer and other configuration fixed.

4.1 Main Results

InfraOcc occupancy benchmark. Table 2 tests whether the proposed routing chain improves occupancy under the InfraOcc evaluator. RoadOcc achieves the highest mIoU among the compared

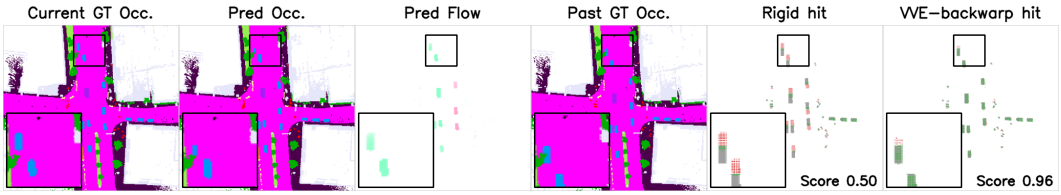


Figure 6: **Velocity-guided historical retrieval.** With queried dynamic targets fixed by current GT, the panels show the end-to-end occupancy prediction and compare a fixed-coordinate read with the VVE backtrace into aligned past GT. Black boxes are enlarged in the insets. ■ same-class matches, ■ unmatched targets, and ■ other queried voxels are shown in green, red, and gray, respectively. The score is the same-class hit rate within the one-voxel matching radius.

methods at 65.29, a +4.44 gain over STCOcc. Dynamic mIoU rises by +4.71 to 32.37, while static mIoU rises by +4.24 to 89.98. Five of six dynamic classes improve, led by Bike (+10.47) and Bus (+6.68), whereas Pedestrian changes by −0.19. Thus the aggregate gain covers both traffic participants and persistent structure, with variation by class.

Joint occupancy–velocity behavior. Table 3 evaluates motion accuracy and semantic coverage alongside occupancy. Direct MAVE measures velocity error over all GT dynamic voxels, while DSR measures dynamic semantic recall. Relative to STCOcc, RoadOcc lowers Direct MAVE from 1.946 to 1.669 and raises DSR from 57.12 to 61.11, improving motion accuracy alongside dynamic recovery. Adding VDSF P/T/R to CRT-Fusion and STCOcc likewise raises Dyn. by +1.28 and +1.14, respectively, while improving Direct MAVE and DSR. The consistent trend supports the module's utility beyond the full RoadOcc architecture. TP-MAVE conditions on semantic matches and must be read with DSR. For CRT-Fusion, the +1.89 DSR gain indicates broader dynamic recovery even though the conditional error over matched voxels increases.

Qualitative occupancy recovery. Figure 5 supports the aggregate dynamic gain with representative matched scenes. STCOcc produces stale or spurious pedestrian occupancy, confuses a vehicle class, distorts a bus footprint, and incompletely recovers a car. RoadOcc more closely recovers the GT class and extent in these examples while preserving the roadway.

Transfer and temporal gaps. On mask-trained Occ3D-nuScenes, RoadOcc improves over STCOcc by 0.41 overall and 0.79 dynamic mIoU points (Table A2). Without fine-tuning the 0.5 s-trained checkpoints, testing at 1.5 s yields 33.49 Dyn. for RoadOcc and 28.58 for STCOcc on the same 1,520 anchors (Table A4). These results extend the empirical support beyond the primary dataset and training cadence. They evaluate the complete model rather than isolate P/T/R.

Table 4: **Component ablation.**

Base	Design				Output ↑		
	DCA	VVE	VDSF	P/T/R	All	Dyn.	Nxt.
✓					60.85	27.66	26.74
✓			✓		61.05	28.01	26.99
✓	✓				61.69	28.55	27.64
✓	✓		✓		63.01	30.10	29.44
✓	✓	✓	✓		64.17	30.97	30.78
✓	✓	✓	✓	✓	**65.29**	**32.37**	**31.72**

4.2 Ablation Studies

Using Liao et al. (2025) as the baseline, we hold the backbone, token budget, temporal-memory configuration, and training protocol fixed unless explicitly varied. Tables 4, 5, and 6 then examine component composition, historical-address correction, and P/T/R source selection, respectively.

Overall ablation. Table 4 follows the routing chain. VDSF alone adds only 0.20 mIoU. Adding VDSF to the DCA variant raises Dyn. by +1.55 and Nxt. by +1.80. VVE then adds +1.16 mIoU, and P/T/R adds a further +1.12 mIoU. This progression supports combining candidate selection, motion-based retrieval, and source routing. Tables below examine latter two design in more detail.

Address correction and source selection. With capacity and token budget fixed, the fixed-coordinate and VVE reads in Table 5 differ only in historical address. VVE raises Dyn. by +0.87, lowers dMAVE from 2.146 to 1.690, and raises DSR by +1.51. P/T/R then adds +1.40 Dyn. and +1.76 DSR, showing that learned source selection improves on motion-corrected retrieval.

Table 5: **Controlled address and source selection.** All variants use the same sparse-token budget and model capacity.

Variant	Design		Output			
	VVE	P/T/R	Dyn.↑	Nxt.↑	dM↓	DSR↑
Fixed read			30.10	29.44	2.146	57.84
VVE read	✓		30.97	30.78	1.690	59.35
Full	✓	✓	**32.37**	**31.72**	**1.669**	**61.11**

Table 6: **P/T/R routing controls.** Mean ± standard deviation is reported for the multi-seed variants. w/o R denotes full w/o Refresh.

Variant	Dyn.↑	Nxt.↑	DSR↑	dM↓
None	30.97±.27	30.78±.34	59.35±.39	1.690±.019
Gate	31.25±.20	30.94±.30	59.75±.45	1.685±.018
State	31.88±.18	31.38±.25	60.52±.40	1.675±.017
w/o R	32.05±.12	31.41±.18	60.69±.05	1.679±.033
Full	**32.37**±.11	**31.72**±.26	**61.11**±.38	**1.669**±.020

P/T/R routing controls. Table 6 compares routing variants within DCA–VVE–VDSF. None omits P/T/R. Gate adaptively mixes the same candidates. State adds P/T/R supervision without executing its probabilities. Full uses the supervised routes in fusion. Full w/o R masks α_R and renormalizes (α_P, α_T). The complete control contract is given in the appendix. Across three seeds, Full reaches 32.37 ± 0.11 Dyn., compared with 30.97 ± 0.27 for None and 31.25 ± 0.20 for Gate. State and Full improve successively by 0.63 and 0.49 points over the preceding variant. The clearest gain is in dynamic occupancy. The dMAVE change from None to Full is smaller (0.021 m/s), with standard deviations of 0.019 and 0.020 m/s. Removing Refresh from Full reduces Dyn., Nxt., and DSR by 0.32, 0.31, and 0.42 points, respectively, while increasing dMAVE by 0.010 m/s. Under the same three-seed protocol, the 0.32-point Dyn. gap exceeds either reported run-to-run standard deviation (0.11 and 0.12). We treat the other metric changes as supporting trends rather than standalone significance claims.

