VGGT-OCC: GEOMETRY-GUIDED REASONING ACROSS VIEWS AND SCALES FOR 3D SEMANTIC OCCUPANCY PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

3D semantic occupancy prediction requires consistent reasoning across camera views, temporal observations, and spatial scales. Yet reference-point projection alone does not carry camera geometry into subsequent sampling, weighting, and cross-camera aggregation. High-resolution decoding must also balance coarse 3D context with the local detail needed for fine-scale prediction. We present VGGT-Occ, a geometry-guided framework for vision-only 3D semantic occupancy prediction across views, time, and scales. Across views and time, Projection-Aware Deformable Attention (PA-DA) carries projection geometry through three stages: it projects learned 3D offsets into each view, applies a Jacobian-based attention bias, and predicts channel-wise fusion weights conditioned on semantic features, projection geometry, and temporal context to aggregate observations across cameras and time. Across scales, a sequential coarse-to-fine decoder concentrates cross-view attention at coarse resolutions and propagates coarse 3D context through gated fusion to guide high-resolution refinement. VGGT-Occ outperforms the compared state-of-the-art methods in semantic mIoU across all three benchmarks. It achieves 37.24% IoU / 25.36% mIoU on SurroundOcc-nuScenes, 44.47% mIoU on Occ3D-nuScenes, and 45.06% IoU / 18.13% mIoU on SSCBench-KITTI-360. Code will be released publicly.

1 INTRODUCTION

3D semantic occupancy prediction reconstructs the surrounding environment as a dense voxel grid with fine-grained semantic labels, providing comprehensive geometric and semantic scene understanding for autonomous driving (Zhang et al., 2026; Xu et al., 2025; Wei et al., 2023; Tian et al., 2023). Recovering this representation from multi-camera images requires consistent reasoning across views and spatial scales to align observations, recover global 3D structure, and preserve local detail at object boundaries and small structures.

Most query-based multi-view occupancy methods use camera geometry to project 3D reference points onto image feature maps. However, using geometry only at this initial step leaves three gaps in the 2D-to-3D lifting process. First, offsets learned in the image plane do not describe a shared 3D displacement across cameras, which can produce inconsistent sampling directions. Sparse4D (Lin et al., 2022) and Far3D (Jiang et al., 2024) address this issue with projected 3D offsets, but consistent sampling alone does not determine how observations should be weighted and fused. Second, query-only attention weights do not explicitly account for local projection sensitivity, although the same 3D displacement can produce different image-plane changes across views. Third, averaging features across visible cameras assigns equal weight to each observation, without distinguishing their projection geometry or temporal relevance.

Scale changes also affect the balance between global structure and local detail. Coarse voxels describe global layout and broad semantic regions, while high-resolution voxels resolve object boundaries and small structures. The number of voxels grows rapidly at fine resolutions, making repeated cross-view lifting increasingly expensive. At the same time, upsampling and direct addition transfer coarse information across scales, but do not provide a learned mechanism to balance it against current-scale

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

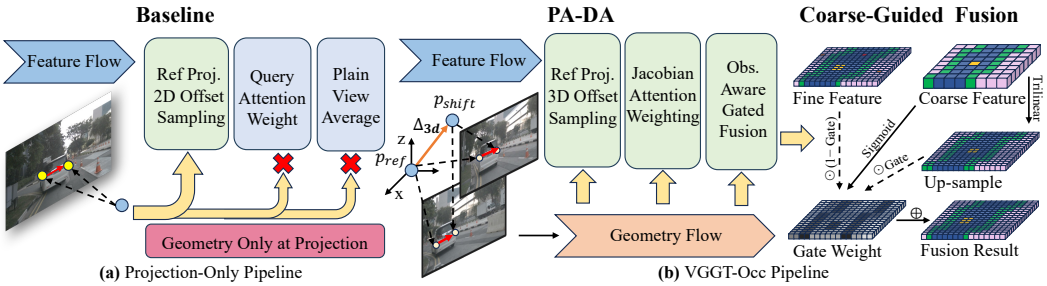

Figure 1: **VGGT-Occ overview.** (a) The baseline projects reference points, predicts 2D offsets and attention weights from queries, and averages visible camera features uniformly. (b) PA-DA carries projection geometry through sampling, weighting, and observation fusion; coarse-guided gated fusion propagates 3D context across scales.

features. A suitable decoder should focus cross-view reasoning on coarse scales, then use coarse 3D context to guide lightweight refinement at the fine scale while preserving complete voxel coverage.

In this paper, we present **VGGT-Occ**, a geometry-guided framework for 3D semantic occupancy prediction (Figure 1). It follows two complementary design choices. First, projection geometry is used throughout sampling, attention weighting, and observation fusion. Second, cross-view reasoning is concentrated at coarse scales, whose 3D context guides high-resolution refinement through gated fusion.

To implement the first choice, we adopt a geometry-grounded Transformer (Wang et al., 2026) to jointly encode the multi-view images and produce features with cross-view context. Within the occupancy head, we introduce **PA-DA (Projection-Aware Deformable Attention)**, which carries projection geometry through cross-attention. It learns 3D offsets and projects them into each camera view for cross-view-consistent, distance-adaptive sampling. A Jacobian-based bias then adjusts the attention weights. Finally, **Observation-Aware Gated Fusion** predicts channel-wise weights from the observation features, projection geometry, and temporal context to combine observations across cameras and time.

For the second choice, we use a sequential coarse-to-fine decoder with **Coarse-Guided Gated Fusion**. A single coarse embedding is refined by successive patch splitting layers, so each scale passes its 3D context to the next. Cross-attention is concentrated at the two coarse scales (10K and 80K voxels), while the fine scale (640K voxels) uses local depthwise convolutions. At each transition, channel-wise gates predicted from coarse features are upsampled to balance coarse context and current-scale features. Depthwise convolution then refines the fused features.

Our main contributions are:

1. We propose **PA-DA**, which carries projection geometry beyond reference-point initialization through cross-view-consistent 3D sampling, Jacobian-guided attention weighting, and observation-aware fusion.
2. We design a **sequential coarse-to-fine decoder with coarse-guided gated fusion**, which uses coarse 3D context to modulate refinement across scales while avoiding cross-view attention on the finest grid.
3. We evaluate VGGT-Occ on SurroundOcc-nuScenes, Occ3D-nuScenes, and SSCBench-KITTI-360. On SurroundOcc-nuScenes, it achieves 37.24% IoU and 25.36% mIoU, leading the compared methods in semantic mIoU (Table 1). Ablations support PA-DA across two backbones and the benefit of explicit coarse guidance.

2 Related Work

Geometry in Multi-View 3D Perception. Camera geometry is fundamental to 2D-to-3D correspondence. Early view transformation methods such as OFT (Roddick et al., 2019), LSS (Philion & Fidler, 2020), and BEVDet (Huang et al., 2021) use camera calibration to lift image features into

3D. Later methods incorporate geometry through projected queries (Li et al., 2022), 3D positional embeddings (Liu et al., 2022a), and multi-view voxel lifting (Wei et al., 2023; Zhang et al., 2023; Li et al., 2023c; Wang et al., 2024). Ray-based Gaussian initialization (Huang et al., 2025) and adaptive superquadric representations (Zuo et al., 2025a) provide alternative geometric priors. Depth and 3D foundation models (Yang et al., 2024; Wang et al., 2025) provide transferable geometric priors for tasks including place recognition (Deng et al., 2025a), relocalization (Deng et al., 2025b), and occupancy prediction (Zhu et al., 2026). Spatial correspondence determines where to sample, but not how strongly each observation should contribute. VGGT-Occ uses a geometry-grounded encoder and carries projection geometry through sampling, attention weighting, and observation fusion.

Deformable Attention for 3D Perception. Deformable DETR (Zhu et al., 2021) introduced sparse feature sampling with learned 2D offsets. DETR3D (Wang et al., 2021) projects 3D queries into camera views, while BEVFormer (Li et al., 2022) aggregates multi-view features into BEV queries through deformable attention. Sparse4D (Lin et al., 2022) combines projected 4D keypoint sampling with learned aggregation across temporal features, while Far3D (Jiang et al., 2024) uses learned 3D offsets for image sampling. GaussianFormer3D (Zhao et al., 2026) applies 3D deformable attention to LiDAR-camera fusion. Building on projected 3D sampling, PA-DA couples cross-view-consistent offsets with a Jacobian-based attention bias and observation-aware gated fusion. Its fusion replaces uniform averaging across visible cameras with per-observation, per-channel gates conditioned on feature semantics, projection geometry, and temporal context.

3D Occupancy Prediction. Early dense methods (Cao & de Charette, 2022; Wei et al., 2023; Zhang et al., 2023) recover semantic voxel grids through multi-scale 3D features; subsequent works advance view transformation (Li et al., 2023c; Wang et al., 2024) and dual representations (Shi et al., 2026). FlashOcc (Yu et al., 2023) accelerates prediction via a channel-to-height plugin, while RenderOcc (Pan et al., 2024) and GaussRender (Chambon et al., 2025) add rendering-based supervision. SparseOcc (Liu et al., 2024) uses sparse voxel and mask decoding, while Gaussian-based (Huang et al., 2024b; 2025; Zuo et al., 2025b) and superquadric (Zuo et al., 2025a) methods represent scenes with compact primitives. Gaussian splatting (Kerbl et al., 2023) also supports sparse-query pretraining (Zhang et al., 2025). SelfOcc (Huang et al., 2024a) explores self-supervision via neural rendering, and Cam4DOcc (Ma et al., 2024a) extends occupancy to 4D forecasting. CTF-Occ (Tian et al., 2023) combines token-selected cross-attention with interpolation across pyramid levels, while ProtoOcc (Oh et al., 2025) enriches low-resolution queries with image prototypes before multi-perspective decoding. VGGT-Occ concentrates cross-view attention at coarse scales and uses coarse context both to initialize finer grids and to regulate feature fusion. Beyond patch splitting, a separate coarse branch predicts gates that balance coarse context against current-scale features, retaining dense coverage with lightweight fine-scale refinement.

