LEARNING MEMORY-GUIDED SCENE GRAPH REASONING FOR TEMPORAL 3D SEMANTIC SCENE COMPLETION

Anonymous authors
Paper under double-blind review

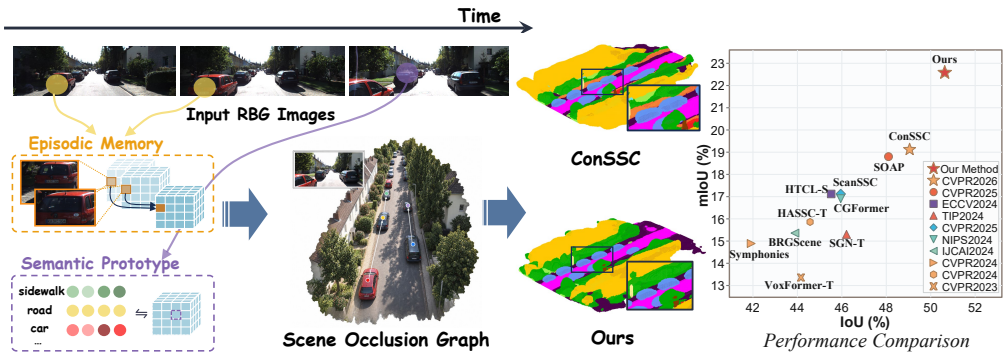

Figure 1: MemGraphSSC adapts scene completion to voxel observation states. For regions lacking current-frame support, it retrieves historical evidence from episodic memory when historical support is available and queries learned semantic prototypes otherwise, while preserving current visual evidence where available. Directed occlusion relations further guide feature refinement in occluded regions. Qualitative and quantitative comparisons illustrate the benefits of combining observation-conditioned memory with relational refinement.

ABSTRACT

3D Semantic Scene Completion (SSC) aims to infer scene geometry and semantics from partial observations, providing dense 3D scene understanding for autonomous driving. Existing completion methods do not consistently adapt their completion strategies to different observation conditions and make limited use of object-level occlusion relations to recover missing scene information. However, occlusion limits the available visual evidence, and simply aggregation multi-frame features does not explicitly model directed occlusion relations between objects. To address these challenges, we propose MemGraphSSC, a temporal SSC framework that combines observation-conditioned memory with occlusion-relation-guided refinement. We introduce two complementary components: (1) an Observation-Conditioned Dual Memory module that retrieves historical region features for history-supported voxels and learned class prototypes for observation-unsupported voxels, while preserving visual initialization for current-supported voxels; and (2) an Occlusion-Relation-Guided Refinement module that uses depth-consistent relations from precomputed scene graphs to construct voxel priors and approximate object supports, then refines features in occluded regions through directed messages between relation endpoints. Experimental results show that the proposed method achieves superior performance on SemanticKITTI and SSCBench-KITTI-360 datasets, outperforming existing state-of-the-art vision-based SSC methods.

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.

1 INTRODUCTION

Humans can infer scene structure and semantics from incomplete visual observations using contextual knowledge. Semantic scene completion (SSC) seeks to endow intelligent systems with a similar capability by jointly predicting dense 3D occupancy and semantic labels across observed and unobserved regions. By extending scene understanding beyond directly visible surfaces, SSC supports applications in autonomous driving (Behley et al., 2019; Wei et al., 2023; Xia et al., 2023) and robotic perception and navigation (Liang et al., 2021; Wu et al., 2020; Schmid et al., 2022).

SSC has been studied using depth images(Zhang et al., 2018; Wang et al., 2019), LiDAR point clouds(Yan et al., 2021; Cheng et al., 2021; Rist et al., 2021), and RGB images(Cao & De Charette, 2022; Liu et al., 2025). SSCNet (Song et al., 2017) pioneered volumetric completion from a single depth image, while LiDAR-based methods such as LMSCNet (Roldao et al., 2020) exploit direct geometric measurements despite sparse input. Camera-based approaches instead use dense appearance cues: MonoScene (Cao & De Charette, 2022) lifts monocular image features into voxels, VoxFormer (Li et al., 2023) densifies depth-supported sparse queries, and FoundationSSC (Chen et al., 2026) incorporates pretrained semantic and geometric priors. Temporal methods further recover missing information from historical observations. HTCL (Li et al., 2024a) selects and refines context through cross-frame affinities, while ConSSC (Xue et al., 2026) combines historical voxel refinement with temporal semantic aggregation. However, enriching temporal context through feature aggregation and pose-based alignment does not by itself ensure that historical features provide relevant evidence for every voxel lacking current-frame support. For regions without surface support in either the current or historical frames, placing greater emphasis on historical features cannot compensate for the lack of direct evidence and may introduce misleading cues. This motivates a central question: how can historical memory provide relevant evidence where available, while learned semantic priors support completion where historical observations are insufficient?

To address this challenge, we propose MemGraphSSC, a camera-based temporal SSC framework that combines observation-conditioned memory with occlusion-guided relational refinement. As illustrated in Figure 1, the key idea is to assign complementary evidence sources according to voxel observation states and incorporate relational context from scene occlusion graphs into memory-enhanced features. Specifically, the Observation-Conditioned Dual-Memory module distinguishes three mutually exclusive states: current-supported, history-supported, and observation-unsupported. It preserves current visual features, retrieves historical region features, and queries learned semantic prototypes for these states, respectively. The Occlusion-Relation-Guided Feature Refinement module further converts precomputed image-derived scene graphs into voxel priors and approximate object supports, using directed occlusion relations to refine features in occluded regions. Experiments on SemanticKITTI and SSCBench-KITTI-360 demonstrate the effectiveness of MemGraphSSC, which achieves state-of-the-art performance among the compared methods. Our main contributions are summarized as follows:

- We introduce MemGraphSSC, a camera-based temporal SSC framework that integrates observation-conditioned memory with directed occlusion relations to improve scene completion under limited observations.
- We propose an Observation-Conditioned Dual Memory module that retrieves historical evidence for history-supported voxels and learned semantic prototypes for observation-unsupported voxels, providing complementary information for completion under different observation conditions.
- We introduce Occlusion-Relation-Guided Feature Refinement, which incorporates scene-graph priors into memory-enhanced voxel features and uses directed object relations to refine occluded-object support regions.
- We evaluate MemGraphSSC on SemanticKITTI and SSCBench-KITTI-360, with quantitative comparisons, component ablations, and qualitative results demonstrating its effectiveness, including improved reconstruction of occluded regions and regions outside the current field of view.