Routing scope. Equation 7 supervises motion-conditioned semantic preferences, not instance correspondence. Labels use the nearest history, and each scale shares its predicted route across valid memory slots. GT-conditioned routing visualizations separate source choices from occupancy detection. They do not measure candidate misses or guarantee support at every predicted historical address. Dyn., Nxt., Direct MAVE, and DSR provide complementary end-to-end metrics.

Visual analysis. Figure 6 compares zero-motion reading with VVE backtracing. Current GT fixes the queried dynamic voxels and their classes, while predicted occupancy provides the end-to-end context. Red-to-green changes show dynamic flow relocating queries through backwarp sampling onto same-class past GT, isolating address quality from the subsequent routing decision.

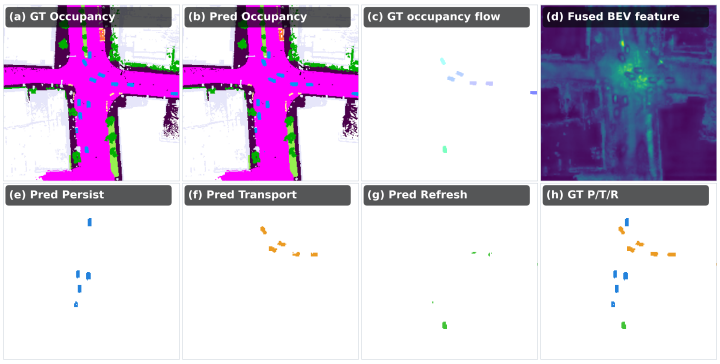


Figure 7: **Predicted P/T/R on GT dynamic support.**

On the same GT dynamic support, Figure 7 then shows Persist following fixed-coordinate support, Transport following displacement, and Refresh favoring current evidence. Together, the visualizations inspect historical addresses and learned source preferences on dynamic targets.

5 Conclusion

In this work, we present RoadOcc, a voxel-level evidence-source routing framework for temporal roadside occupancy. Beyond implicit temporal aggregation, RoadOcc explicitly learns where history should be retrieved and how strongly it should be reused. The proposed DCA uses current image evidence to identify motion-sensitive voxels, while VVE estimates their historical source addresses, and VDSF combines evidence through Persist, Transport, or Refresh routing. On the InfraOcc protocol, the unified formulation achieves the highest occupancy accuracy among the compared methods while improving dynamic recovery and velocity estimation. Extensive experiments further demonstrate the superiority of our design. Future work will extend this principle to broader moving-platform settings, longer temporal horizons, and instance-level modeling.

AI USE STATEMENT

Generative AI tools were used for literature organization, reviewer-style feedback, and language editing. The authors verified all citations, technical descriptions, analyses, and revisions, and take responsibility for the final content of this work.

REFERENCES

Anh-Quan Cao and Raoul de Charette. MonoScene: Monocular 3D Semantic Scene Completion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3991–4001, 2022.

Ming Chang, Xishan Zhang, Rui Zhang, Zhipeng Zhao, Guanhua He, and Shaoli Liu. Recurrent-BEV: A Long-term Temporal Fusion Framework for Multi-view 3D Detection. In *European Conference on Computer Vision (ECCV)*, pp. 131–147, 2024.

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 *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4156–4166, 2025a.

Dubing Chen, Huan Zheng, Jin Fang, Xingping Dong, Xianfei Li, Wenlong Liao, Tao He, Pai Peng, and Jianbing Shen. Rethinking Temporal Fusion with a Unified Gradient Descent View for 3D Semantic Occupancy Prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1505–1515, 2025b.

Junjie Huang and Guan Huang. BEVDet4D: Exploit Temporal Cues in Multi-camera 3D Object Detection. *arXiv preprint arXiv:2203.17054*, 2022.

Junjie Huang, Guan Huang, Zheng Zhu, Ye Yun, and Dalong Du. BEVDet: High-Performance Multi-Camera 3D Object Detection in Bird-Eye-View. *arXiv preprint arXiv:2112.11790*, 2021.

Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Tri-Perspective View for Vision-Based 3D Semantic Occupancy Prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9223–9232, 2023.

Yuanhui Huang, Wenzhao Zheng, Borui Zhang, Jie Zhou, and Jiwen Lu. SelfOcc: Self-Supervised Vision-Based 3D Occupancy Prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19946–19956, 2024a.

Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. GaussianFormer: Scene as Gaussians for Vision-Based 3D Semantic Occupancy Prediction. In *European Conference on Computer Vision (ECCV)*, pp. 376–393, 2024b.

Ziyang Leng, Jiawei Yang, Wenlong Yi, and Bolei Zhou. Occupancy Learning with Spatiotemporal Memory. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 26569–26578, 2025.

Jinke Li, Xiao He, Chonghua Zhou, Xiaoqiang Cheng, Yang Wen, and Dan Zhang. ViewFormer: Exploring Spatiotemporal Modeling for Multi-view 3D Occupancy Perception via View-guided Transformers. In *European Conference on Computer Vision (ECCV)*, 2024.

Wuyang Li, Zhu Yu, and Alexandre Alahi. VoxDet: Rethinking 3D Semantic Scene Completion as Dense Object Detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, 2025a. doi: 10.52202/085713-2707.

Yiming Li, Zhiding Yu, Christopher Choy, Chaowei Xiao, Jose M. Alvarez, Sanja Fidler, Chen Feng, and Anima Anandkumar. VoxFormer: Sparse Voxel Transformer for Camera-Based 3D Semantic Scene Completion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9087–9098, 2023a.

Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. BEVDepth: Acquisition of Reliable Depth for Multi-View 3D Object Detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(2):1477–1485, 2023b. doi: 10.1609/aaai.v37i2.25233.

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*, 2023c. doi: 10.48550/arXiv.2307.01492. Technical report for the 3D Occupancy Prediction Challenge at the CVPR 2023 Workshops.

Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. BEVFormer: Learning Bird's-Eye-View Representation From LiDAR-Camera via Spatiotemporal Transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 47 (3):2020–2036, 2025b. doi: 10.1109/TPAMI.2024.3515454.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1516–1526, 2025.

Yili Liu, Linzhan Mou, Xuan Yu, Chenrui Han, Sitong Mao, Rong Xiong, and Yue Wang. Let Occ Flow: Self-Supervised 3D Occupancy Flow Prediction. In *Proceedings of the 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 2895–2912. PMLR, 2025.

Junyi Ma, Xieyuanli Chen, Jiawei Huang, Jingyi Xu, Zhen Luo, Jintao Xu, Weihao Gu, Rui Ai, and Hesheng Wang. Cam4DOcc: Benchmark for Camera-Only 4D Occupancy Forecasting in Autonomous Driving Applications. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21486–21495, 2024a.

Qihang Ma, Xin Tan, Yanyun Qu, Lizhuang Ma, Zhizhong Zhang, and Yuan Xie. COTR: Compact Occupancy TRansformer for Vision-based 3D Occupancy Prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19936–19945, 2024b.

Riya Mohan, Juana Valeria Hurtado, Rohit Mohan, and Abhinav Valada. ForecastOcc: Vision-Based Semantic Occupancy Forecasting. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2026.