3 METHOD

3.1 OVERVIEW

VGGT-Occ predicts a dense semantic occupancy grid in a reference coordinate frame from images captured by N cameras over S temporal steps. Each observation is indexed by (s, n), with a signed time offset $\Delta t_{s,n} = t_{s,n} - t_{\text{ref}}$ between its capture time and the occupancy reference time.

In Figure 2, the unified encoder jointly processes the input images using a geometry-grounded Transformer (Wang et al., 2026), producing multi-scale image features with cross-view context. PA-DA (Section 3.2) lifts these features into the coarse voxel grids. The resulting 3D features pass through a sequential coarse-to-fine hierarchy, where patch splitting initializes each finer scale from the preceding scale's output. Coarse-guided gated fusion (Section 3.3) combines the coarse context with current-scale features, and the finest scale uses local convolutions rather than cross-view attention.

3.2 PA-DA: PROJECTION-AWARE DEFORMABLE ATTENTION

In the baseline image-to-voxel lifting pipeline, deformable cross-attention (Zhu et al., 2021) projects the reference point of each voxel query onto image feature maps and predicts 2D sampling offsets and attention weights from the query. The resulting features are averaged across visible camera

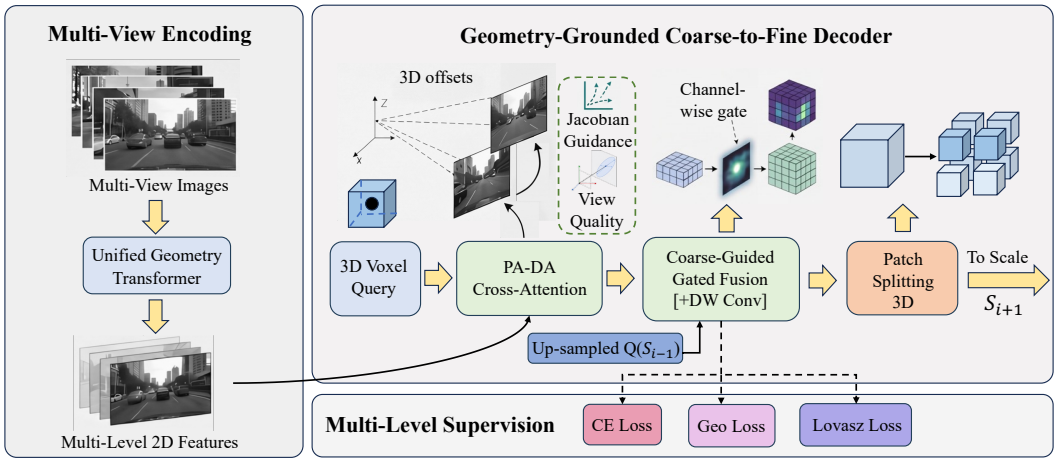

Figure 2: **VGGT-Occ architecture.** Unified geometry encoding produces multi-scale image features. PA-DA performs geometry-guided cross-view lifting at S_0 and S_1; S_2 uses local convolutions. Coarse-guided gated fusion combines current-scale features with the preceding scale's output $Q(S_{i-1})$, while patch splitting initializes S_{i+1}.

observations. PA-DA carries projection geometry through this pipeline (Figure 3): shared 3D offsets define sampling locations, a Jacobian-based bias adjusts pointwise attention weights, and observation-aware gated fusion combines camera-time features using projection geometry, semantic features, and temporal context.

3.2.1 3D Offset Learning with Projection

To resolve the cross-view inconsistency of 2D offset learning, we predict 3D offsets $\mathbf{\Delta}_{3d}^{(h,l,k)} \in \mathbb{R}^3$ from the query feature via a linear projection, with one offset per attention head h, feature level l, and sampling point k. The offsets are shared across camera-time observations. Given a query's reference point $\mathbf{p}_{\text{ref}}$ in the occupancy coordinate frame, each shifted point is projected into observation (s, n):

$$\mathbf{p}_{\text{shift}}^{(h,l,k)} = \mathbf{p}_{\text{ref}} + \mathbf{\Delta}_{3d}^{(h,l,k)}, \qquad (u_{s,n}^{(h,l,k)}, v_{s,n}^{(h,l,k)}) = \pi_{s,n}(\mathbf{p}_{\text{shift}}^{(h,l,k)}), \tag{1}$$

where $\pi_{s,n}$ uses camera intrinsics $\mathbf{K}_{s,n}$ and the transform $\mathbf{E}_{s,n}$ from the occupancy reference frame to the camera frame at capture time. It includes the image-coordinate transformation and returns normalized image coordinates. These projected positions are the sampling locations, without additional per-view 2D offsets. Each (h, l, k) therefore defines a shared 3D displacement across views, while the corresponding image-plane displacement varies with viewing geometry and depth.

3.2.2 Jacobian-Guided Attention Weighting

The projection Jacobian $\mathbf{J}_{s,n}(\mathbf{p}) = \partial \pi_{s,n}(\mathbf{p})/\partial \mathbf{p} \in \mathbb{R}^{2\times 3}$ describes local sensitivity in normalized image coordinates. We evaluate it at each shifted point from Eq. (1) and use its smaller non-zero singular value to bias the attention logits:

$$\text{obs} = \sigma_{\min}\big(\mathbf{J}_{s,n}(\mathbf{p}_{\text{shift}})\big).\text{detach}(), \qquad \mathbf{a} = \text{softmax}\big(\mathbf{W}_q\mathbf{f} + s_{h,l} \cdot \log(\text{obs} + \epsilon)\big), \tag{2}$$

where $\mathbf{W}_q\mathbf{f}$ gives query logits and $s_{h,l}$ is learned per head and level. Softmax normalizes levels and points within each observation and head. The Jacobian is detached; feature sampling remains differentiable. Under the stated pinhole assumptions (Appendix B.2), this bias is a learned log-depth prior with numerical stabilization. The Jacobian formulation applies the same rule to the actual normalized projection, including camera intrinsics and image transforms, and extends to other differentiable projection models. We compute the Jacobian analytically alongside projection. All 12 learned scales in the SurroundOcc 8f model are positive (0.0267–0.5497).

3.2.3 Observation-Aware Gated Fusion

With multiple temporal steps, the same voxel may receive features from several camera-time observations with different projection geometry and temporal relevance. Rather than averaging these

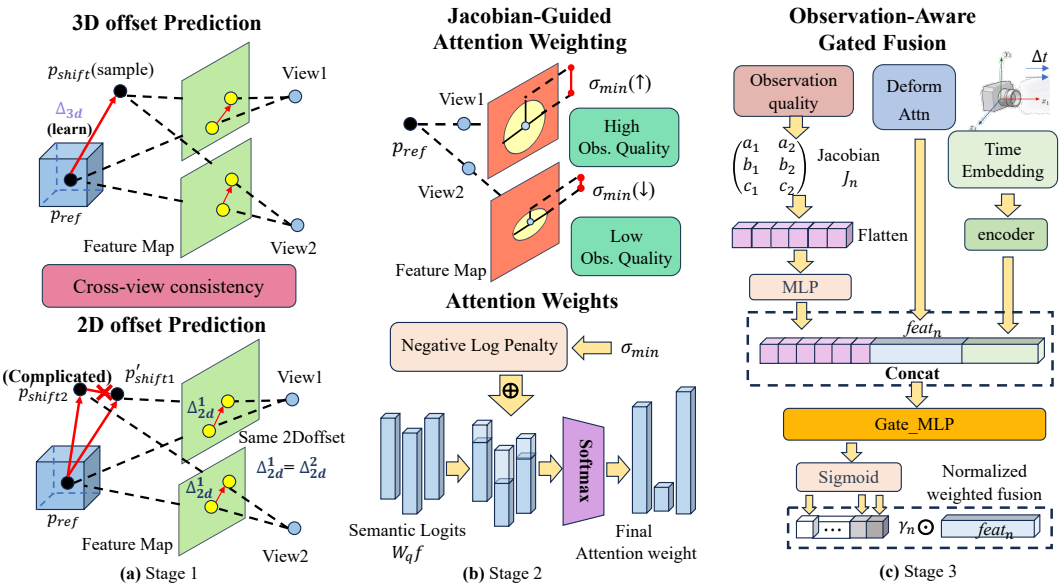

Figure 3: **PA-DA: three-stage projection-aware deformable attention.** Stage 1 projects points displaced by shared 3D offsets into each view. Stage 2 applies a Jacobian-based attention bias. Stage 3 predicts channel-wise fusion weights from per-observation features, the full projection Jacobian, and temporal metadata, then normalizes the weighted features across valid camera-time observations.