2 Related Work

2.1 Camera-Based Semantic Scene Completion

Semantic scene completion (SSC) jointly predicts volumetric occupancy and semantic labels from partial observations. Following the depth-based SSCNet (Song et al., 2017), camera-based methods develop image-to-voxel representations: MonoScene (Cao & De Charette, 2022) projects features along viewing rays, VoxFormer (Li et al., 2023) densifies depth-supported sparse queries, and OccFormer (Zhang et al., 2023) combines local and global transformer pathways. CGFormer (Yu et al., 2024) further reduces projection ambiguity through context-dependent queries and depth-aware aggregation. Foundation models provide complementary priors: FoundationStereo (Wen et al., 2025) offers transferable stereo representations, while FoundationSSC (Chen et al., 2026) adapts semantic and geometric priors through decoupled pathways and complementary view transformations. Building on FoundationSSC, we supplement its voxel features with observation-conditioned memory and explicit occlusion relations.

2.2 Temporal Evidence and Observation-Aware Completion

Temporal SSC exploits complementary observations across frames. VoxFormer-T (Li et al., 2023) aggregates historical image features, HTCL (Li et al., 2024a) selects temporal context through cross-frame affinities and dynamic refinement, and Hi-SOP (Li et al., 2026) separately aligns geometric and temporal context. FlowScene (Wang et al., 2025b) combines optical-flow alignment with occlusion-guided voxel refinement. Observation-aware methods further distinguish available evidence: VisHall3D (Lu et al., 2025) separates visible-region reconstruction from invisible-region hallucination, while ProtoFormer (Fang et al., 2026) introduces prototype-guided attention for point-cloud SSC. Our method assigns different evidence sources according to current and historical surface support, retrieving historical region tokens for history-supported voxels and semantic prototypes for observation-unsupported voxels. These depth-derived states describe surface support rather than physical visibility or occupancy.

2.3 Semantic Context and Object-Relation Reasoning

Contextual SSC methods model voxel relations (MonoScene (Cao & De Charette, 2022)) or use instance queries to connect image and scene representations (Symphonies (Jiang et al., 2024)). Vision-language approaches introduce additional semantic context: VLScene (Wang et al., 2025a) combines distilled vision-language guidance with geometric and semantic interactions, while LOMA (Cui et al., 2025) integrates language and visual features through a triplane Mamba architecture. Explicit scene graphs encode objects and their relations; 3DSSG (Wald et al., 2020), for example, learns such graphs from reconstructed indoor point clouds. Our method uses precomputed occlusion relations to construct voxel priors and approximate endpoint supports. Directed messages derived from endpoint features, classes, and relation attributes refine object-support regions, complementing the evidence supplied by observation-conditioned memory.

3 Method

3.1 Problem Formulation and Framework Overview

In this work, camera-based semantic scene completion (SSC) predicts a semantic volume from stereo image sequences. Let $\mathcal{T} = \{t\} \cup \mathcal{H}$ index the current frame t and valid historical frames $\mathcal{H}$, with stereo inputs $\mathcal{I}_\mathcal{T} = \{(\mathbf{I}_i^L, \mathbf{I}_i^R)\}_{i\in\mathcal{T}}$. Given camera calibration $\mathcal{K}$, ego poses $\mathcal{P}$, and precomputed image-derived scene graphs $\mathcal{G}_\mathcal{T}$, MemGraphSSC predicts:

$$\hat{\mathbf{Y}} = \mathcal{F}_\theta(\mathcal{I}_\mathcal{T}, \mathcal{K}, \mathcal{P}, \mathcal{G}_\mathcal{T}), \tag{1}$$

where θ denotes the model parameters and $\hat{\mathbf{Y}} \in \{0, \dots, C\}^{X\times Y\times Z}$ is defined in the current coordinate frame. Label 0 denotes empty space, and $1, \dots, C$ denote occupied categories. Scene graphs are used during both training and inference; ground-truth voxel labels are used only for supervision.

As shown in Figure 2, pose-aligned lifting maps current and historical image features onto a common voxel grid, and temporal aggregation produces $\mathbf{F}_{\text{temp}}$ (Sec. 3.2). The aligned historical features are

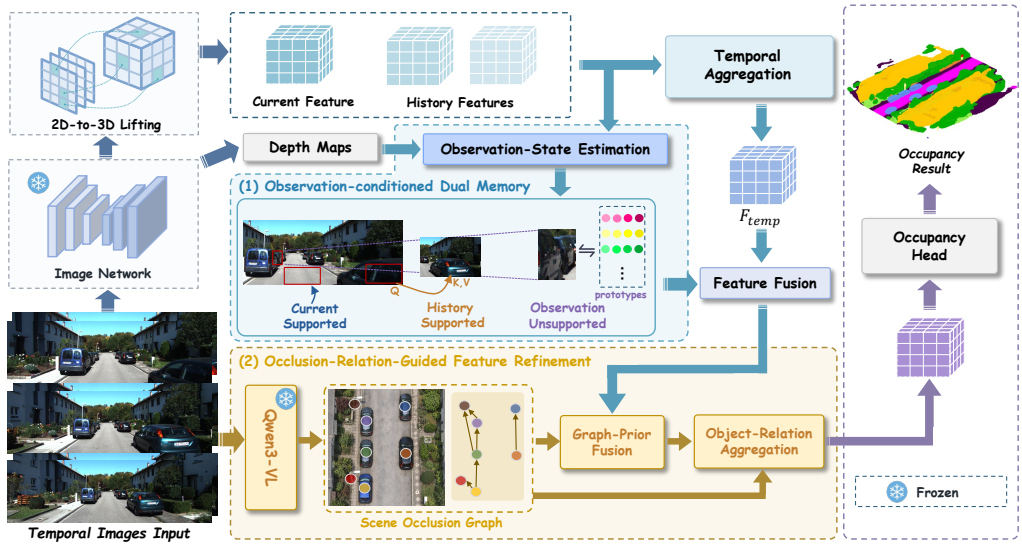

Figure 2: Overview of MemGraphSSC. (1) Observation-Conditioned Dual Memory preserves current visual evidence and retrieves historical features or semantic prototypes according to voxel observation states. (2) Occlusion-Relation-Guided Feature Refinement integrates scene-graph priors and directed occlusion relations into memory-enhanced features. The occupancy head predicts the completed semantic volume.

used to construct episodic memory, while $\mathbf{F}_{\text{temp}}$ provides queries for semantic prototype retrieval and a prior for voxel initialization. Observation-conditioned dual memory (Sec. 3.3) preserves visual seeds or retrieves historical evidence or semantic prototypes according to voxel observation support. The resulting voxel features are fused with $\mathbf{F}_{\text{temp}}$ to obtain memory-enhanced features $\mathbf{F}_{\text{mem}}$. Occlusion-relation-guided refinement (Sec. 3.4) incorporates priors from scene graphs precomputed using Qwen3-VL-8B and updates object-support features through directed relations. The refined features are then passed to the inherited 3D encoder and SSC head.