Youshaa Murhij and Dmitry Yudin. OFMPNet: Deep End-to-End Model for Occupancy and Flow Prediction in Urban Environment. *Neurocomputing*, 586:127649, 2024. doi: 10.1016/j.neucom.2024.127649.

Wenzhe Ouyang, Zenglin Xu, Bin Shen, Jinghua Wang, and Yong Xu. LinkOcc: 3D Semantic Occupancy Prediction With Temporal Association. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(2):1374–1384, 2025. doi: 10.1109/TCSVT.2024.3486019.

Pin Tang, Zhongdao Wang, Guoqing Wang, Jilai Zheng, Xiangxuan Ren, Bailan Feng, and Chao Ma. SparseOcc: Rethinking Sparse Latent Representation for Vision-Based Semantic Occupancy Prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15035–15044, 2024.

Xiaoyu Tian, Tao Jiang, Longfei Yun, Yucheng Mao, Huitong Yang, Yue Wang, Yilun Wang, and Hang Zhao. Occ3D: A large-scale 3d occupancy prediction benchmark for autonomous driving. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 64318–64330, 2023. doi: 10.52202/075280-2809.

Arpitsinh Vaghela, Duo Lu, Aayush Atul Verma, Bharatesh Chakravarthi, Hua Wei, and Yezhou Yang. MC-BEVRO: Multi-Camera Bird Eye View Road Occupancy Detection for Traffic Monitoring. In *IEEE International Conference on Intelligent Transportation Systems (ITSC)*, pp. 4143–4149, 2025. doi: 10.1109/ITSC60802.2025.11423511.

Shiming Wang, Holger Caesar, Liangliang Nan, and Julian F. P. Kooij. AsyncBEV: Cross-modal Flow Alignment in Asynchronous 3D Object Detection. In *International Conference on Learning Representations (ICLR)*, 2026. URL `https://openreview.net/forum?id=JINemP2BQP`.

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 *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17850–17859, 2023.

Yuping Wang, Xiangyu Huang, Xiaokang Sun, Mingxuan Yan, Shuo Xing, Zhengzhong Tu, and Jiachen Li. UniOcc: A Unified Benchmark for Occupancy Forecasting and Prediction in Autonomous Driving. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 25560–25570, 2025.

Yuqi Wang, Yuntao Chen, Xingyu Liao, Lue Fan, and Zhaoxiang Zhang. PanoOcc: Unified Occupancy Representation for Camera-based 3D Panoptic Segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17158–17168, 2024.

Yi Wei, Linqing Zhao, Wenzhao Zheng, Zheng Zhu, Jie Zhou, and Jiwen Lu. SurroundOcc: Multi-Camera 3D Occupancy Prediction for Autonomous Driving. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21729–21740, 2023.

Hao Xiang, Zhaoliang Zheng, Xin Xia, Runsheng Xu, Letian Gao, Zewei Zhou, Xu Han, Xinkai Ji, Mingxi Li, Zonglin Meng, Jin Li, Mingyue Lei, Zhaoyang Ma, Zihang He, Haoxuan Ma, Yunshuang Yuan, Yingqian Zhao, and Jiaqi Ma. V2X-Real: A Largs-Scale Dataset for Vehicle-to-Everything Cooperative Perception. In *European Conference on Computer Vision (ECCV)*, pp. 455–470, 2024. doi: 10.1007/978-3-031-72943-0_26.

Lei Yang, Xiaokai Bai, Boqi Li, Chunmian Lin, Li Wang, Ziying Song, Jiahuan Zhang, Enhui Ma, Haibao Yu, Jiaqi Ma, et al. InfraOcc: An Infrastructure Occupancy Benchmark with Static-to-Dynamic Reasoning. *arXiv preprint arXiv:2608.30657*, 2026.

Xiaoqing Ye, Mao Shu, Hanyu Li, Yifeng Shi, Yingying Li, Guangjie Wang, Xiao Tan, and Errui Ding. Rope3D: The Roadside Perception Dataset for Autonomous Driving and Monocular 3D Object Detection Task. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21341–21350, 2022.

Haibao Yu, Yizhen Luo, Mao Shu, Yiyi Huo, Zebang Yang, Yifeng Shi, Zhenglong Guo, Hanyu Li, Xing Hu, Jirui Yuan, and Zaiqing Nie. DAIR-V2X: A Large-Scale Dataset for Vehicle-Infrastructure Cooperative 3D Object Detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21361–21370, 2022.

Shanshuai Yuan, Julong Wei, Muer Tie, Xiangyun Ren, Zhongxue Gan, and Wenchao Ding. LMPOcc: 3D Semantic Occupancy Prediction Utilizing Long-Term Memory Prior from Historical Traversals. *arXiv preprint arXiv:2504.13596*, 2025.

Yunpeng Zhang, Zheng Zhu, and Dalong Du. OccFormer: Dual-path Transformer for Vision-based 3D Semantic Occupancy Prediction. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9433–9443, 2023.

Lianqing Zheng, Long Yang, Qunshu Lin, Wenjin Ai, Minghao Liu, Shouyi Lu, Jianan Liu, Hongze Ren, Jingyue Mo, Xiaokai Bai, Jie Bai, Zhixiong Ma, and Xichan Zhu. OmniHD-Scenes: A Next-Generation Multimodal Dataset for Autonomous Driving. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 48(6):7032–7049, 2026. doi: 10.1109/TPAMI.2026.3663672.

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 (ECCV)*, pp. 55–72, 2024.

APPENDIX

A OVERVIEW

This supplement documents the implementation and controlled evidence underlying RoadOcc. (1) Sections B–B.2 define the fixed-roadside data and velocity targets used to construct historical addresses. (2) Section C specifies the coupled occupancy, motion, and coverage diagnostics used to assess those addresses and their routed sources. (3) Sections E–E.1 provide the scale-wise objectives and exact DCA–VVE–VDSF execution contract. (4) Sections F–H report efficiency, controlled studies, and scope under the complete model configuration.

B ROADSIDE SENSOR CONFIGURATION

InfraOcc reorganizes synchronized V2X-Real streams from object-level cooperative perception into dense roadside occupancy supervision (Yang et al., 2026). Two fixed infrastructure units each carry one LiDAR and two high-resolution cameras under GPS time synchronization. Calibration maps the four camera streams, two point clouds, and occupancy annotations into a common roadside coordinate system. The four cameras are the only exteroceptive model inputs. LiDAR is used solely for benchmark construction and ground-truth generation. Thus the reported setting remains camera-only despite the multimodal data-collection platform.

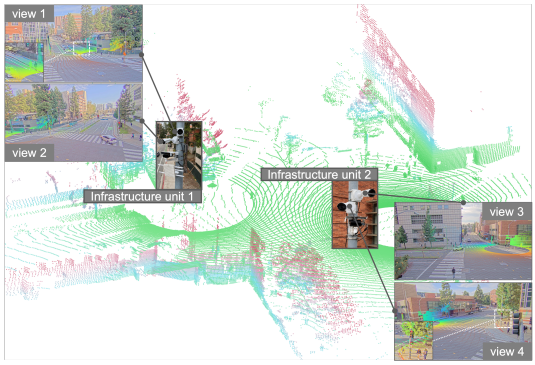


Figure A1: Infrastructure-side sensor configuration. Two fixed units provide four synchronized cameras and two LiDARs calibrated to a common roadside frame.