observations uniformly, we predict a channel-wise gate for each one. Let $\mathbf{v}_{s,n} \in \mathbb{R}^C$ denote the cross-attention feature from observation (s, n). The full 2×3 projection Jacobian at the query position, denoted $\mathbf{J}_{s,n}$, is detached and embedded alongside signed temporal metadata to condition the gate:

$$\mathbf{g}^{\text{geo}}_{s,n} = \text{MLP}_{\text{geo}}\big(1000\,\text{flatten}(\mathbf{J}_{s,n})\big), \tag{3}$$

$$\boldsymbol{\tau}_{s,n} = [\,\Delta t_{s,n},\ \log(1+|\Delta t_{s,n}|),\ r_{s,n},\ m_{s,n}\,], \qquad \mathbf{g}^{\text{time}}_{s,n} = \text{MLP}_{\text{time}}(\boldsymbol{\tau}_{s,n}), \tag{4}$$

$$\boldsymbol{\gamma}_{s,n} = \text{MLP}_{\text{gate}}([\,\mathbf{v}_{s,n} \,\|\, \mathbf{g}^{\text{geo}}_{s,n} \,\|\, \mathbf{g}^{\text{time}}_{s,n}\,]) \in (0,1)^C, \tag{5}$$

where $\Delta t_{s,n}$ is the signed capture-time offset from the occupancy reference, $r_{s,n}$ indicates the reference step, and $m_{s,n}$ marks a valid frame. The gate MLP ends with a sigmoid and scores each observation independently, combining its semantic features with projection geometry and temporal context.

For a voxel query $\mathbf{q}$, let $\mathcal{V}(\mathbf{q})$ contain the observations that are both geometrically visible and temporally valid. PA-DA normalizes the per-channel gates across this set:

$$\mathbf{o}(\mathbf{q}) = \frac{\sum_{(s,n)\in\mathcal{V}(\mathbf{q})} \boldsymbol{\gamma}_{s,n} \odot \mathbf{v}_{s,n}}{\sum_{(s,n)\in\mathcal{V}(\mathbf{q})} \boldsymbol{\gamma}_{s,n}}, \tag{6}$$

where $\odot$ and the division are element-wise. The weights sum to one per channel, giving a weighted average of observation features. In implementation, the denominator is floored at 10^{-5} for numerical stability, yielding zero when no valid observations exist (Appendix B.3).

3.3 Coarse-Guided Gated Fusion

The decoder starts from a single learnable voxel grid at S_0 and refines it sequentially through S_1 and S_2. Each scale produces an occupancy prediction, and learned patch splitting expands every voxel token into eight child tokens to double the resolution along all three axes.

PA-DA is applied at S_0 and S_1 to aggregate cross-view image features on the smaller grids. The finest scale S_2 uses local 3D convolutions for refinement, retaining complete dense coverage without fine-scale cross-view attention.

At each transition $S_{i-1} \to S_i$, the current-scale blocks transform the patch-split initialization into $\mathbf{h}_{\text{curr}}$. The previous scale's output $\mathbf{h}_{\text{prev}}$ also supplies a separate coarse branch that predicts channel-

wise gates at low resolution. The gates and projected coarse features are upsampled to balance coarse context and current-scale features:

$$\mathbf{g} = \sigma\big(\mathrm{MLP}(\mathbf{h}_{\mathrm{prev}})\big), \tag{7}$$

$$\hat{\mathbf{g}} = \mathrm{Up}(\mathbf{g}), \qquad \hat{\mathbf{h}}_{\mathrm{prev}} = \mathrm{Up}\big(\mathrm{Proj}(\mathbf{h}_{\mathrm{prev}})\big), \tag{8}$$

$$\mathbf{h}_{\mathrm{fused}} = \hat{\mathbf{g}} \odot \hat{\mathbf{h}}_{\mathrm{prev}} + (1 - \hat{\mathbf{g}}) \odot \mathbf{h}_{\mathrm{curr}}, \tag{9}$$

where σ is the sigmoid function, Proj aligns the channel dimensions when needed, and Up denotes trilinear upsampling. Both gate prediction and feature projection operate on the coarse grid. Values of $\hat{\mathbf{g}}$ close to one favor coarse context, while values close to zero retain the current-scale features. This fusion is applied at both $S_0 \to S_1$ and $S_1 \to S_2$ transitions.

After fusion, a depthwise 3D convolution (Liu et al., 2022b) followed by LayerNorm provides local spatial refinement:

$$\mathbf{h}_{\mathrm{out}} = \mathrm{LayerNorm}\big(\text{DW-Conv3d}(\mathbf{h}_{\mathrm{fused}})\big), \tag{10}$$

where the convolution uses kernel size 3 with groups equal to the channel dimension. The refined output $\mathbf{h}_{\mathrm{out}}$ feeds the occupancy classifier and, except at the final scale, initializes the next scale through patch splitting.

4 Experiments

4.1 Setup

Datasets. We evaluate on the 6,019-sample nuScenes validation split (Caesar et al., 2020) using SurroundOcc (Wei et al., 2023) and Occ3D (Tian et al., 2023) annotations, and on the SSCBench-KITTI-360 test split (Li et al., 2023a). SurroundOcc and Occ3D use $200\times200\times16$ grids at 0.5 m and 0.4 m, respectively; SSCBench uses a $256\times256\times32$ grid at 0.2 m. We report geometric IoU and semantic mIoU:

$$\mathrm{IoU} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}, \qquad \mathrm{mIoU} = \frac{1}{C}\sum_{c=1}^{C} \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FP}_c + \mathrm{FN}_c}, \tag{11}$$

where TP, FP, and FN denote true positives, false positives, and false negatives. For SurroundOcc and SSCBench, IoU treats all semantic classes as occupied. Semantic mIoU excludes empty/free, with $C = 16$, 17, and 18 for SurroundOcc, Occ3D, and SSCBench, respectively. Occ3D mIoU uses camera-visible voxels.

Architecture. We adapt **VGGT-Omega-1B** (Wang et al., 2026) with LoRA (r=8), keeping the patch encoder frozen. The nuScenes inputs comprise six 704×256 images per step ($W\times H$), with $T \in \{1, 8\}$ for SurroundOcc and $T = 8$ for Occ3D. SSCBench uses four `image_00` frames at 1408×384, sampled every fifth source frame. PA-DA uses four heads and four sampling points per head and feature level. The backbone ablation also uses DINOv3-L (Siméoni et al., 2025); adapter details are given in Appendix A.

Training. The occupancy objective combines cross-entropy, semantic and geometric scaling losses, and Lovász-Softmax. Occ3D uses camera-mask supervision. Additional settings are given in Appendix A.

4.2 Main Results

SurroundOcc-nuScenes. Table 1 compares single-step and multi-step methods. With one temporal step, VGGT-Occ achieves 33.29% IoU and 21.45% mIoU, exceeding the strongest single-step baselines by 0.63 mIoU. With eight steps, it reaches 37.24% IoU and 25.36% mIoU, ranking first in semantic mIoU and surpassing SuperOcc-L (8f) by 0.81 mIoU. It also leads on 10 of the 16 semantic classes.

Occ3D-nuScenes. Under the camera-visible protocol, VGGT-Occ achieves the highest mIoU in Table 2 at 44.47%, surpassing COTR (+Dr.Occ) by 0.37 mIoU.

Method	mIoU	IoU	barrier	bicycle	bus	car	const. veh.	motorcycle	pedestrian	traffic cone	trailer	truck	drive. surf.	other flat	sidewalk	terrain	manmade	vegetation
MonoScene (1f)	7.31	23.96	4.03	0.35	8.00	8.04	2.90	0.28	1.16	0.67	4.01	4.35	27.72	5.20	15.13	11.29	9.03	14.86
BEVFormer (1f)	16.75	30.50	14.22	6.58	23.46	28.28	8.66	10.77	6.64	4.05	11.20	17.78	37.28	18.00	22.88	22.17	13.80	22.21
TPVFormer (1f)	17.10	30.86	15.96	5.31	23.86	27.32	9.79	8.74	7.09	5.20	10.97	19.22	38.87	21.25	24.26	23.15	11.73	20.81
OccFormer (1f)	19.03	31.39	18.65	10.41	23.92	30.29	10.31	14.19	13.59	10.13	12.49	20.77	38.78	19.79	24.19	22.21	13.48	21.35
SurroundOcc (1f)	20.30	31.49	20.59	11.68	28.06	30.86	10.70	15.14	14.09	12.06	14.38	22.26	37.29	23.70	24.49	22.77	14.89	21.86
GaussianFormer-2 (1f)	20.82	31.74	21.39	13.44	28.49	30.82	10.92	15.84	13.55	10.53	14.04	22.92	40.61	24.36	26.08	24.27	13.83	21.98
GaussRender (1f)	20.82	32.61	20.32	13.22	28.32	31.05	10.92	15.65	12.84	8.91	13.29	22.76	41.22	24.48	26.38	25.20	15.31	23.25
QuadricFormer (1f)	20.12	31.22	19.58	13.11	27.27	29.64	11.25	16.26	12.65	9.15	12.51	21.24	40.20	24.34	25.69	24.24	12.95	21.86
GaussianWorld (stream)	22.13	33.40	21.38	14.12	27.71	31.84	13.66	17.43	13.66	11.46	15.09	23.94	42.98	24.86	28.84	26.74	15.69	24.74
S2GO-Base (4f)	22.72	35.46	21.93	13.36	27.47	32.08	14.86	15.31	12.91	11.79	13.42	23.98	46.85	29.14	30.30	29.05	14.69	26.40
SuperOcc-L (8f)	24.55	38.13	23.92	12.43	29.77	33.62	17.36	17.74	14.95	14.79	13.42	26.25	48.22	29.22	32.16	30.20	18.67	30.04
VGGT-Occ (1f)	21.45	33.29	21.39	16.73	27.04	29.35	10.24	18.09	15.62	14.27	13.41	23.06	39.03	24.77	25.41	25.13	15.33	24.28
VGGT-Occ (8f)	25.36	37.24	25.39	18.63	29.88	32.90	15.57	22.38	17.37	17.14	15.43	26.73	45.21	29.65	31.19	29.34	19.56	29.45

Table 1: **3D semantic occupancy on SurroundOcc-nuScenes.** The labels 1f, 4f, and 8f denote the number of temporal steps, each comprising six surround-view images; "stream" denotes recurrent temporal propagation across successive steps. Dark, medium, and light purple indicate the best, second-best, and third-best results, respectively.