3.2 Pose-Aligned Temporal Evidence Aggregation

To integrate features across frames, we design a Pose-Aligned Temporal Evidence Aggregation module. For each stereo pair, the feature extraction pipeline built on FoundationStereo (Wen et al., 2025) produces semantic context features and a depth distribution. Historical features are lifted directly into the current ego frame using

$$\mathbf{T}_{i\to t} = \mathbf{T}_{w\leftarrow t}^{-1}\mathbf{T}_{w\leftarrow i}, \tag{2}$$

where $\mathbf{T}_{w\leftarrow i}$ maps ego-frame coordinates at time i to world coordinates. Composing this transform with the camera-to-ego calibration before lifting yields aligned voxel features $\mathbf{F}_i(v)$ without an additional voxel-grid warp. Here, v indexes the $X_f \times Y_f \times Z_f$ feature grid unless otherwise stated. Alignment compensates for ego-motion but does not explicitly model independent object motion.

Voxel-wise attention scores each historical frame using the current and historical features, their absolute difference, and an embedding of $\log(1+|t-i|)$. Softmax normalization over valid frames yields weights for the aggregated history $\mathbf{H}(v)$. A sigmoid gate computed from $\mathbf{F}_t(v)$, $\mathbf{H}(v)$, and their absolute difference controls fusion:

$$\mathbf{F}_{\text{temp}}(v) = \mathbf{F}_t(v) + g_{\text{temp}}(v)\big(\mathbf{H}(v) - \mathbf{F}_t(v)\big). \tag{3}$$

Without valid history, the output defaults to $\mathbf{F}_t(v)$.

3.3 Observation-Conditioned Dual-Memory

To adapt completion strategies to different observation states, we introduce an Observation-Conditioned Dual-Memory module.

Observation-conditioned dual memory selects evidence for voxel initialization using depth-derived binary surface-support masks $V_i(v)$ on the current feature grid. Observation-State Estimation assigns each voxel to one of three mutually exclusive states: current-supported (CS), with current-frame support; history-supported (HS), with historical but no current support; or observation-unsupported (OU), with neither within the sampled window. These states describe estimated surface support, not ground-truth occupancy or ray-traced visibility. OU can therefore include empty space or missed surfaces and does not necessarily indicate occlusion or a lack of observations outside the window.

3.3.1 Region-Based Episodic Memory

For HS voxels, an episodic bank retrieves evidence from supported historical regions. A token $\mathbf{z}_{ij}$ averages aligned features $\mathbf{F}_i(v)$ over supported voxels in region j of frame i. Unsupported regions are excluded, and those with the highest support fractions are retained within a per-frame token budget. Retained tokens form a bank indexed by n. Each normalized key $\mathbf{k}_n$ combines projected token content with spatial-position and temporal-offset embeddings, using $\log(1 + |t - i|)$ for the offset; its value $\mathbf{u}_n$ is a learned token projection. Each HS voxel forms a normalized query $\mathbf{q}_v$ from its current-frame feature $\mathbf{F}_t(v)$ and a spatial-position embedding. Retrieval selects up to K keys by dot-product similarity and aggregates their values:

$$\mathbf{F}_{\text{epi}}(v) = \sum_{n \in \mathcal{N}_K(v)} \frac{\exp(\mathbf{q}_v^\top \mathbf{k}_n / \tau)}{\sum_{n' \in \mathcal{N}_K(v)} \exp(\mathbf{q}_v^\top \mathbf{k}_{n'} / \tau)} \mathbf{u}_n, \tag{4}$$

where $\mathcal{N}_K(v)$ indexes the selected tokens and τ is the retrieval temperature. The bank is rebuilt for each sample and permits retrieval beyond the same spatial coordinate without maintaining a persistent map.

3.3.2 Class-Structured Semantic Prototypes

For OU voxels, semantic memory provides M learnable key–value prototypes per class $c \in \{0, \ldots, C\}$, including empty space. Let $\boldsymbol{\kappa}_{cm}$ and $\boldsymbol{\nu}_{cm}$ denote the normalized key and learnable value of prototype m for class c. A normalized query $\mathbf{q}_v^p$, projected from $\mathbf{F}_{\text{temp}}(v)$, is compared with all keys using dot-product similarity divided by τ. A softmax over all classes and prototypes weights their values to produce $\mathbf{F}_{\text{pro}}(v)$, without using ground-truth labels to select prototypes. Class logits aggregate similarities within each class:

$$\ell_c^p(v) = \log \sum_{m=1}^{M} \exp\left(\frac{(\mathbf{q}_v^p)^\top \boldsymbol{\kappa}_{cm}}{\tau} \right). \tag{5}$$

These logits receive auxiliary supervision during training (Sec. 3.5) to encourage class structure.

3.3.3 State-Conditioned Feature Fusion

For voxels without current proposal support (HS and OU), the inherited head constructs a prior $\mathbf{P}_0(v) = \text{MLP}(\mathbf{F}_{\text{temp}}(v))$ and blends it with retrieved memory:

$$\mathbf{P}_0'(v) = (1 - \gamma(v))\mathbf{P}_0(v) + \gamma(v)\mathbf{M}(v). \tag{6}$$

Here, $\mathbf{M}(v)$ is $\mathbf{F}_{\text{epi}}(v)$ for HS voxels and $\mathbf{F}_{\text{pro}}(v)$ for OU voxels, with $\gamma(v)$ taking the corresponding fixed value γ_{HS} or γ_{OU}. CS voxels retain visual seed features without memory injection. After voxel self-attention, the inherited fusion module combines the resulting $\mathbf{F}_{\text{vox}}$ with $\mathbf{F}_{\text{temp}}$ to obtain $\mathbf{F}_{\text{mem}}$ for relational refinement.