Roadside benchmarks such as Rope3D and DAIR-V2X also study fixed or cooperative perception (Ye et al., 2022; Yu et al., 2022), but primarily evaluate object-level detection. InfraOcc instead retains a fixed coordinate system for dense occupancy, exposing voxel-level persistence and displacement. This property is essential for testing whether a current voxel retrieves its historical feature from a source-consistent location.

B.1 BENCHMARK SCOPE AND CONSTRUCTION CONTRACT

InfraOcc provides a real-world protocol that combines fixed roadside sensing, temporally continuous dense semantic voxel labels, and camera-only evaluation. Ego-centric occupancy benchmarks assume a moving sensor frame, while existing roadside and cooperative datasets predominantly expose object-level detection or tracking labels. We therefore use the InfraOcc protocol for primary evaluation and complement it with matched P/T/R transfer to two temporal occupancy backbones.

The benchmark contains 290 temporally continuous roadside sequences, with 215 used for training and 75 for evaluation. Each sequence spans approximately 10–24 seconds. Raw streams are captured at 10 Hz and dense occupancy is labeled at 2 Hz, giving the 0.5 s interval used by RoadOcc. Each keyframe contains four calibrated camera views and a semantic occupancy volume over $[-64, 64] \times [-64, 64] \times [-4.8, 1.6]$ m at 0.4 m resolution, or $320 \times 320 \times 16$ voxels. Labels distinguish occupied, free, and unobserved space. Unobserved voxels are excluded from supervision and evaluation.

Occupancy construction separates persistent structure from transient objects. A semantic static scaffold is accumulated from registered vehicle-side LiDAR after dynamic-point removal. Dynamic geometry is reconstructed from infrastructure-side LiDAR using annotated object tracklets, box-guided alignment, and local registration. The two sources are recomposed at each keyframe in the fixed roadside frame and voxelized. Ray traversal labels free space before the first observed surface, while voxels behind surfaces or outside valid rays remain unobserved. These LiDAR sources construct supervision only, and RoadOcc receives no LiDAR feature as model input.

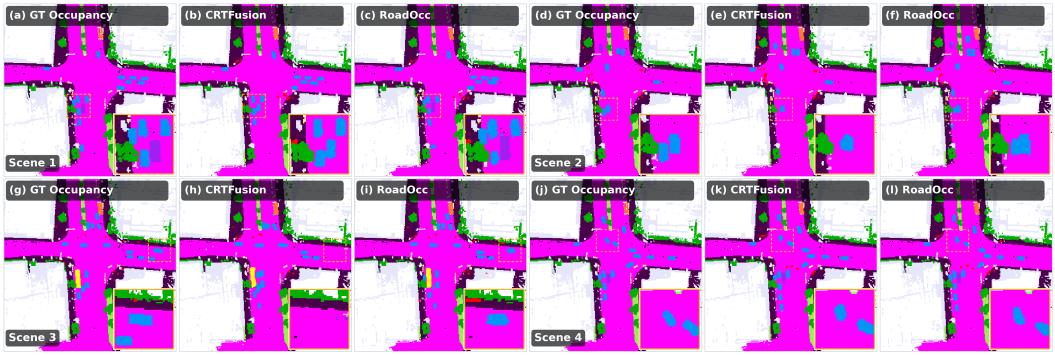


Figure A2: Matched-token occupancy comparison across four scenes. Each triplet shows GT, CRT-Fusion, and RoadOcc in a common BEV frame and semantic palette. The lower-right insets enlarge a representative region that RoadOcc recovers but CRT-Fusion mispredicts.

Qualitative occupancy recovery. Figure A2 compares both methods on identical tokens and coordinates. In the highlighted regions, CRT-Fusion either omits a participant or assigns an incorrect footprint, whereas RoadOcc more closely follows the GT extent and class. The surrounding road structure remains visually unchanged. The gain is therefore concentrated on transient participants rather than obtained by altering the dominant static background.

B.2 Velocity-Target Construction

InfraOcc occupancy labels do not by themselves define the transport field used by RoadOcc. We therefore generate planar voxel-velocity targets from annotated dynamic boxes and their temporal tracklets. Raw box velocity and adjacent-frame tracklet displacement provide two candidates in the fixed roadside frame.

For track j, let $\mathbf{R}_{\mathrm{rs}\leftarrow\mathrm{raw}}$ transform planar vectors from the annotation frame into the fixed roadside frame, and let $\mathbf{c}_{j,t}^{\mathrm{rs}}$ be its roadside-frame center. The two candidate velocities are

$$\widetilde{\mathbf{v}}_{j,t} = \mathbf{R}_{\mathrm{rs}\leftarrow\mathrm{raw}}\mathbf{v}_{j,t}^{\mathrm{raw}}, \qquad \mathbf{d}_{j,t} = \frac{\mathbf{c}_{j,t+\Delta t}^{\mathrm{rs}} - \mathbf{c}_{j,t}^{\mathrm{rs}}}{\Delta t}, \qquad \Delta t = 0.5\ \mathrm{s}. \tag{11}$$

We denote the selected candidate by $\overline{\mathbf{v}}_{j,t}$. A valid tracklet displacement replaces the raw estimate when the raw estimate is invalid, when their speeds differ by at most $10\ \mathrm{m/s}$, or when their speed ratio is at most two. The corrected velocity is rasterized over dynamic occupancy inside each box, whose vertical extent is expanded by a factor of 1.5 to absorb sparse boundary misalignment. Let $\widetilde{B}_{j,t}$ denote this expanded box. The native target assigned to a dynamic voxel is

$$\mathbf{v}_t^{\mathrm{gt}}(\mathbf{x}) = \begin{cases} \mathbf{0}, & \|\overline{\mathbf{v}}_{j,t}\|_2 < 0.8\ \mathrm{m/s}, \\ \mathrm{clip}_{60}(\overline{\mathbf{v}}_{j,t}), & \mathbf{x} \in \widetilde{B}_{j,t} \cap \mathcal{O}_t^{\mathrm{dyn}}, \end{cases} \tag{12}$$

where clip_{60} limits the magnitude to $60\ \mathrm{m/s}$. For candidates no faster than 2 m/s, a one-step (0.5 s) future-occupancy check compares zero-motion and transported same-class support. The check is applied only with at least eight support voxels. Zero motion is selected when its score exceeds the transported score by at least 0.05. Finally, face-connected component repair propagates valid motion within the corresponding raw dynamic occupancy component, and targets are generated at native, $1/2$, $1/4$, and $1/8$ scales.