Method	mIoU	others	barrier	bicycle	bus	car	const. veh.	motorcycle	pedestrian	traffic cone	trailer	truck	drive. surf.	other flat	sidewalk	terrain	manmade	vegetation
BEVFormer	26.90	5.90	37.80	17.90	40.40	42.40	7.40	23.90	21.80	21.00	22.40	30.70	55.40	28.40	36.00	28.10	20.00	17.70
CTF-Occ	28.50	8.10	39.30	20.60	38.30	42.20	16.90	24.50	22.70	21.10	23.00	31.10	53.30	33.80	38.00	33.20	20.80	18.00
TPVFormer	27.80	7.20	38.90	13.70	40.80	45.90	17.20	20.00	18.90	14.30	26.70	34.20	55.70	35.50	37.60	30.70	19.40	16.80
SparseOcc	30.90	10.60	39.20	20.20	32.90	43.30	19.40	23.80	23.40	29.30	21.40	29.30	67.70	36.30	44.60	40.90	22.00	21.90
BEVFormer+OSP*	41.21	10.95	49.00	27.68	50.24	55.99	22.96	31.02	30.91	30.25	35.60	41.23	82.09	42.59	51.90	55.10	44.82	38.17
FlashOcc*	37.80	9.10	46.30	17.70	42.70	50.60	23.70	20.10	22.30	24.10	30.30	37.40	81.70	40.10	52.30	56.50	47.70	40.60
FB-Occ*	39.10	13.60	44.70	27.00	45.40	49.10	25.20	26.30	27.90	27.80	32.30	36.80	80.10	42.80	51.20	55.10	42.20	37.50
ViewFormer*	41.90	12.90	50.10	27.90	44.60	52.90	22.40	29.60	28.00	29.20	35.20	39.40	84.70	49.40	57.40	59.70	47.40	40.60
BEVDet4D*	36.00	8.20	44.20	10.30	42.10	49.60	23.40	17.40	21.50	19.70	31.30	37.10	80.10	37.40	50.40	54.30	45.60	39.60
BEVDet4D (+Dr.Occ)*	43.43	14.90	52.80	30.70	45.60	55.10	30.30	30.80	31.00	35.90	39.10	42.50	81.10	43.40	54.00	57.60	49.00	44.50
COTR*‡	43.10	12.80	51.60	30.30	45.40	54.90	30.40	30.40	30.00	33.90	36.90	42.10	83.40	44.40	55.30	58.00	49.60	44.00
COTR (+Dr.Occ)*	44.10	14.00	53.00	32.60	45.80	54.90	29.70	31.50	30.90	35.70	38.00	42.50	83.80	44.90	56.10	58.80	51.10	45.60
VGGT-Occ*	44.47	14.43	51.26	31.99	50.03	54.62	30.10	34.16	32.42	33.98	37.15	41.76	84.38	49.63	56.69	59.68	49.60	44.13

Table 2: **3D semantic occupancy on Occ3D-nuScenes.** Comparison values follow Dr.Occ (Zhu et al., 2026), except BEVFormer+OSP (Shi et al., 2024), which uses the original reported scores. * denotes training with camera-mask supervision, and ‡ denotes results retrained in Dr.Occ because official pretrained weights were unavailable. Colors follow Table 1.

Method	mIoU	IoU	car	bicycle	motorcycle	truck	other veh.	person	road	parking	sidewalk	other ground	building	fence	vegetation	terrain	pole	traffic sign	other struct.	other object
MonoScene (1f)	12.30	37.90	19.30	0.40	0.60	8.00	2.00	0.90	48.40	11.40	28.10	3.20	32.90	3.50	26.20	16.80	6.90	5.70	4.20	3.10
VoxFormer (1f)	11.90	38.80	17.80	1.20	0.90	4.60	2.10	1.60	47.00	9.70	27.20	2.90	31.20	5.00	29.00	14.70	6.50	6.90	3.80	2.40
TPVFormer (1f)	13.60	40.20	21.60	1.10	1.40	8.10	2.60	2.40	53.00	12.00	31.10	3.80	34.80	4.80	30.10	17.50	7.50	5.90	5.50	2.70
OccFormer (1f)	13.80	40.30	22.60	0.70	0.30	9.90	3.80	2.80	54.30	13.40	31.50	3.60	36.40	4.80	31.00	19.50	7.80	8.50	7.00	4.60
GaussianFormer (1f)	12.90	35.40	18.90	1.00	4.60	18.10	7.60	3.40	45.50	10.90	25.00	5.30	28.40	5.70	29.50	8.60	3.00	2.30	9.50	5.10
GaussianFormer-2 (1f)	13.90	38.40	21.10	2.60	4.20	12.40	5.70	1.60	54.10	11.00	32.30	3.30	32.00	5.00	28.90	17.30	3.60	5.50	5.90	3.50
S2GO-Base (4f)	15.10	40.80	22.70	1.30	1.70	15.90	5.10	2.10	53.80	13.30	33.40	3.80	35.30	7.20	31.20	21.10	6.40	6.50	6.00	4.20
VGGT-Occ (4f)	18.13	45.06	27.38	4.31	6.70	17.15	9.04	3.15	57.98	12.65	36.64	4.19	42.04	8.43	35.42	21.46	10.91	11.21	9.39	8.24

Table 3: **3D semantic scene completion on SSCBench-KITTI-360.** Results use **monocular** camera input on the test set. Comparison values are reproduced as reported in S2GO (Park et al., 2026). The labels 1f and 4f denote temporal steps, each with one monocular image. Colors follow Table 1.

SSCBench-KITTI-360. With four frames, VGGT-Occ leads Table 3 with 45.06% IoU and 18.13% mIoU. It exceeds S2GO-Base (4f), the strongest semantic baseline, by 4.26 IoU and 3.03 mIoU, and GaussianFormer-2 (1f) by 4.23 mIoU. It ranks first on 13 of 18 semantic classes. Across all three benchmarks, VGGT-Occ outperforms the compared state-of-the-art methods in semantic mIoU.

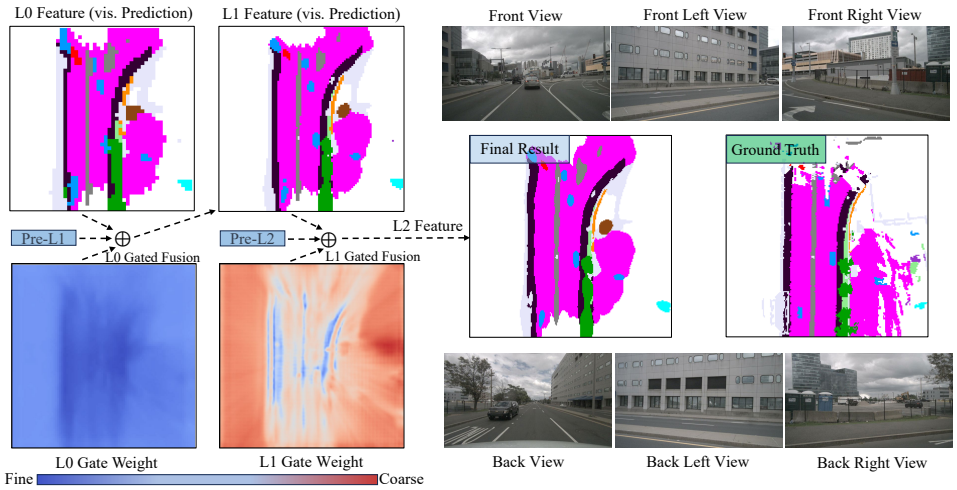

Figure 4: **Visualization of the coarse-to-fine gated fusion.** (Left) Coarse features (L_0, L_1), visualized through occupancy predictions, are fused with current-scale features (Pre-L_1, Pre-L_2). (Right) Camera images, final prediction (L_2), and ground truth. Warmer gate colors indicate larger coarse-feature weights.