3.4 Occlusion-Relation-Guided Feature Refinement

To incorporate image-derived occlusion relations into scene completion, we introduce Occlusion-Relation-Guided Feature Refinement (OFR), which projects directed scene-graph relations into the voxel grid and uses them to refine voxel features.

3.4.1 From Scene Graphs to Voxel Supports

Precomputed Qwen3-VL-8B scene graphs provide entity classes, image boxes, depth estimates, and directed relations with confidence and depth-consistency information. Entities form nodes, and edges $r = (s, \text{occludes}, o)$ point from foreground occluders s to background objects o. Relations are filtered by supported endpoint classes, confidence, depth ordering, and projected support.

For a retained current-frame relation, let $B_s(v)$ and $B_o(v)$ indicate projection into the padded subject and object boxes, $A(v)$ indicate a valid camera projection, and $z(v)$ denote camera depth. The endpoint depths are d_s, d_o. The image-region weight is $b_r(v) = B_s(v)B_o(v)$ when the projected box masks have a nonempty intersection, and $b_r(v) = \frac{1}{2}B_o(v)$ otherwise. Spatial quality q_r is 1 in the overlap case and 0.5 in the fallback case. Depth constraints define the relation-region weight and endpoint supports:

$$\begin{aligned} w_r(v) &= c_r b_r(v) A(v) \mathbf{1}[d_s + \delta \le z(v) \le d_o + \Delta_o], \\ U_s^r(v) &= c_r B_s(v) A(v) \mathbf{1}[|z(v) - d_s| \le \Delta_s], \\ U_o^r(v) &= w_r(v) \mathbf{1}[z(v) \ge d_o - \delta]. \end{aligned} \tag{7}$$

Here, c_r is relation confidence and $\mathbf{1}[\cdot]$ is the indicator function. The margin δ sets the lower depth thresholds, Δ_s is the subject depth-band half-width, and Δ_o controls the extension beyond the object depth. The weight w_r constructs the graph prior, while U_s^r and U_o^r support endpoint pooling and object-region updates. These regions require no voxel annotations and approximate entities rather than exact instance boundaries.

3.4.2 Graph-Prior Fusion

For subject and object classes c_s, c_o, a fixed ordered class-pair encoding $\mathbf{e}(c_s, c_o)$ is weighted by $w_r(v)$. Summing contributions yields the frame-level prior; their maximum spatial weight defines confidence. Historical 2D relations are independently lifted directly onto the current unified voxel grid using their corresponding camera projection parameters.

Let $\mathbf{P}_g^{\text{cur}}, Q_g^{\text{cur}}$ denote the current-frame prior and confidence, and $\mathbf{P}_g^f, Q_g^f$ their historical counterparts. For $\mathcal{H}^+ \subseteq \mathcal{H}$ containing frames with nonzero confidence after filtering and projection, we combine them as

$$\begin{aligned} \mathbf{P}_g &= \mathbf{P}_g^{\text{cur}} + \frac{\eta_h}{|\mathcal{H}^+|} \sum_{f \in \mathcal{H}^+} \mathbf{P}_g^f, \\ Q_g(v) &= \max\left\{ Q_g^{\text{cur}}(v), \max_{f \in \mathcal{H}^+} Q_g^f(v) \right\}. \end{aligned} \tag{8}$$

We set $\eta_h = 0.5$. If $\mathcal{H}^+$ is empty, $\mathbf{P}_g = \mathbf{P}_g^{\text{cur}}$ and $Q_g = Q_g^{\text{cur}}$. Empty graph sets contribute zero; confidence is not scaled by η_h.

Gated residual fusion incorporates the prior into memory-enhanced features:

$$\mathbf{F}_g = \mathbf{F}_{\text{mem}} + \alpha_g \left[W_g \mathbf{P}_g \odot \sigma(G_g([\mathbf{P}_g, Q_g])) \odot Q_g \right]. \tag{9}$$

Here, W_g, G_g are voxel-wise projections, $[\cdot,\cdot]$ denotes channel concatenation, σ is sigmoid, and $\odot$ is element-wise multiplication with confidence broadcast across channels. The learnable scalar α_g is initialized to zero.

3.4.3 Directed Object-Relation Aggregation

Object-relation aggregation (ORA) refines $\mathbf{F}_g$ using accepted current-frame relations $\mathcal{R}_{\text{ORA}}$ with sufficient endpoint support and spatial quality. Each endpoint support is normalized by its maximum. Support-weighted pooling followed by layer normalization yields descriptors $\mathbf{h}_s^r, \mathbf{h}_o^r$. A learned function ϕ generates a directed message:

$$\mathbf{m}_r = \tanh(\phi([\mathbf{h}_s^r, \mathbf{h}_o^r, \mathbf{h}_s^r - \mathbf{h}_o^r, \mathbf{e}_s, \mathbf{e}_o, \mathbf{a}_r])), \tag{10}$$

where $\mathbf{e}_s, \mathbf{e}_o$ are learned role-specific class embeddings. The attributes $\mathbf{a}_r$ comprise c_r, q_r, and the depth gap $d_o - d_s$, clipped to $[0, D_{\max}]$ and divided by $D_{\max}$. A learned sigmoid gate g_r, conditioned

on $\mathbf{h}_o^r$, $\mathbf{m}_r$, and $\mathbf{a}_r$, controls injection into normalized object supports $\bar{U}_o^r(v)$:

$$\mathbf{R}_o(v) = \frac{\sum_{r\in\mathcal{R}_{\mathrm{ORA}}} \bar{U}_o^r(v)\, g_r\, c_r q_r\, \mathbf{m}_r}{\max\big(1, \sum_{r\in\mathcal{R}_{\mathrm{ORA}}} \bar{U}_o^r(v)\big)},$$
$$\mathbf{F}_{\mathrm{out}}(v) = \mathbf{F}_g(v) + \beta \tanh(\alpha_o)\mathbf{R}_o(v). \tag{11}$$

The scalar α_o is initialized to zero, and β bounds the update magnitude per channel. Updates are confined to object supports; without accepted relations, $\mathbf{F}_{\mathrm{out}} = \mathbf{F}_g$. The inherited 3D encoder and SSC head predict voxel labels from $\mathbf{F}_{\mathrm{out}}$.