Future occupancy is used only to construct offline supervision and evaluation targets. Neither future images nor future features are provided to RoadOcc at training or inference time. The two velocity thresholds serve different roles. The 0.8 m/s threshold denoises the physical target, whereas $\epsilon = 10^{-3}$ m/s is the numerical zero test used by the P/T/R rule after target construction.

C Evaluation Contract

Aggregate occupancy is dominated by persistent structure and can therefore conceal missed or stale traffic participants. Standard protocols emphasize semantic and geometric IoU (Wang et al., 2023;

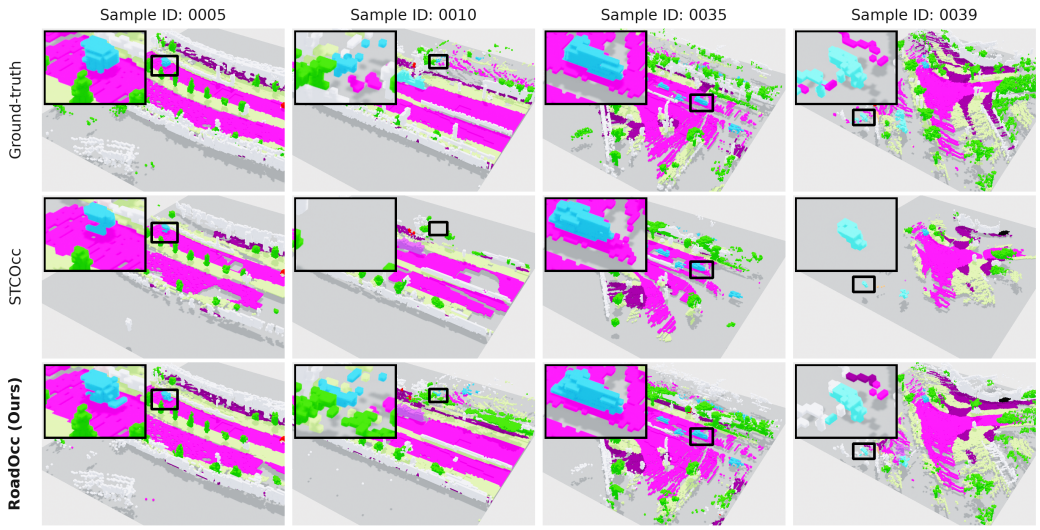


Figure A3: **Qualitative comparison on Occ3D-nuScenes.** We show four validation frames under the official camera-visible mask, semantic palette, voxel grid, and viewing transform. Black boxes and the upper-left insets show the same GT dynamic component in all three rows. Each region is selected where RoadOcc recovers more of that component than the released STCOcc checkpoint.

Tian et al., 2023), whereas dynamic occupancy and flow require motion-aware diagnostics (Ma et al., 2024a; Liu et al., 2025). Table A1 formalizes our evaluation contract on InfraOcc (Xiang et al., 2024; Yang et al., 2026). Dynamic and next-frame IoU measure semantic recovery at the current and flow-advected next-frame locations. Direct MAVE and DSR jointly measure motion accuracy and coverage. Static, geometric, and ray metrics measure whether dynamic recovery preserves the fixed scene. Together, these diagnostics distinguish improved correspondence from reduced dynamic support or retained stale evidence.

Dynamic-object comparison. From left to right, RoadOcc raises recall within the highlighted GT component from 58.5 to 92.5, 0.0 to 87.5, 76.4 to 92.7, and 59.0 to 90.2%, respectively. The insets show the corresponding recovery of missed or truncated traffic participants.

Table A1: Metric and diagnostic scope.

Metric	Support	Failure	Role
Dyn. mIoU ↑	Dynamic GT voxels	Omission / class error	Dynamic semantic recovery
Static / gIoU ↑	Static / occupied support	Static-structure damage	Static-scene guardrail
Next Dyn. ↑	Advected dynamic GT	Displacement error	Next-frame consistency
Direct MAVE ↓	All dynamic GT voxels	Miss-induced motion bias	Unconditional flow error
TP-MAVE ↓ / DSR ↑	Matched semantics / dynamic GT	Conditional-match bias	Matched accuracy / coverage
RayIoU ↑	Visible evaluation rays	Surface-depth mismatch	Geometric fidelity

Direct MAVE evaluates every valid GT dynamic voxel, whereas TP-MAVE is conditional on a correct semantic match and must therefore be interpreted with DSR. A valid routing improvement should increase dynamic IoU and DSR, decrease Direct MAVE, and preserve the static and geometric guardrails. Reporting these quantities together distinguishes improved semantic correspondence from a lower conditional error caused by reduced semantic support. For completeness, let $\mathcal{D}$ be the set of evaluated dynamic classes, $V_c = \{i : y_i = c\}$ the valid GT voxels of class c, and $V_c^{\mathrm{TP}} = \{i \in V_c : \widehat{y}_i = c\}$. Classes without valid support are omitted from the corresponding class macro-average. The sets $\mathcal{D}'$ and $\mathcal{D}''$ denote the supported classes for Direct and TP-MAVE, respectively. The two velocity metrics and semantic recall are

$$\text{Direct MAVE} = \frac{1}{|\mathcal{D}'|} \sum_{c \in \mathcal{D}'} \frac{1}{|V_c|} \sum_{i \in V_c} \|\widehat{\mathbf{v}}_i - \mathbf{v}_i\|_2, \tag{13}$$

$$\text{TP-MAVE} = \frac{1}{|\mathcal{D}''|} \sum_{c \in \mathcal{D}''} \frac{1}{|V_c^{\mathrm{TP}}|} \sum_{i \in V_c^{\mathrm{TP}}} \|\widehat{\mathbf{v}}_i - \mathbf{v}_i\|_2. \tag{14}$$

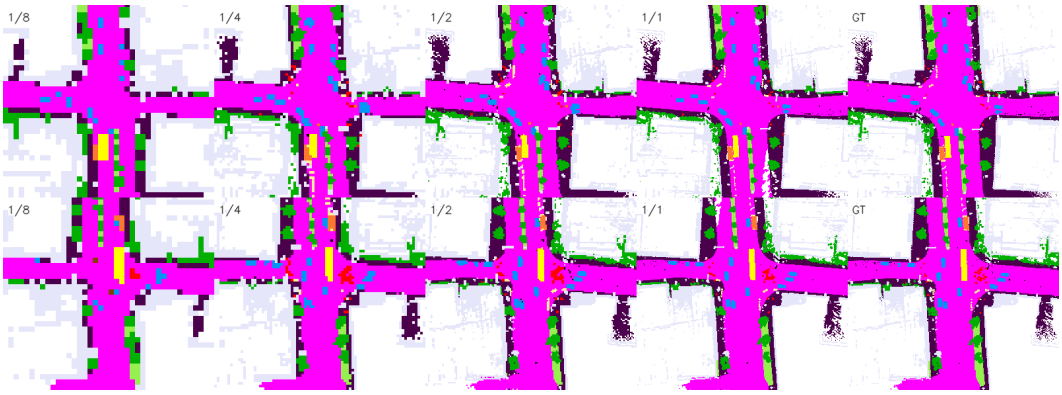


Figure A4: Coarse-to-fine occupancy decoding for two representative samples. Columns show predictions from $1/8$ scale through full resolution and the corresponding GT. Successive stages sharpen road boundaries and recover traffic participants that remain unresolved at coarse resolution.