Gating Behavior. Figure 4 shows the gates at $S_0 \rightarrow S_1$ and $S_1 \rightarrow S_2$, labeled L_0 and L_1, respectively. The predominantly blue L_0 map favors current-scale features, whereas the warmer L_1 map gives greater weight to coarse context, with cooler regions near some boundaries. This pattern is consistent with the scale allocation: S_1 retains PA-DA cross-view attention, while S_2 uses local refinement. The maps illustrate spatially varying balances between coarse context and current-scale detail.

4.3 Ablation Studies

All ablations use SurroundOcc-nuScenes validation with eight temporal steps. Unless noted, they share the VGGT-Omega configuration, decoder, input history, and training recipe (Appendix A). The PA-DA tables vary only the indicated stages or gate inputs.

4.3.1 Stage-wise Analysis of PA-DA

The three PA-DA stages introduce projection geometry into sampling, attention weighting, and observation fusion, respectively.

Table 4 compares **Standard** (2D offsets, uniform observation averaging), S1 (shared 3D offsets), S2 (Jacobian bias), S3 (observation gate), and full PA-DA under the matched setup. Full PA-DA improves over Standard by 2.10 IoU and 1.63 mIoU; S1 and S2 add 0.24 and 0.36 mIoU, respectively.

S3 adds 1.18 mIoU to S1 and 1.03 mIoU to S1+S2. Thus the gain is not solely from 3D sampling or the Jacobian bias.

Configuration	S1	S2	S3	IoU	mIoU
Standard				35.14	23.73
S1	✓			35.82	23.97
S1 + S2	✓	✓		36.37	24.33
S1 + S3	✓		✓	36.89	25.15
Full PA-DA	✓	✓	✓	**37.24**	**25.36**

Table 4: **PA-DA stage ablation (8f).**

Stage 3 conditioning. Table 5 keeps S1+S2 fixed and varies only gate inputs. **Feature** is $\mathbf{v}_{s,n}$; the feature-only row is a generic gate. Full conditioning improves over it by 0.91 mIoU. Omitting feature, time, or Jacobian gives 24.97, 24.68, or 25.02 mIoU, respectively, versus 25.36 with all three inputs.

Feature	Jacobian	Time	IoU	mIoU
✓			36.51	24.45
	✓	✓	36.64	24.97
✓	✓		36.71	24.68
✓		✓	36.79	25.02
✓	✓	✓	**37.24**	**25.36**

Table 5: **Stage 3 gate conditioning (8f).**

PA-DA across backbones. With the same coarse-to-fine decoder, PA-DA improves IoU/mIoU by 1.70/1.24 points on DINOv3-L and 2.10/1.63 on VGGT-Omega (Table 6), supporting its use across both encoders.

Backbone	Attention	IoU	mIoU
VGGT-Omega	Standard	35.14	23.73
VGGT-Omega	PA-DA	**37.24**	**25.36**
DINOv3-L	Standard	32.86	21.07
DINOv3-L	PA-DA	**34.56**	**22.31**

Table 6: **PA-DA across backbones (8f).**

4.3.2 Coarse-Guided Refinement Across Scales

All variants retain patch splitting; we vary coarse guidance, fusion weights, and depthwise convolution (DW).

No guidance disables both coarse branches; single-transition variants enable only the named branch. **Fixed fusion** uses equal coarse/current weights at both transitions. Only **Both gates** omits DW.

The **Full decoder** gains 1.26 IoU and 1.32 mIoU over **No guidance**. Individual transitions add 0.45/0.83 mIoU; both together perform best. Fixed fusion adds 0.79 mIoU, learned gating adds another 0.53, and DW adds 0.09. Thus adaptive coarse guidance improves on patch splitting, especially at the finer transition.

Configuration	$S_0 \to S_1$	$S_1 \to S_2$	DW	IoU	mIoU
No guidance			✓	35.98	24.04
S_0-guided	✓		✓	36.47	24.49
S_1-guided		✓	✓	36.59	24.87
Fixed fusion	0.5	0.5	✓	36.41	24.83
Both gates	✓	✓		37.03	25.27
Full decoder	✓	✓	✓	**37.24**	**25.36**

Table 7: **Coarse-guided refinement ablation.** All variants retain patch splitting.

4.4 Inference Cost

To put inference cost in context, we compare VGGT-Occ with two commonly used occupancy methods, BEVDetOcc and GaussianFormer-2, retaining each method's original input configuration (Table 8). VGGT-Occ has 955.7M parameters without the VGGT-Omega output heads; 47.8M are trainable, including the 41.1M decoder. On RTX PRO 6000 and RTX 5090, backbone latency is 433.05 and 663.64 ms, respectively. PA-DA decoder latency is 148.03 and 161.11 ms, versus 144.11 and 156.64 ms for standard deformable attention. PA-DA adds 2.7% and 2.9% to decoder latency; the backbone accounts for 74.5% and 80.5% of total latency.

Method	Input ($T \times C \times H \times W$)	RTX PRO 6000		RTX 5090	
		Latency (ms)	Memory (GB)	Latency (ms)	Memory (GB)
BEVDetOcc	$8 \times 6 \times 384 \times 704$	666.82	4.60	610.63	5.09
GaussianFormer-2	$1 \times 6 \times 864 \times 1600$	729.48	23.99	768.05	20.03
VGGT-Occ	$8 \times 6 \times 256 \times 704$	581.08	13.76	824.75	14.68

Table 8: **Inference cost.** C denotes cameras per step. Batch 1; CUDA 12.8, TensorRT 10.15, mixed FP16. Mean of 20 full forward passes after 5 warm-ups. Memory reports measured CUDA device usage in GB.

5 Conclusion

We introduced VGGT-Occ for vision-only 3D semantic occupancy prediction. PA-DA incorporates projection geometry throughout sampling, weighting, and observation fusion, while coarse-guided decoding balances 3D context with local detail. The model achieves the highest semantic mIoU among the compared methods on three benchmarks. Ablations show gains from adaptive coarse guidance and from PA-DA with both evaluated backbones. The large encoder remains the main computational bottleneck, making its acceleration a priority for real-time deployment.

AI use statement

Generative AI assisted writing, formatting, and interpretation of reported results, but not dataset generation. The authors reviewed the AI-assisted content for accuracy and consistency and take responsibility for the final manuscript.

REFERENCES

Maxim Berman, Amal Rannen Triki, and Matthew B. Blaschko. The Lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In *CVPR*, 2018.

Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020.

Anh-Quan Cao and Raoul de Charette. MonoScene: Monocular 3D semantic scene completion. In *CVPR*, 2022.

Loïck Chambon, Eloi Zablocki, Alexandre Boulch, Mickaël Chen, and Matthieu Cord. GaussRender: Learning 3D occupancy with Gaussian rendering. In *ICCV*, 2025.

Tianchen Deng, Xun Chen, Ziming Li, Hongming Shen, Danwei Wang, Javier Civera, and Hesheng Wang. UniPR-3D: Towards universal visual place recognition with visual geometry grounded transformer. *arXiv preprint arXiv:2512.21078*, 2025a.

Tianchen Deng, Wenhua Wu, Kunzhen Wu, Guangming Wang, Siting Zhu, Shenghai Yuan, Xun Chen, Guole Shen, Zhe Liu, and Hesheng Wang. Reloc-VGGT: Visual re-localization with geometry grounded transformer. *arXiv preprint arXiv:2512.21883*, 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, Yun Ye, 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 *CVPR*, 2023.

Yuanhui Huang, Wenzhao Zheng, Borui Zhang, Jie Zhou, and Jiwen Lu. SelfOcc: Self-supervised vision-based 3D occupancy prediction. In *CVPR*, 2024a.

Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. GaussianFormer: Scene as Gaussians for vision-based 3D semantic occupancy prediction. In *ECCV*, 2024b.

Yuanhui Huang, Amonnut Thammatadatrakoon, Wenzhao Zheng, Yunpeng Zhang, Dalong Du, and Jiwen Lu. GaussianFormer-2: Probabilistic Gaussian superposition for efficient 3D occupancy prediction. In *CVPR*, 2025.

Xiaohui Jiang, Shuailin Li, Yingfei Liu, Shihao Wang, Fan Jia, Tiancai Wang, Lijin Han, and Xiangyu Zhang. Far3D: Expanding the horizon for surround-view 3D object detection. In *AAAI*, 2024.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D Gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4), 2023.

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 *ECCV*, 2024.

Yiming Li, Sihang Li, Xinhao Liu, Moonjun Gong, Kenan Li, Nuo Chen, Zijun Wang, Zhiheng Li, Tao Jiang, Fisher Yu, Yue Wang, Hang Zhao, Zhiding Yu, and Chen Feng. SSCBench: A large-scale 3D semantic scene completion benchmark for autonomous driving. *arXiv preprint arXiv:2306.09001*, 2023a.

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 *CVPR*, 2023b.

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 multi-camera images via spatiotemporal transformers. In *ECCV*, 2022.

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. In *CVPR Workshop on End-to-End Autonomous Driving*, 2023c.

Xuewu Lin, Tianwei Lin, Zixiang Pei, Lichao Huang, and Zhizhong Su. Sparse4D: Multi-view 3D object detection with sparse spatial-temporal fusion. *arXiv preprint arXiv:2211.10581*, 2022.

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

Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. PETR: Position embedding transformation for multi-view 3D object detection. In *ECCV*, 2022a.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. In *CVPR*, 2022b.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 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 *CVPR*, 2024b.