3.5 Training Objective

We augment the baseline objective with prototype and semantic-modulated graph-local supervision:

$$\mathcal{L} = \mathcal{L}_{\mathrm{base}} + \lambda_p \mathcal{L}_{\mathrm{pro}} + \lambda_g \mathcal{L}_{\mathrm{graph}}, \tag{12}$$

where $\mathcal{L}_{\mathrm{base}}$ retains the weighted voxel cross-entropy, semantic and geometric scaling, depth-distribution, and auxiliary image-semantic losses. The prototype loss $\mathcal{L}_{\mathrm{pro}}$ applies class-weighted cross-entropy at OU voxels with valid labels, resized to the feature grid by nearest-neighbor interpolation.

On the prediction grid, graph-local supervision is weighted by confidence and semantic agreement:

$$\mathcal{L}_{\mathrm{graph}} = \frac{\sum_{v\in\mathcal{V}_{\mathrm{valid}}} \omega(v) a_{Y(v)} [-\log p_{Y(v)}(v)]}{\sum_{v\in\mathcal{V}_{\mathrm{valid}}} \omega(v) a_{Y(v)}},$$
$$\omega(v) = \tilde{Q}_g(v)\big(1 + \mu E(v)\big). \tag{13}$$

Here, $Y(v)$ is the ground-truth label, $\mathcal{V}_{\mathrm{valid}} = \{v : Y(v) \neq 255\}$, a_c is the baseline class weight, and $p_c(v)$ is the predicted class probability. Confidence $\tilde{Q}_g$ is interpolated to the prediction grid. Eligible relations are current-frame relations with $q_r = 1$ and sufficient object-class agreement on the feature grid. To obtain $E(v)$, their object supports U_o^r are interpolated to the prediction grid, clipped to $[0, 1]$, and masked by agreement between $Y(v)$ and the relation's object class; the voxel-wise maximum is taken across eligible relations. These supports are used before ORA's per-support normalization. Empty maxima and zero-denominator losses are zero.

Semantic agreement multiplies confidence-based voxel weights by a factor in $[1, 1 + \mu]$, without changing labels or inference-time support selection. No separate relation-prediction loss is used. We set $\lambda_p = 0.1$, $\lambda_g = 1$, and $\mu = 0.2$.

4 Experiments

We evaluate the proposed method on SemanticKITTI(Behley et al., 2019) and SSCBench-KITTI-360(Li et al., 2024b), two large-scale benchmarks for outdoor semantic scene completion. Following the standard evaluation protocol for each benchmark, we compare its performance with recent state-of-the-art methods. We build our framework upon FoundationSSC (Chen et al., 2026), using its pretrained FoundationStereo backbone (Wen et al., 2025) for stereo feature extraction. Occlusion scene graphs are precomputed using a frozen Qwen3-VL-8B model (Bai et al., 2025), which is not updated during SSC training. We optimize the trainable parameters using AdamW (Loshchilov & Hutter, 2017) with a weight decay of 0.01. Training is conducted on three NVIDIA A100 GPUs. Further implementation details are provided in Appendix B.

4.1 Quantitative Results

Table 6 reports results on the SemanticKITTI hidden test set. MemGraphSSC achieves the highest IoU and mIoU among the compared methods, reaching 51.30% and 20.90%, respectively. Compared with the temporal method ConSSC, it improves IoU by 3.13 percentage points and mIoU by 1.70 percentage points. It also surpasses FoundationSSC, the strongest single-frame method in the table, by 3.18 and 1.58 percentage points, respectively. At the category level, MemGraphSSC achieves the highest IoU for road, building, car, motorcycle, other-vehicle, vegetation, trunk, terrain, person, and

Methods	Input	IoU	road	sidewalk	parking	other-grnd	building	car	truck	bicycle	motocycle	other-vehicle	vegetation	trunk	terrain	person	bicylist	motorcyclist	fence	pole	traf.-sign	mIoU
MonoScene	S	34.16	54.7	27.1	24.8	5.7	14.4	18.8	3.3	0.5	0.7	4.4	14.9	2.4	19.5	1.0	1.4	0.4	11.1	3.3	2.1	11.08
TPVFormer	S	34.25	55.1	27.2	27.4	6.5	14.8	19.2	3.7	1.0	0.5	2.3	13.9	2.6	20.4	1.1	2.4	0.3	11.0	2.9	1.5	11.26
OccFormer	S	34.53	55.9	30.3	31.5	6.5	15.7	21.6	1.2	1.5	1.7	3.2	16.8	3.9	21.3	2.2	1.1	0.2	11.9	3.8	3.7	12.32
VoxFormer	T	43.21	54.1	26.9	25.1	7.3	23.5	21.7	3.6	1.9	1.6	4.1	24.4	8.1	24.2	1.6	1.1	0.0	13.1	6.6	5.7	13.41
H2GFormer	T	44.20	56.4	28.6	26.5	4.9	22.8	23.4	4.8	0.8	0.9	4.1	24.6	9.1	23.8	1.2	2.5	0.1	13.3	6.4	6.3	13.72
HASSC-T	T	42.87	55.3	29.6	25.9	11.3	23.1	23.0	2.9	1.9	1.5	4.9	24.8	9.8	26.5	1.4	3.3	0.0	14.3	7.0	7.1	14.38
Symphonize	S	42.19	58.4	29.3	26.9	11.7	24.7	23.6	3.2	3.6	2.6	5.6	24.2	10.0	23.1	3.2	1.9	2.0	16.1	7.7	8.0	15.04
BRGScene	S	43.34	61.9	31.2	30.7	10.7	24.2	22.8	2.8	3.4	2.4	6.1	23.8	8.4	27.0	2.9	2.2	0.5	16.5	7.0	7.2	15.36
SOAP	T	47.54	63.2	36.6	35.8	**16.3**	30.3	28.4	4.3	4.6	2.9	4.9	31.6	15.0	33.2	2.3	0.9	0.1	20.6	11.5	**13.3**	18.72
FoundationSSC	S	48.12	65.7	36.3	33.5	11.6	30.5	28.2	**9.0**	5.6	6.9	10.0	28.6	13.7	32.8	2.8	**4.7**	**2.5**	23.1	9.8	11.8	19.32
ConSSC	T	48.17	64.5	**37.9**	**37.0**	15.5	31.8	29.5	0.2	**5.7**	2.7	6.6	31.7	15.4	33.9	2.0	0.4	0.0	22.5	**12.5**	**13.3**	19.20
Ours	T	**51.30**	**67.1**	37.6	34.8	15.4	**34.2**	**29.8**	8.1	5.4	**9.7**	**11.1**	**33.1**	**15.7**	**35.2**	**3.8**	2.4	1.7	**26.5**	12.3	**13.3**	**20.90**

Table 1: Quantitative results on the SemanticKITTI hidden test set. The "S" and "T" represent the single-frame images and temporal images inputs, respectively. **Bold** and underlined highlight the best and second-best results, respectively.