Thus MAVE is class-macro while DSR is a micro recall over GT dynamic voxels. For next-frame dynamic IoU, centers of predicted current dynamic voxels are advanced by the predicted planar velocity for 0.5 s, transformed to the next roadside frame, and voxelized. Collisions use deterministic per-class majority assignment before class-macro IoU is computed against next-frame GT. This flow-advected diagnostic is distinct from the same-class radius-one hit score used only in the backwarp visualization.

$$\mathrm{DSR} = \frac{\sum_{c\in\mathcal{D}} |V_c^{\mathrm{TP}}|}{\sum_{c\in\mathcal{D}} |V_c|}. \tag{15}$$

D Occ3D-nuScenes Transfer

Table A2 compares representative camera-only temporal methods under the mask-trained Occ3D-nuScenes protocol (Tian et al., 2023; Li et al., 2023c; Chen et al., 2025a; Li et al., 2024; Ma et al., 2024b; Liao et al., 2025). Dynamic mIoU averages bicycle, bus, car, construction vehicle, motorcycle, pedestrian, trailer, and truck, while static mIoU covers the other nine occupied classes. Overall scores are quoted from the papers. For FB-Occ, ViewFormer, COTR, and STCOcc, group means are recomputed from the per-class IoUs in Liao et al. (2025), Table 2. ALOcc's 35.40 dynamic mIoU is reported directly. its 46.73 static mIoU is computed from the corresponding per-class row in Chen et al. (2025a), Table A.6. The source tables round values to one decimal place, so recomputed group means, displayed to two decimals, need not exactly recover the separately reported overall mIoU under $(8\,\mathrm{Dyn.} + 9\,\mathrm{Sta.})/17$. RoadOcc improves over STCOcc by 0.41 overall, 0.79 dynamic, and 0.06 static mIoU points. These results support transfer to an ego-centric benchmark, with the larger improvement in dynamic classes.

Table A2: Mask-trained comparison on Occ3D-nuScenes. All metrics are percentages.

Method	BB	Res.	mIoU	Dynamic mIoU	Static mIoU
FB-Occ	R50	256×704	39.10	33.75	43.89
ViewFormer	R50	256×704	41.90	35.00	47.93
COTR	R50	256×704	44.50	38.60	49.63
ALOcc	R50	256×704	41.40	35.40	46.73
STCOcc	R50	256×704	44.60	39.01	49.58
RoadOcc	R50	256×704	**45.01**	**39.80**	**49.64**

E Training Objectives

RoadOcc assigns each loss to one decision in the DCA–VVE–VDSF chain. Camera-depth supervision uses weight 0.5. Semantic occupancy is supervised from full resolution to $1/8$ scale with weights $[1.0, 0.5, 0.25, 0.125]$, while the static, dynamic, and DCA-candidate auxiliaries each use 0.5. VVE receives class- and speed-balanced robust flow supervision at all four scales with weight 0.1. Its local-connectivity term also uses 0.1 and starts after epoch 3 to limit noisy early correspondences. The P/T/R route loss spans the native, $1/2$, $1/4$, and $1/8$ scales with relative weights

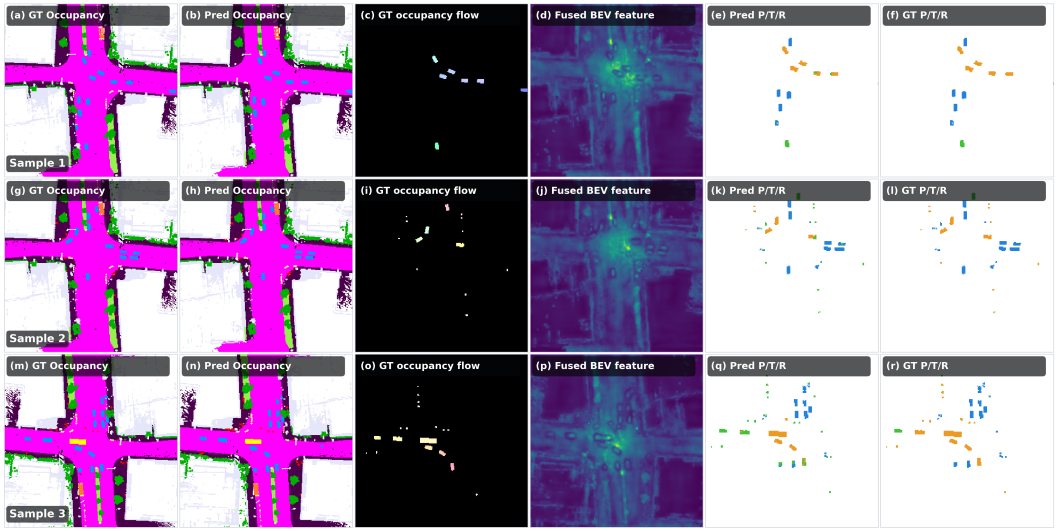


Figure A5: Native-scale P/T/R routing across three samples. Predicted and GT routes share GT dynamic support. Blue, orange, and green denote Persist, Transport, and Refresh, respectively.

$[1, 0.5, 0.25, 0.125]$, normalized under $\lambda_{\text{route}} = 0.5$. Each scale constructs the semantic-support target using its physical cell size. Equal-class-weight cross-entropy is augmented from epoch 2 by a soft-overlap term of weight 0.5 with state weights $[1, 1, 0.25]$. Each executed routing field therefore receives resolution-matched supervision.

E.1 Candidate, Motion, and Routing Contract

DCA score. Let $\boldsymbol{\pi}_{\text{cur}}$ and $\boldsymbol{\pi}_{\text{sta}}$ denote current and static-reference class probabilities, and let $\mathcal{C}^{+}_{\text{sta}}$ contain the static classes together with empty class. The executed change and dynamic scores are

$$\delta(\mathbf{x}) = \frac{1}{|\mathcal{C}^{+}_{\text{sta}}|} \sum_{c \in \mathcal{C}^{+}_{\text{sta}}} |\pi_{\text{cur},c}(\mathbf{x}) - \pi_{\text{sta},c}(\mathbf{x})|, \tag{16}$$

$$p_{\text{dyn}}(\mathbf{x}) = \sum_{c \in \mathcal{C}_{\text{dyn}}} \pi_{\text{cur},c}(\mathbf{x}), \tag{17}$$

$$c(\mathbf{x}) = \frac{\max\{\delta(\mathbf{x}), p_{\text{dyn}}(\mathbf{x})\}}{\max_{\mathbf{y}} \max\{\delta(\mathbf{y}), p_{\text{dyn}}(\mathbf{y})\} + 10^{-6}}. \tag{18}$$

DCA ranks this normalized score. Before fixed top-K selection, VDSF uses $\text{clip}(c + 0.5e, 0, 1)$, where e is nonempty support and hence $\eta = 0.5$ in Eq. 6. For an anchor $\mathbf{r} = (\mathbf{x}, z_{\mathbf{r}})$, $c(\mathbf{r})$ is the score of the corresponding voxel-height bin and is shared across cameras before visibility masking. Let $d_{n,\mathbf{r}}$ be its projected depth, $\widehat{d}_n$ the winning predicted depth bin, and $\Delta_d = 1$ m the bin width. The executed depth-consistency factor is

$$\beta_{n,\mathbf{r}} = \exp\left[-\frac{\sigma^2}{2} \min_{\xi \in \{-\Delta_d, +\Delta_d\}} \left| d_{n,\mathbf{r}} - (\widehat{d}_n + \xi) \right|^2\right], \qquad \sigma = 2. \tag{19}$$

Deformable-attention weights are first softmax-normalized over sampled points, then multiplied by β. They are not renormalized afterward, and camera outputs are averaged over cameras retaining at least one visible anchor.