Gyeongrok Oh, Sungjune Kim, Heeju Ko, Hyung-gun Chi, Jinkyu Kim, Dongwook Lee, Daehyun Ji, Sungjoon Choi, Sujin Jang, and Sangpil Kim. 3D occupancy prediction with low-resolution queries via prototype-aware view transformation. In *CVPR*, 2025.

Mingjie Pan, Jiaming Liu, Renrui Zhang, Peixiang Huang, Xiaoqi Li, Hongwei Xie, Bing Wang, Li Liu, and Shanghang Zhang. RenderOcc: Vision-centric 3D occupancy prediction with 2D rendering supervision. In *ICRA*, 2024.

Jinhyung Park, Chensheng Peng, Yihan Hu, Wenzhao Zheng, Kris Kitani, and Wei Zhan. S2GO: Streaming sparse Gaussian occupancy. In *ICLR*, 2026.

Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3D. In *ECCV*, 2020.

Thomas Roddick, Alex Kendall, and Roberto Cipolla. Orthographic feature transform for monocular 3D object detection. In *BMVC*, 2019.

Yiang Shi, Tianheng Cheng, Qian Zhang, Wenyu Liu, and Xinggang Wang. Occupancy as set of points. In *ECCV*, 2024.

Yunxiao Shi, Hong Cai, Jisoo Jeong, Yinhao Zhu, Shizhong Han, Amin Ansari, and Fatih Porikli. BePo: Dual representation for 3D occupancy prediction. In *CVPR Workshop on Autonomous Driving*, 2026.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

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 *NeurIPS*, 2023.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *CVPR*, 2025.

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. VGGT-Ω. *arXiv preprint arXiv:2605.15195*, 2026.

Yue Wang, Vitor Campagnolo Guizilini, Tianyuan Zhang, Yilun Wang, Hang Zhao, and Justin Solomon. DETR3D: 3D object detection from multi-view images via 3D-to-2D queries. In *CoRL*, 2021.

Yuqi Wang, Yuntao Chen, Xingyu Liao, Lue Fan, and Zhaoxiang Zhang. PanoOcc: Unified occupancy representation for camera-based 3D panoptic segmentation. In *CVPR*, 2024.

Yi Wei, Linqing Zhao, Wenzhao Zheng, Zheng Zhu, Jie Zhou, and Jiwen Lu. SurroundOcc: Multi-camera 3D occupancy prediction for autonomous driving. In *ICCV*, 2023.

Huaiyuan Xu, Junliang Chen, Shiyu Meng, Yi Wang, and Lap-Pui Chau. A survey on occupancy perception for autonomous driving: The information fusion perspective. *Information Fusion*, 114: 102671, 2025.

Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, 2024.

Zichen Yu, Changyong Shu, Jiajun Deng, Kangjie Lu, Zongdai Liu, Jiangyong Yu, Dawei Yang, Hui Li, and Yan Chen. FlashOcc: Fast and memory-efficient occupancy prediction via channel-to-height plugin. *arXiv preprint arXiv:2311.12058*, 2023.

Zichen Yu, Quanli Liu, Wei Wang, Liyong Zhang, and Xiaoguang Zhao. SuperOcc: Toward cohesive temporal modeling for superquadric-based 3D occupancy prediction. *arXiv preprint arXiv:2601.15644*, 2026.

Haiming Zhang, Yiyao Zhu, Wending Zhou, Xu Yan, Yingjie Cai, Bingbing Liu, Shuguang Cui, and Zhen Li. SQS: Enhancing sparse perception models via query-based splatting in autonomous driving. In *NeurIPS*, 2025.

Yanan Zhang, Jinqing Zhang, Zengran Wang, Junhao Xu, and Di Huang. Vision-based 3D occupancy prediction in autonomous driving: a review and outlook. *Frontiers of Computer Science*, 20: 2001301, 2026.

Yunpeng Zhang, Zheng Zhu, and Dalong Du. OccFormer: Dual-path transformer for vision-based 3D semantic occupancy prediction. In *ICCV*, 2023.

Lingjun Zhao, Sizhe Wei, James Hays, and Lu Gan. GaussianFormer3D: Multi-modal Gaussian-based semantic occupancy prediction with 3D deformable attention. In *ICRA*, 2026.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable transformers for end-to-end object detection. In *ICLR*, 2021.

Xubo Zhu, Haoyang Zhang, Fei He, Rui Wu, Yanhu Shan, Wen Yang, and Huai Yu. Dr.Occ: Depth- and region-guided 3D occupancy from surround-view cameras for autonomous driving. In *CVPR*, 2026.

Sicheng Zuo, Wenzhao Zheng, Xiaoyong Han, Longchao Yang, Yong Pan, and Jiwen Lu. QuadricFormer: Scene as superquadrics for 3D semantic occupancy prediction. In *NeurIPS*, 2025a.

Sicheng Zuo, Wenzhao Zheng, Yuanhui Huang, Jie Zhou, and Jiwen Lu. GaussianWorld: Gaussian world model for streaming 3D occupancy prediction. In *CVPR*, 2025b.

A Implementation and Training Details

A.1 Architecture Hyperparameters

Table 9 lists network dimensions and nuScenes grids; SSCBench input and grid sizes are in Table 10.

Table 9: **Architecture hyperparameters.**

Component	Parameter	Value
VGGT-Omega backbone	Pretrained parameters	1B
	Adaptation	LoRA (r=8); patch encoder frozen
	Attention depth	24 frame/global block pairs
	Input resolution	256×704 (height × width)
	Patch size	16
	Embed dim	1024
OccHead (Coarse)	Grid size	50×50×4 (10K)
	Feature dim	256
	Blocks	cross, conv, cross, conv
OccHead (Medium)	Grid size	100×100×8 (80K)
	Feature dim	256
	Blocks	cross, conv, conv
OccHead (Fine)	Grid size	200×200×16 (640K)
	Feature dim	64
	Blocks	conv, conv
PA-DA	n_heads	4
	n_points	4
	Offset predictor	Linear, 3D output per head/level/point
	Jacobian bias scale $s_{h,l}$	Learnable, initialized to 0.1
Observation-Aware Gated Fusion	Geometry MLP	$6 \to 128 \to 256$ (LN, ReLU)
	Temporal MLP	$4 \to 128 \to 256$ (LN, ReLU)
	Gate MLP	$768 \to 128 \to 256$ (LN, ReLU, Sigmoid)
Coarse-Guided Gated Fusion	$S_0 \to S_1$ gate	$256 \to 64 \to 256$ (ReLU, Sigmoid)
	$S_1 \to S_2$ gate	$256 \to 64 \to 64$ (ReLU, Sigmoid)
	Post-fusion refinement	Depthwise Conv3d (k=3) + LayerNorm

DINOv3-L adapter. The backbone comparison uses standalone DINOv3 ViT-L/16 with rank-8 LoRA. Features from blocks 5, 12, 18, and 24 feed a 1024-channel DPT input adapter, without Omega frame/global attention blocks or camera-token conditioning. Both attention variants retain the same coarse-to-fine decoder.

VGGT-Omega pretraining data. The public pretraining datasets listed in Section 3.5.1 of VGGT-Omega (Wang et al., 2026) include Waymo and the synthetic Virtual KITTI, but do not include nuScenes, KITTI, or KITTI-360. Virtual KITTI is distinct from the real KITTI and KITTI-360 benchmarks.

A.2 Dataset, Evaluation, and Training Protocols

Table 10 summarizes the benchmark protocols. Training and evaluation follow the label and visibility conventions used in the comparisons. Occ3D uses camera-visible supervision and evaluation, matching the camera-mask-supervised baselines (*) in Table 2. SurroundOcc and SSCBench use full-volume evaluation over valid voxels, excluding ignored labels.

Optimization. For each benchmark, we train for 30 epochs with AdamW, linear learning-rate warmup, and cosine decay from 10^{-5} to 5×10^{-6}. Training uses EMA and four RTX PRO 6000 GPUs, with batch size 4 per GPU and four gradient-accumulation steps (effective batch size 64). Full configurations will be released with the code.

Checkpoint selection and ablation controls. All reported main and ablation results use the EMA checkpoint with the best validation mIoU during the 30-epoch training run. Within each backbone, compared ablation variants use the same initialization for shared parameters and the same training recipe. The cross-backbone comparison uses each backbone's own pretrained initialization.

Setting	SurroundOcc-nuScenes	Occ3D-nuScenes	SSCBench-KITTI-360
Cameras	6 surround	6 surround	image_00 (mono)
Input ($W \times H$)	704×256	704×256	1408×384
Temporal	$1f/8f$	$8f$	$4f$, gap=5
Voxel grid	$200 \times 200 \times 16$, 0.5 m	$200 \times 200 \times 16$, 0.4 m	$256 \times 256 \times 32$, 0.2 m
Evaluation	16-class IoU/mIoU	17-class mIoU, camera-visible mask	18-class mIoU/completion IoU
Backbone adaptation	LoRA (r=8); patch encoder frozen	LoRA (r=8); patch encoder frozen	LoRA (r=8); patch encoder frozen
Training objective	Occupancy losses	Occupancy losses	Occupancy losses

Table 10: **Dataset, evaluation, and training protocols.** Input sizes are width × height. For nuScenes, 8f uses the reference keyframe and seven preceding keyframes from the same scene, consistent with the compared history-based methods; no future keyframes are used. Unavailable history is padded and marked invalid. SSCBench uses every fifth source frame.