Methods	IoU	car	bicycle	motocycle	truck	other-vehicle	person	road	parking	sidewalk	other-grnd	building	fence	vegetation	terrain	pole	traf.-sign	other-struct.	other-obj.	mIoU
MonoScene	37.87	19.3	0.4	0.6	8.0	2.0	0.9	48.4	11.4	28.1	3.3	32.9	3.5	26.2	16.8	6.9	5.7	4.2	3.1	12.31
VoxFormer	38.76	17.8	1.2	0.9	4.6	2.1	1.6	47.0	9.7	27.2	2.9	31.2	5.0	29.0	14.7	6.5	6.9	3.8	2.4	11.91
TPVFormer	40.22	21.6	1.1	1.4	8.1	2.6	2.4	53.0	12.0	31.1	3.8	34.8	4.8	30.1	17.5	7.5	5.9	5.5	2.7	13.64
OccFormer	40.27	22.6	0.7	0.3	9.9	3.8	2.8	54.3	13.4	31.5	3.6	36.4	4.8	31.0	19.5	7.8	8.5	7.0	4.6	13.81
Symphonies	44.12	30.0	1.9	5.9	25.1	12.1	8.2	54.9	13.8	32.8	6.9	35.1	8.6	38.3	11.5	14.0	9.6	14.4	11.3	18.58
SOAP	48.48	30.0	3.3	4.4	7.8	6.0	5.9	60.7	17.5	**40.1**	6.3	**45.5**	10.9	40.9	**24.9**	17.2	19.4	12.9	9.7	20.17
FoundationSSC	48.61	30.5	**4.9**	8.4	24.9	13.0	7.2	63.6	17.8	39.8	5.7	43.6	10.5	39.1	24.8	16.2	**20.3**	11.5	10.2	21.78
ConSSC	48.79	32.0	1.1	3.3	22.8	13.1	9.0	62.3	**18.2**	37.2	**8.8**	43.3	9.7	42.4	13.3	15.4	13.1	16.6	13.8	20.85
Ours	**49.73**	**35.7**	1.3	**9.0**	**26.0**	18.1	**11.0**	**64.4**	18.1	38.3	**8.8**	41.8	**11.4**	**42.6**	14.4	**17.6**	**14.5**	**18.2**	**15.1**	**22.55**

Table 2: Quantitative results on the SSCBench-KITTI360 test set. **Bold** and underlined highlight the best and second-best results, respectively.

fence, and ties for the best traffic-sign result. These results support the effectiveness of combining observation-conditioned memory with occlusion-guided relational refinement for both geometric completion and semantic prediction.

As shown in Table 2, we achieve the best results on SSCBench-KITTI-360 among the compared methods, with an IoU of 49.72 and an mIoU of 22.55. Its consistently strong performance across both benchmarks supports its effectiveness in outdoor semantic scene completion.

4.2 Qualitative Results

Figure 3 presents qualitative comparisons with ConSSC and HTCL-S on the SemanticKITTI validation set. In the first row, our method produces a more complete scene reconstruction, with better-preserved structures of buildings occluded by vehicles and vegetation, as well as improved reconstruction of distant vehicles. In the second row, it delineates clearer boundaries between successive, partially occluded vehicles. In the third row, it reconstructs and semantically labels road regions and other objects outside the current field of view, illustrating the benefit of historical evidence. Together, these examples support the effectiveness of our method in exploiting temporal information and reconstructing occluded regions.

4.3 Ablation Studies

All ablation experiments are conducted on the SemanticKITTI validation set. Table 3 evaluates the contributions of temporal aggregation, memory recall, OFR, and ORA on the SemanticKITTI validation set. Temporal aggregation improves IoU and mIoU over the baseline by 1.74 and 0.92 percentage points, respectively. Adding memory further increases these metrics by 0.92 and 0.83

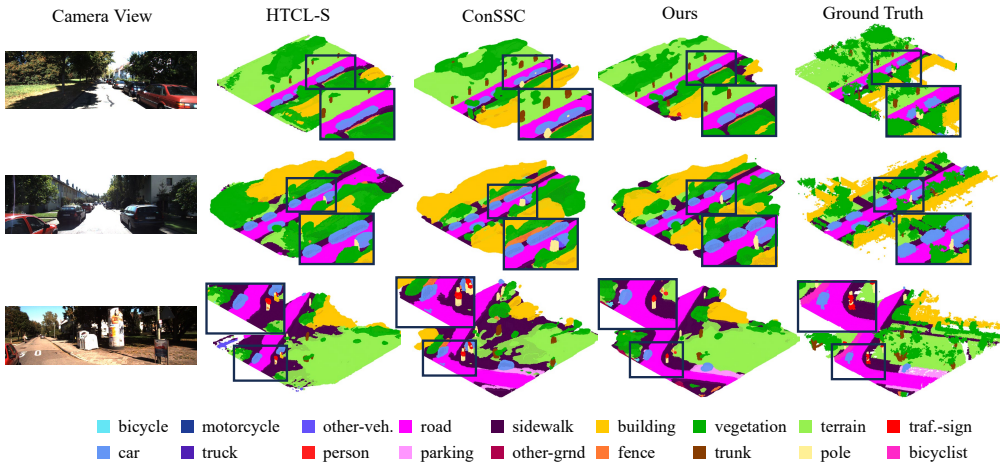

Figure 3: Visualization results of our MemGraphSSC and the state-of-the-art methods, include HTCL and ConSSC.

Table 3: Ablation of architecture components on SemanticKITTI validation set.

Method	Temporal	Mem.	OFR	ORA	**IoU**	**mIoU**
Baseline					47.91	20.36
(a)	✓				49.65	21.28
(b)	✓	✓			50.57	22.11
(c)	✓	✓	✓		50.30	22.39
(d)	✓		✓	✓	50.12	22.27
(e)	✓	✓	✓	✓	**50.63**	**22.58**

Table 4: Ablation of observation-conditioned memory routing on SemanticKITTI validation set.