Recursive occupancy decoding. Figure A4 shows the progressive output of the multi-scale decoder. Coarse predictions establish scene layout, while finer stages restore object boundaries and sparse traffic participants before the full-resolution output.

VVE correspondence range. At scales $1/8$, $1/4$, and $1/2$, the 5×5 correlation volume ($r = 2$) spans $\pm$ 6.4, $\pm$ 3.2, and $\pm$ 1.6 m on the native 0.4 m grid. Over the 0.5 s interval, the equivalent speed ranges are twice these values.

Each correlation operand is normalized to unit ℓ_2 norm with $\epsilon = 10^{-6}$. The resulting volume has 25 channels and is concatenated with current and historical BEV embeddings, processed by two 3×3 convolution blocks, and decoded as a two-dimensional velocity residual in m/s. The window bounds only the explicit matching evidence. The coarse velocity is added to each finer prediction and the final continuous residual is not clipped, so it does not impose a hard bound on predicted speed. The GT-flow study in the main paper diagnoses constructed transport targets across speed ranges. In the temporal blend, the history branch receives weight $0.5\,\mathbf{1}_{\text{hist}}\max_z p_{\text{dyn}}(\mathbf{x}, z)$. This is current dynamic support gated by history validity, rather than a calibrated measure of historical correspondence.

Native P/T/R target. For a current dynamic voxel, the Transport source is

$$\mathbf{x}_{\text{src}} = \text{round}\left(\mathbf{x} - \frac{\Delta t\,\mathbf{v}_t(\mathbf{x})}{0.4\text{ m}}\right), \qquad \Delta t = 0.5\text{ s}. \tag{20}$$

Persist tests the fixed-coordinate address and Transport tests $\mathbf{x}_{\text{src}}$. Support means that the nearest previous GT contains the same semantic class at the same height within a 3×3 XY neighborhood (L_∞ radius one). Support is Boolean. One or more matching voxels suffice, with no instance-level tie-breaking. Equation 7 tests the fixed address for stationary targets and the Transport address for moving targets. If the motion-selected candidate is out of grid or unsupported, the target is Refresh. Thus support at the alternative address does not override the motion condition. These GT-derived labels supervise source preferences without requiring instance identities.

Coarse-to-fine FIFO routing. At each decoder scale, a shared controller projects current features and the nearest fixed-coordinate history to eight channels, adds embeddings of detached VVE flow (normalized by 10 m/s) and detached upsampled coarse P/T/R, and predicts three route logits after a $3 \times 3 \times 3$ refinement. A lightweight native head combines upsampled $1/2$ context, native flow, and the finest P/T/R distribution for full-resolution routing. Each scale shares its route distribution across valid FIFO slots, whose ages and cached VVE fields determine source-consistent inverse-trajectory Transport addresses. Routed slots are concatenated before fusion. Invalid slots are masked and queues reset at scene boundaries. Fusion uses soft routes, with argmax only for diagnostics.

F Capacity and Efficiency Context

Table A3 reports parameters, synchronized model latency measured on a single NVIDIA RTX A6000 with batch size 1, and occupancy accuracy for the reproduced temporal configurations. RoadOcc contains more parameters, but its reported configuration reduces latency from 333.44 to 253.25 ms and improves all-class and dynamic mIoU by +4.44 and +4.71 points. DCA, VVE, and P/T/R operate on sparse or low-resolution features. The temporal path also uses shorter histories and smaller token budgets. This is a comparison of complete configurations, rather than an isolated measurement of routing overhead.

Table A3: Accuracy–efficiency comparison.

Setting	Params.↓	Lat.↓	All↑	Dyn.↑
STCOcc '25	60.47	333.44	60.85	27.66
RoadOcc (**ours**)	78.69	253.25	65.29	32.37

Class-wise improvement. Figure A6 resolves the official result in Table 2 into dynamic and static classes. Relative to STCOcc, RoadOcc improves five of the six dynamic classes, led by Bike at +10.47 and Bus at +6.68, while Pedestrian changes by −0.19. Every static class also improves. The gains therefore extend across dynamic participants and persistent structure, although their magnitude varies by category.

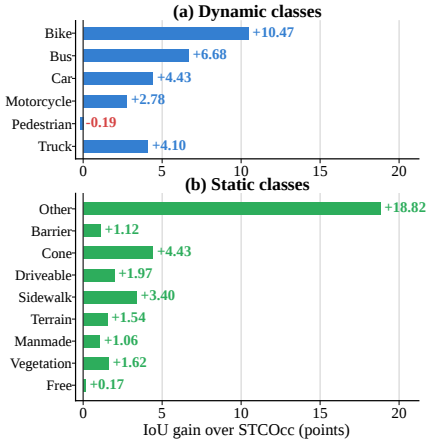


Figure A6: Class-wise IoU gains.

G Controlled Design Studies

All studies hold the backbone, data protocol, optimizer, training schedule, scene-boundary queue reset, and VDSF configuration fixed. Unless varied explicitly, the decoder stages use $K = [2000, 512, 128]$ sparse tokens and retain $[8, 4, 2]$ history samples, respectively.

Claim–evidence map. Table 4 tests the component chain, Table 5 isolates VVE-addressed reading, and Table 6 isolates supervised source execution and Refresh. Tables A4–A7 cover temporal gaps, history-reuse routes, token saturation, and historical-address configurations.

Temporal-gap zero-shot generalization. Table A4 evaluates checkpoints trained only at the 0.5 s cadence without fine-tuning. Phase-offset streams retain all 1,970 validation inputs, while the same 1,520 anchors with valid past and future 1.5 s context are scored at every gap. The elapsed time is recovered from timestamps and used by temporal transport.

Table A4: Temporal-gap zero-shot evaluation on matched anchors. Both checkpoints are trained only at a 0.5 s interval.