A.3 Loss Formulation

For logits $\mathbf{z}_{r,i}$, probabilities $\mathbf{p}_{r,i} = \mathrm{softmax}(\mathbf{z}_{r,i})$, and labels $y_{r,i}$, scales $r \in \{0, 1, 2\}$ have coarse-to-fine weights $(w_0, w_1, w_2) = (0.1, 0.2, 1.0)$. All sums use batch-wide valid voxels $\mathcal{I}_r$, excluding label 255 and masked voxels. The empty index e is 17 for Occ3D and 0 for SurroundOcc/SSCBench.

Cross-Entropy Loss. With label smoothing $\eta = 0.1$ and class weights $\boldsymbol{\omega}$,

$$\mathcal{L}_{\text{ce}} = \sum_r w_r \ \mathrm{CE}_{\eta,\boldsymbol{\omega}}(\mathbf{z}_r, y_r; \mathcal{I}_r). \tag{12}$$

Here CE uses mean reduction. Given class frequencies f_c^{ce}, $\widetilde{\omega}_c = 1/\log(f_c^{\text{ce}} + \delta)$ and $\omega_c = \widetilde{\omega}_c / \operatorname{mean}_d \widetilde{\omega}_d$, with default $\delta = 10^{-3}$; otherwise $\omega_c = 1$.

Geometric Scaling Loss (geo_scal). Soft precision, recall, and specificity for probabilities a_i and binary targets t_i are

$$P = \frac{\sum_i a_i t_i}{\sum_i a_i}, \qquad R = \frac{\sum_i a_i t_i}{\sum_i t_i}, \qquad S = \frac{\sum_i (1 - a_i)(1 - t_i)}{\sum_i (1 - t_i)}.$$

Let $\Phi(\mathbf{a}, \mathbf{t}) = -\log P - \log R - \log S$, with sums over $\mathcal{I}_r$ and numerical safeguards. With $a_{r,i} = 1 - p_{r,i,e}$ and $t_{r,i} = \mathbf{1}[y_{r,i} \neq e]$,

$$\mathcal{L}_{\text{geo}} = \sum_r w_r \, \Phi(\mathbf{a}_r, \mathbf{t}_r). \tag{13}$$

Semantic Scaling Loss (sem_scal). Averaging the same penalties over selected present classes $\mathcal{C}_r^{\text{sem}}$ gives

$$\mathcal{L}_{\text{sem}} = \sum_r \frac{w_r}{|\mathcal{C}_r^{\text{sem}}|} \sum_{c \in \mathcal{C}_r^{\text{sem}}} \Phi(\mathbf{p}_{r,c}, \mathbf{1}[y_r = c]). \tag{14}$$

The set includes all present classes, except e when empty-class exclusion is enabled. Undefined precision/specificity terms are omitted; an empty class set contributes zero.

Lovász-Softmax Loss. For errors $m_{r,i,c} = |p_{r,i,c} - \mathbf{1}[y_{r,i} = c]|$, the Lovász extension of the Jaccard loss (Berman et al., 2018) gives

$$\mathcal{L}_{\text{lovasz}} = \sum_r w_r \, \frac{\sum_{c \in \mathcal{C}_r^{\text{occ}}} \beta_{r,c} \, \overline{\Delta_{J_c}}(\mathbf{m}_{r,c})}{\sum_{c \in \mathcal{C}_r^{\text{occ}}} \beta_{r,c}}. \tag{15}$$

Here $\mathcal{C}_r^{\text{occ}}$ contains present non-empty classes; an empty set contributes zero. Weights are uniform unless frequencies f_c^{L} are supplied:

$$u_c = \max(f_c^{\text{L}}, 1)^{-\kappa}, \qquad \beta_c = \mathrm{clip}\left(\frac{u_c}{\operatorname{mean}_{d \neq e} u_d}, \, b_{\min}, \, b_{\max} \right).$$

Defaults are $(\kappa, b_{\min}, b_{\max}) = (0.25, 0.65, 1.75)$. Class balancing applies only at the finest scale by default ($\beta_{2,c} = \beta_c$, $\beta_{0,c} = \beta_{1,c} = 1$), or at all scales when this restriction is disabled.

Total Occupancy Loss.

$$\mathcal{L}_{\text{occ}} = \mathcal{L}_{\text{ce}} + \lambda_{\text{sem}} \mathcal{L}_{\text{sem}} + \lambda_{\text{geo}} \mathcal{L}_{\text{geo}} + \lambda_{\text{lovasz}} \mathcal{L}_{\text{lovasz}}. \tag{16}$$

The configuration sets λ_{sem} and λ_{geo} via `alpha` and `gamma`. At epoch E, $\lambda_{\text{lovasz}} = \lambda_{\text{target}} \mathbf{1}[E \geq E_{\text{start}}]$, with configured target/start defaults of 1/0.

B PA-DA: MATHEMATICAL DETAILS

We use the notation of Section 3.2.

B.1 STAGE 1: 3D OFFSET LEARNING WITH PROJECTION

For query $\mathbf{q} \in \mathbb{R}^d$ and reference point $\mathbf{p}_{\text{ref}} \in \mathbb{R}^3$, a linear predictor generates one 3D offset per head/level/point (h, l, k), shared across observations:

$$\boldsymbol{\Delta}_{3d}^{(h,l,k)} = \mathbf{W}_{\text{offset}}^{(h,l,k)}\mathbf{q} + \mathbf{b}_{\text{offset}}^{(h,l,k)} \in \mathbb{R}^3, \tag{17}$$

$$\mathbf{p}_{\text{shift}}^{(h,l,k)} = \mathbf{p}_{\text{ref}} + \boldsymbol{\Delta}_{3d}^{(h,l,k)}, \tag{18}$$

$$(u_{s,n}^{(h,l,k)}, v_{s,n}^{(h,l,k)}) = \pi_{s,n}(\mathbf{p}_{\text{shift}}^{(h,l,k)}). \tag{19}$$

With $\mathbf{R}_{s,n}, \mathbf{t}_{s,n}$ from $\mathbf{E}_{s,n}$, Eq. (1) uses

$$\pi_{s,n}(\mathbf{p}) = \mathbf{D}_{s,n}\,\mathcal{A}_{s,n}(\rho(\mathbf{K}_{s,n}(\mathbf{R}_{s,n}\mathbf{p} + \mathbf{t}_{s,n}))),$$

Here $\rho([x, y, z]^\top) = [x/z, y/z]^\top$ performs perspective division, $\mathcal{A}_{s,n}$ is the image affine transform, and $\mathbf{D}_{s,n} = \text{diag}(1/W_{s,n}, 1/H_{s,n})$ uses the transformed image dimensions.

These normalized locations are the final samples, without per-view 2D offsets; image-plane displacements vary with viewing geometry and depth.

B.2 STAGE 2: JACOBIAN-GUIDED ATTENTION WEIGHTING

For the full normalized projection, the Jacobian with respect to the occupancy-frame point and its SVD are

$$\mathbf{J}_{s,n}(\mathbf{p}) = \frac{\partial \pi_{s,n}(\mathbf{p})}{\partial \mathbf{p}} = \frac{\partial (u_{s,n}, v_{s,n})}{\partial \mathbf{p}} \in \mathbb{R}^{2\times 3}, \tag{20}$$

$$\mathbf{J}_{s,n} = \mathbf{U}\begin{bmatrix}\sigma_{\max} & 0 & 0\\ 0 & \sigma_{\min} & 0\end{bmatrix}\mathbf{V}^\top. \tag{21}$$

The smaller non-zero singular value $\sigma_{\min}$ is the minimum local projection sensitivity outside the viewing-ray null space, in inverse meters. For a zero-skew pinhole camera, identity image transform, and $f_x/W = f_y/H = \bar{f}$, it equals $\bar{f}/Z$ at camera-frame depth $Z > 0$; otherwise the full Jacobian is used.

Let $\sigma_{\min}^{(s,n,h,l,k)}$ be evaluated at $\mathbf{p}_{\text{shift}}^{(h,l,k)}$. Bias and attention are

$$b_{s,n}^{(h,l,k)} = s_{h,l} \cdot \log\Big(\text{sg}(\sigma_{\min}^{(s,n,h,l,k)}) + \epsilon\Big), \tag{22}$$

$$A_{s,n}^{(h,l,k)} = \text{softmax}_{l,k}\Big(\mathbf{w}_{\text{attn}}^{(h,l,k)\top}\mathbf{q} + b_{s,n}^{(h,l,k)}\Big). \tag{23}$$

Here $s_{h,l}$ is learned from an initial value of 0.1, $\epsilon = 10^{-5}$, and sg stops gradients through geometry, not feature sampling. Query-projected logits are normalized over levels/points within each observation/head; invalid samples receive a large negative bias. Positive $s_{h,l}$ suppresses more distant samples in the pinhole case. All 12 learned scales in the SurroundOcc 8f model are positive (0.0267–0.5497).