Routing	**Memory source**		**IoU**	**mIoU**
	HS	OU		
No memory	–	–	50.12	22.27
Episodic only	Episodic	Episodic	50.07	22.08
Prototype only	Prototype	Prototype	50.10	22.13
Ours	Episodic	Prototype	**50.63**	**22.58**

points. Introducing OFR yields an additional 0.28-point mIoU gain, although IoU decreases by 0.27 points. Adding ORA to this configuration improves both metrics, achieving the best results among the evaluated variants: 50.63 IoU and 22.58 mIoU. Removing memory from the full model reduces IoU and mIoU by 0.51 and 0.31 points, respectively, indicating that memory remains beneficial alongside relation-guided refinement. Overall, the full model exceeds the baseline by 2.72 IoU and 2.22 mIoU points.

Table 4 evaluates the effect of memory routing. Assigning episodic memory to HS voxels and semantic prototypes to OU voxels achieves the best performance, with 50.63% IoU and 22.58% mIoU, improving over the no-memory variant by 0.51 and 0.31 percentage points, respectively. Using episodic memory or prototypes for both states reduces mIoU to 22.08% and 22.13%, respectively, below the no-memory result of 22.27%. These results indicate that memory injection alone does not guarantee better completion and support matching the memory source to the observation state: historical evidence for HS voxels and category-level priors for OU voxels.

5 Conclusion

We introduce MemGraphSSC, a camera-based 3D semantic scene completion framework that integrates historical evidence with occlusion relations. The Observation-Conditioned Dual Memory module addresses different observation limitations by retrieving historical evidence for history-supported voxels and drawing on globally learned semantic prototypes for observation-unsupported voxels. Scene graphs encoding directed occlusion relations further guide feature refinement in occluded regions. MemGraphSSC achieves state-of-the-art performance on SemanticKITTI, with qualitative results showing improved reconstruction of regions outside the current field of view and those affected by occlusion. These findings highlight the value of combining temporal evidence with relational context for semantic scene completion.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Jurgen Gall. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9297–9307, 2019.

Anh-Quan Cao and Raoul De Charette. Monoscene: Monocular 3d semantic scene completion. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3981–3991. IEEE, 2022.

Shiyuan Chen, Wei Sui, Bohao Zhang, Zeyd Boukhers, John See, and Cong Yang. Unleashing semantic and geometric priors for 3d scene completion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 3020–3028, 2026.

Ran Cheng, Christopher Agia, Yuan Ren, Xinhai Li, and Liu Bingbing. S3cnet: A sparse semantic scene completion network for lidar point clouds. In *Conference on robot learning*, pp. 2148–2161. PMLR, 2021.

Yubo Cui, Zhiheng Li, Jiaqiang Wang, and Zheng Fang. Loma: Language-assisted semantic occupancy network via triplane mamba. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 2609–2617, 2025.

Chenghao Fang, Jianqing Liang, Jiye Liang, Zijin Du, and Feilong Cao. Point cloud semantic scene completion with prototype-guided transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 3822–3830, 2026.

Haoyi Jiang, Tianheng Cheng, Naiyu Gao, Haoyang Zhang, Tianwei Lin, Wenyu Liu, and Xinggang Wang. Symphonize 3d semantic scene completion with contextual instance queries. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20258–20267. IEEE, 2024.

Bohan Li, Jiajun Deng, Wenyao Zhang, Zhujin Liang, Dalong Du, Xin Jin, and Wenjun Zeng. Hierarchical temporal context learning for camera-based semantic scene completion. In *European Conference on Computer Vision*, pp. 131–148. Springer, 2024a.

Bohan Li, Jiajun Deng, Yasheng Sun, Xiaofeng Wang, Xin Jin, and Wenjun Zeng. Hierarchical context alignment with disentangled geometric and temporal modeling for semantic occupancy prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

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 *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9087–9098. IEEE, 2023.

Yiming Li, Sihang Li, Xinhao Liu, Moonjun Gong, Kenan Li, Nuo Chen, Zijun Wang, Zhiheng Li, Tao Jiang, Fisher Yu, et al. Sscbench: A large-scale 3d semantic scene completion benchmark for autonomous driving. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 13333–13340. IEEE, 2024b.

Yiqing Liang, Boyuan Chen, and Shuran Song. Sscnav: Confidence-aware semantic scene completion for visual semantic navigation. In *2021 IEEE international conference on robotics and automation (ICRA)*, pp. 13194–13200. IEEE, 2021.

Xianzhu Liu, Haozhe Xie, Shengping Zhang, Hongxun Yao, Rongrong Ji, Liqiang Nie, and Dacheng Tao. 2d semantic-guided semantic scene completion. *International Journal of Computer Vision*, 133(3):1306–1325, 2025.

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

Haoang Lu, Yuanqi Su, Xiaoning Zhang, Longjun Gao, Yu Xue, and Le Wang. Vishall3d: Monocular semantic scene completion from reconstructing the visible regions to hallucinating the invisible regions. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 28674–28684. IEEE, 2025.

Christoph B Rist, David Emmerichs, Markus Enzweiler, and Dariu M Gavrila. Semantic scene completion using local deep implicit functions on lidar data. *IEEE transactions on pattern analysis and machine intelligence*, 44(10):7205–7218, 2021.

Luis Roldao, Raoul De Charette, and Anne Verroust-Blondet. Lmscnet: Lightweight multiscale 3d semantic completion. In *2020 International Conference on 3D Vision (3DV)*, pp. 111–119. IEEE, 2020.

Lukas Schmid, Mansoor Nasir Cheema, Victor Reijgwart, Roland Siegwart, Federico Tombari, and Cesar Cadena. Sc-explorer: Incremental 3d scene completion for safe and efficient exploration mapping and planning. *arXiv preprint arXiv:2208.08307*, 2022.

Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 190–198. IEEE, 2017.

Johanna Wald, Helisa Dhamo, Nassir Navab, and Federico Tombari. Learning 3d semantic scene graphs from 3d indoor reconstructions. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3960–3969. IEEE, 2020.

Meng Wang, Huilong Pi, Ruihui Li, Yunchuan Qin, Zhuo Tang, and Kenli Li. Vlscene: vision-language guidance distillation for camera-based 3d semantic scene completion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7808–7816, 2025a.

Meng Wang, Fan Wu, Ruihui Li, Qin Yunchuan, Zhuo Tang, and Li Ken Li. Learning temporal 3d semantic scene completion via optical flow guidance. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 5745–5771. Curran Associates, Inc., 2025b. doi: 10.52202/085713-0203.

Yida Wang, David Joseph Tan, Nassir Navab, and Federico Tombari. Forknet: Multi-branch volumetric semantic completion from a single depth image. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8607–8616. IEEE, 2019.