	Dyn. ↑			Nxt. ↑		
Method	0.5 s	1.0 s	1.5 s	0.5 s	1.0 s	1.5 s
STCOcc (Liao et al., 2025)	28.04	28.26	28.58	25.21	20.90	19.12
RoadOcc (**ours**)	**33.71**	**33.57**	**33.49**	**30.23**	**25.18**	**22.78**

RoadOcc preserves a 4.91-point Dyn. advantage when the test gap triples. Nxt. is 22.78 versus 19.12. Elapsed time scales the transport displacement, and the complete model remains effective at the longer intervals tested here. This comparison evaluates temporal-gap generalization of the full system, without isolating the contribution of P/T/R.

VVE training design. The main-paper control establishes the value of reading from a corrected address. Figure A7 follows a progressive sequence that introduces multi-scale flow, candidate-map guidance, connectivity regularization, and flow-addressed history retrieval. Dyn. improves monotonically while Direct MAVE decreases across this progression. Overall, Dyn. rises by +3.63 and Direct MAVE falls by 0.271. The final flow-addressed read contributes another +1.65 Dyn. while Direct MAVE decreases by only 0.021, indicating that its main benefit comes from retrieving useful history.

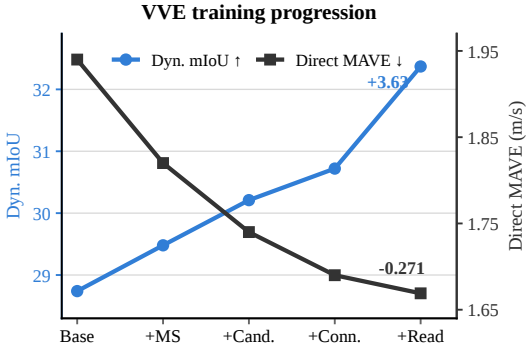


Figure A7: Dynamic occupancy and motion-error progression under the VVE training signals.

Routing-control contract. Table 6 fixes DCA, VVE, VDSF, candidate sources, token and history budgets, optimizer, schedule, and seeds. Gate mixes the same three candidates without route supervision. State adds P/T/R supervision but does not use its probabilities in fusion, while Full uses the supervised soft routes. Full w/o R retains Full, masks α_R, and renormalizes (α_P, α_T).

History-reuse route analysis. Table A5 audits the two routes that reuse historical evidence on valid GT dynamic voxels. Persist emphasizes precise reuse at the fixed coordinate, while Transport provides broader coverage after displacement. The current-only fallback is evaluated end to end by Full versus Full w/o R in Table 6.

Table A5: History-reuse route statistics on valid GT dynamic voxels. Percentages are computed against the full routing population.

State	Target	Pred.	Precision	Recall
Persist	52.34	38.84	95.50	70.87
Transport	44.30	50.40	72.98	83.03

Table A6: Sensitivity to sparse-token allocation.

Budget K	Dyn. ↑	Nxt. ↑	dM ↓	DSR ↑
$[1000, 256, 64]$	31.74	30.96	1.716	60.28
$[2000, 512, 128]$	32.37	31.72	1.669	61.11
$[4000, 1024, 256]$	32.39	31.69	1.665	61.18

Qualitative routing behavior. Figure A5 relates each route to observable scene evidence. Persist coincides with aligned objects, Transport follows nonzero displacement, and Refresh favors current evidence when the motion-selected historical candidate lacks support in the training target. Full versus Full w/o R in Table 6 measures the end-to-end contribution of the current-evidence fallback.

Sparse-token budget. VDSF processes only the top-K selected voxels. Table A6 tests whether the reported behavior depends on a larger compute budget or on identifying the right locations. All other model and routing settings remain fixed while K is scaled at every decoder stage. Increasing

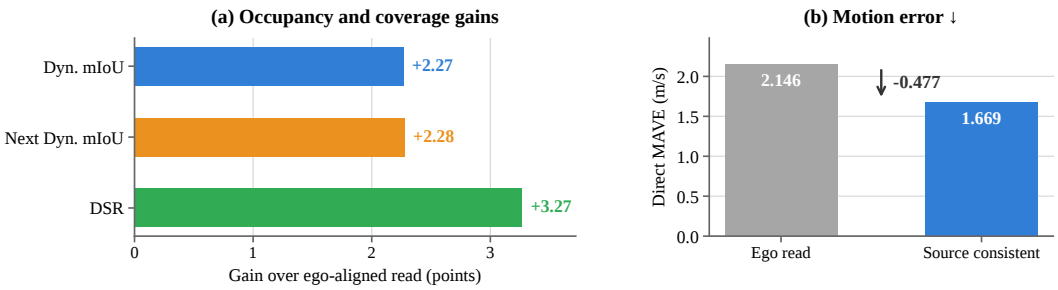


Figure A8: Measured gains of the complete source-consistent configuration over the ego-aligned reference. Values are derived from the configurations in Table A7.

the small allocation to the default improves Dyn. by 0.63 and DSR by 0.83. Doubling the default allocation changes Dyn. by only 0.02, dMAVE by 0.004, and DSR by 0.07, while next-frame IoU fluctuates by -0.03. This saturation shows that the reported result does not rely on continually expanding the sparse-token budget.

Velocity-address construction. Table A7 summarizes a fixed-coordinate read, target-vector backtracing, and the complete source-consistent configuration. The fixed-read row repeats the reference in Table 5, while the full row reports RoadOcc. These results describe retrieval configurations alongside the main-paper routing controls and flow-based occupancy methods (Liu et al., 2025; Liao et al., 2025).

Target-vector backtracing improves Dyn. by 1.31 and reduces dMAVE by 0.442 over the ego-aligned read. The complete configuration is higher by another 0.96 Dyn., 1.06 Nxt., and 1.29 DSR, with dMAVE lower by 0.035. The results relate the retrieval configurations to occupancy, coverage, and motion error. They do not directly measure the semantic validity of each transported feature.

Table A7: Historical-address configurations.

Address	Dyn. ↑	Nxt. ↑	dM ↓	DSR ↑
Ego-aligned read	30.10	29.44	2.146	57.84
Target-vector trace	31.41	30.66	1.704	59.82
Source-consistent	32.37	31.72	1.669	61.11

Stepwise address benefit. Figure A8 summarizes the difference between the fixed-read reference and the complete configuration in Table A7. Dyn., Next Dyn., and DSR rise by 2.27, 2.28, and 3.27 points, respectively, while Direct MAVE falls by 0.477.

H Implementation and Scope

The production configuration uses four cameras, a ResNet-50 backbone, three decoder stages, a 0.4 m voxel grid, and a 0.5 s sampling interval. Models are optimized with AdamW for 24 epochs. Controlled variants share the same initialization policy, optimizer, schedule, scene-boundary queue reset, and InfraOcc evaluator. Only the designated factor is changed.

P/T/R operates on semantic voxel support, matching the InfraOcc label space. Nearby same-class objects can share support within the local radius. Refresh is supervised when the historical candidate selected by the target motion state lacks support. It uses the current-frame feature. The labels encode a semantic source preference, not instance identity or guaranteed validity at a predicted address. Nearest-history supervision and a shared route across FIFO slots also leave history-age-specific source selection unresolved. We report dynamic IoU, Direct MAVE, and DSR jointly to assess semantic recovery, motion error, and coverage beyond the GT-conditioned route diagnostic.