B.3 Stage 3: Observation-Aware Gated Fusion

For observation (s, n), the full query-position Jacobian is detached, scaled, and embedded by a two-layer MLP with internal LayerNorm. Temporal metadata comprises the signed offset $\Delta t_{s,n} = t_{s,n} - t_{\text{ref}}$ in seconds, its log magnitude, reference-step indicator $r_{s,n}$, and frame-validity indicator $m_{s,n}$. A sigmoid gate scores each observation independently. Over visible, temporally valid observations $\mathcal{V}(\mathbf{q})$,

$$\mathbf{g}^{\text{geo}}_{s,n} = \text{MLP}_{\text{geo}}\big(1000\,\text{flatten}(\text{sg}(\mathbf{J}_{s,n}))\big) \in \mathbb{R}^C, \tag{24}$$

$$\boldsymbol{\tau}_{s,n} = [\,\Delta t_{s,n},\ \log(1+|\Delta t_{s,n}|),\ r_{s,n},\ m_{s,n}\,], \qquad \mathbf{g}^{\text{time}}_{s,n} = \text{MLP}_{\text{time}}(\boldsymbol{\tau}_{s,n}), \tag{25}$$

$$\boldsymbol{\gamma}_{s,n} = \text{MLP}_{\text{gate}}([\,\mathbf{v}_{s,n} \,\|\, \mathbf{g}^{\text{geo}}_{s,n} \,\|\, \mathbf{g}^{\text{time}}_{s,n}\,]) \in (0,1)^C, \tag{26}$$

$$\mathbf{o}(\mathbf{q}) = \frac{\sum_{(s,n)\in\mathcal{V}(\mathbf{q})} \boldsymbol{\gamma}_{s,n} \odot \mathbf{v}_{s,n}}{\max\Big(\sum_{(s,n)\in\mathcal{V}(\mathbf{q})} \boldsymbol{\gamma}_{s,n},\ \epsilon_{\text{f}}\Big)}. \tag{27}$$

The element-wise denominator floor $\epsilon_{\text{f}} = 10^{-5}$ stabilizes Eq. (6) and gives zero output when no valid observations exist.

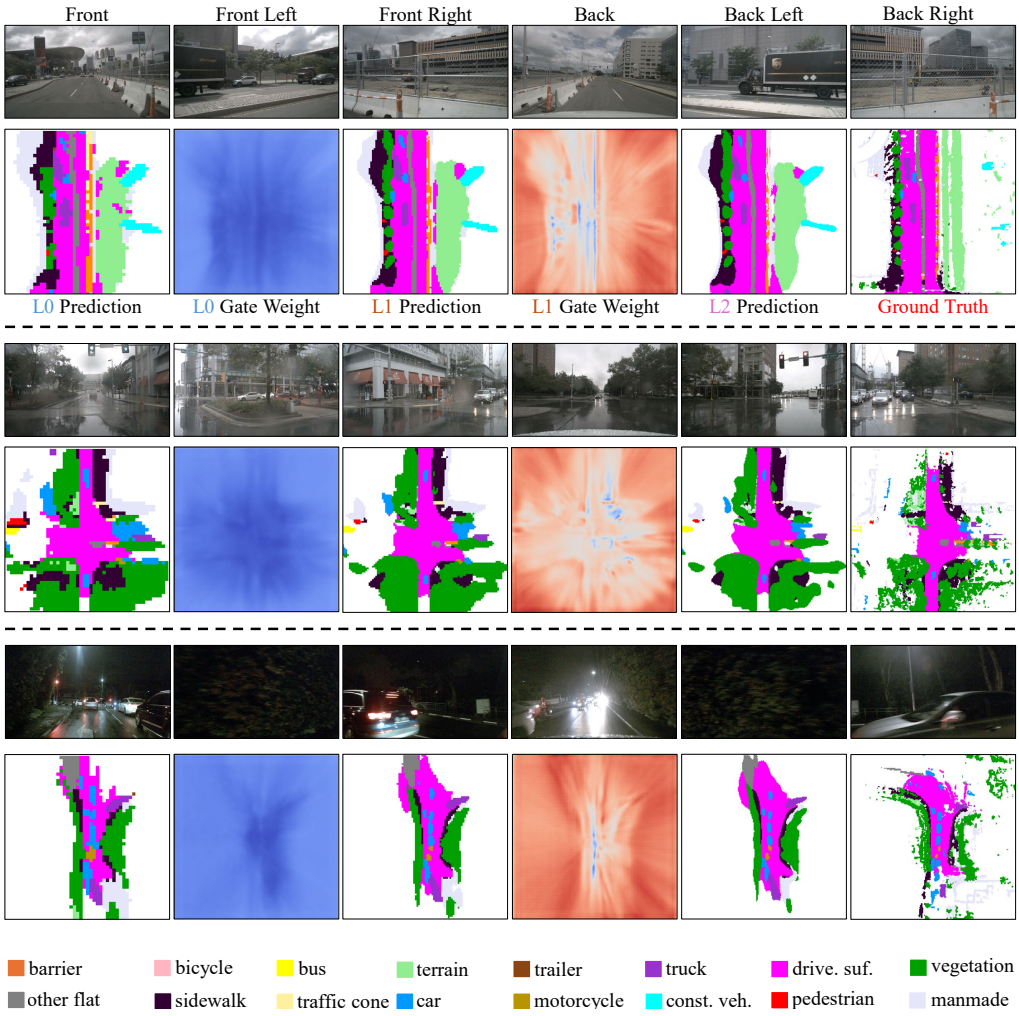

Figure 5: **Gate heatmaps under challenging conditions.** Daytime clutter (top), heavy rain (middle), and nighttime (bottom). Each scene includes three-scale predictions and ground truth. The L_0 and L_1 maps correspond to $S_0 \to S_1$ and $S_1 \to S_2$. Warmer colors indicate larger coarse-feature weights, while cooler colors indicate greater current-scale contributions.

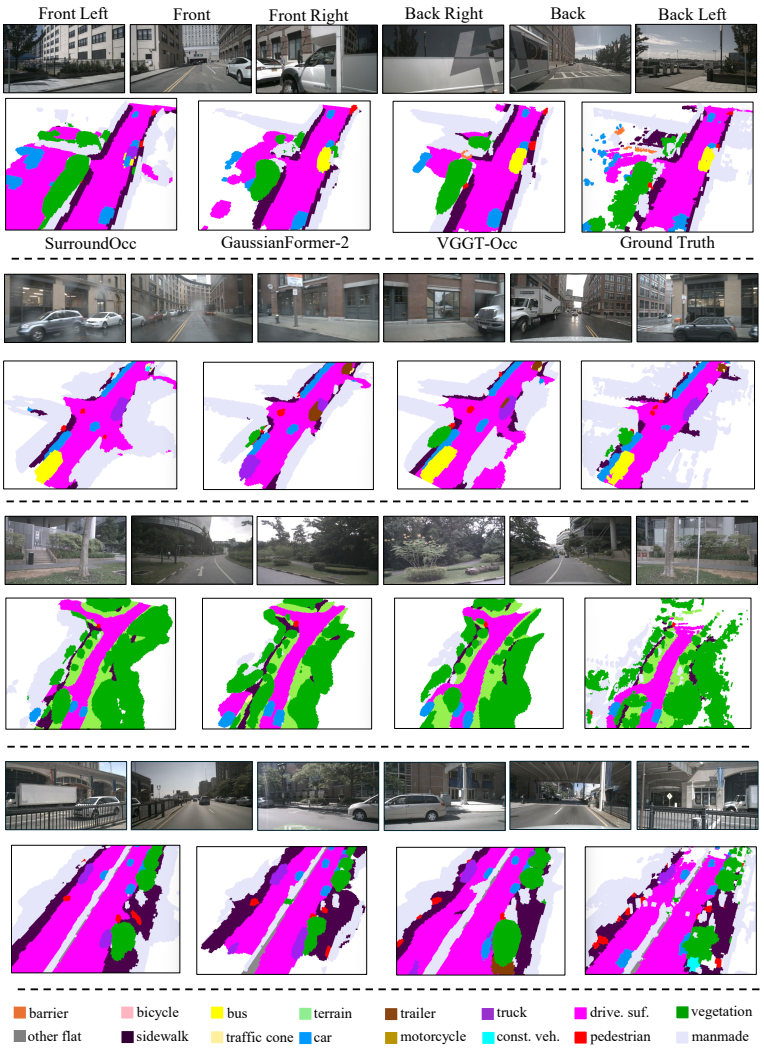

Figure 6: **Qualitative comparison on SurroundOcc-nuScenes.** From left to right: SurroundOcc, GaussianFormer-2, VGGT-Occ, and ground truth. VGGT-Occ shows clearer road-sidewalk boundaries and more coherent roadside semantic regions in the illustrated scenes.

C Gate Heatmap Analysis

Figure 5 extends the gating visualization to daytime construction clutter, heavy rain, and nighttime. In these examples, L_0 remains predominantly blue, favoring current-scale features, while L_1 assigns larger weights to coarse context over broad regions. Cooler bands within the otherwise warm L_1 maps indicate greater current-scale contributions in localized regions. This combination of broad coarse weighting and localized fine-scale weighting is visible in both the cluttered daytime scene and the rain/night examples. The maps illustrate spatially varying fusion of coarse context and local detail across the shown scenes.

D Additional Qualitative Results

Figure 6 compares VGGT-Occ with SurroundOcc and GaussianFormer-2 on urban streets and roadside vegetation. In the upper scenes, VGGT-Occ shows more continuous sidewalks and boundaries closer to ground truth; in the lower scenes, road and sidewalk remain distinguishable beside vegetation, including narrow road edges. These examples complement the semantic mIoU results in Table 1.