Yi Wei, Linqing Zhao, Wenzhao Zheng, Zheng Zhu, Jie Zhou, and Jiwen Lu. Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21672–21683. IEEE, 2023.

Bowen Wen, Matthew Trepte, Joseph Aribido, Jan Kautz, Orazio Gallo, and Stan Birchfield. Foundationstereo: Zero-shot stereo matching. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5249–5260. IEEE, 2025.

Shun-Cheng Wu, Keisuke Tateno, Nassir Navab, and Federico Tombari. Scfusion: Real-time incremental scene reconstruction with semantic completion. In *2020 International Conference on 3D Vision (3DV)*, pp. 801–810. IEEE, 2020.

Zhaoyang Xia, Youquan Liu, Xin Li, Xinge Zhu, Yuexin Ma, Yikang Li, Yuenan Hou, and Yu Qiao. Scpnet: Semantic scene completion on point cloud. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17642–17651. IEEE, 2023.

Yujie Xue, Meng Wang, Ruihui Li, Fan Wu, Zhizhong Liu, Zhuo Tang, and Kenli Li. Learning spatial-temporal consistency for 3d semantic scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28567–28577, 2026.

Xu Yan, Jiantao Gao, Jie Li, Ruimao Zhang, Zhen Li, Rui Huang, and Shuguang Cui. Sparse single sweep lidar point cloud segmentation via learning contextual shape priors from scene completion. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 3101–3109, 2021.

Zhu Yu, Runmin Zhang, Jiacheng Ying, Junchen Yu, Xiaohai Hu, Lun Luo, Si-Yuan Cao, and Hui-Liang Shen. Context and geometry aware voxel transformer for semantic scene completion. *Advances in Neural Information Processing Systems*, 37:1531–1555, 2024.

Jiahui Zhang, Hao Zhao, Anbang Yao, Yurong Chen, Li Zhang, and Hongen Liao. Efficient semantic scene completion network with spatial group convolution. In *European conference on computer vision*, pp. 749–765. Springer, 2018.

Yunpeng Zhang, Zheng Zhu, and Dalong Du. Occformer: Dual-path transformer for vision-based 3d semantic occupancy prediction. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9399–9409. IEEE, 2023.

Appendix

Outline. Appendix A details the observation-state routing and occlusion-relation pipeline. Appendix B records implementation choices. Appendices C and D provide extended results and controlled analyses. Appendix E presents further visual comparisons, followed by limitations and future work in Appendix F.

TABLE OF CONTENTS

A Additional Method Details 14

A.1 Observation-State and Dual-Memory Implementation 14

A.2 Scene-Graph Construction Pipeline 14

A.3 Projection and Accepted Voxel Supports 17

A.4 Relation Refinement and Ground-Truth Calibration 17

B Implementation Details 17

B.1 Datasets 17

B.2 Metrics 17

C Additional Quantitative Results 18

C.1 Results in SemanticKITTI Validation 18

C.2 Observation-State Statistics 19

C.3 Scene-Graph Reliability and Support Coverage 21

C.4 SemanticKITTI Validation by Class 21

D Additional Ablation Studies 22

D.1 Semantic Modulation Strength 22

D.2 Semantic Prototype Number 22

D.3 Relation Filtering and Model Parameters 23

E Additional Qualitative Results 23

F Limitations and Future Work 32

A Additional Method Details

This appendix supplies reproducibility details and analyses that do not fit in the main text. We use the main-paper terminology throughout: MemGraphSSC, Observation-Conditioned Dual Memory and Occlusion-Relation-Guided Feature Refinement (OFR). OFR consists of Graph-Prior Fusion and Object-Relation Aggregation (ORA). A voxel is *directly supported* only when an estimated depth surface falls inside it; free space along the viewing ray is not counted as direct surface support or as ground-truth occupancy.

A.1 Observation-State and Dual-Memory Implementation

For current and valid historical depth-surface supports $V_t(v)$ and $V_i(v)$, the three routing masks are

$$\begin{aligned} M_{\mathrm{CS}}(v) &= V_t(v), \\ M_{\mathrm{HS}}(v) &= (1 - V_t(v))\mathbf{1}\left[\bigvee_{i\in\mathcal{H}} V_i(v)\right], \\ M_{\mathrm{OU}}(v) &= 1 - M_{\mathrm{CS}}(v) - M_{\mathrm{HS}}(v). \end{aligned} \tag{14}$$

Invalid historical frames are excluded before constructing the masks. The masks are mutually exclusive. In particular, OU means that no direct depth surface was found *within the sampled window*; it does not imply that the voxel is occupied or necessarily occluded.

The episodic bank divides each valid historical feature grid into $4 \times 4 \times 2$ regions. It averages features only over supported voxels, then retains at most 256 tokens per historical frame in order of support fraction. An HS query retrieves the three most similar tokens with cosine similarity and temperature 0.1. The bank is rebuilt for every sample and is not a persistent scene map. OU queries four learnable key–value prototypes per class, including the empty class among SemanticKITTI's 20 labels. No voxel ground truth is used to choose a prototype at inference. The fixed masked-prior mixing gates are 0.25 for episodic retrieval and 0.05 for prototypes. Current-supported visual seeds retain the inherited path. The auxiliary prototype classification loss uses ground truth only at valid-label OU voxels during training.

A.2 Scene-Graph Construction Pipeline

The image-to-voxel pipeline has five steps. Its detector and relation model are frozen; the JSON graph is generated for each image before SSC training.

1. Entity detection and category mapping. For each left RGB image, Grounding DINO base (the pinned `IDEA-Research/grounding-dino-base` checkpoint) uses the text prompt `car . truck . bus . van . trailer . tram . person . bicycle . motorcycle .`. Box and text thresholds are 0.25; class-agnostic duplicate suppression uses box IoU 0.8. Duplicate or overmerged boxes are removed, vehicle synonyms are mapped to the corresponding SemanticKITTI thing class, and person–cycle pairs are merged into rider classes where applicable. Each remaining entity has an `obj_N` identifier, class, detection confidence, and pixel-coordinate box in `entities`.

2. Visual relation proposal. The RGB image is annotated with thin boxes and numeric identifiers, then sent with an entity list (identifier, class, box) to frozen Qwen3-VL-8B-Instruct. The local Transformers server loads `Qwen/Qwen3-VL-8B-Instruct` in `bfloat16` with automatic device placement and exposes a localhost OpenAI-compatible endpoint. Detector and VLM inference run on separate GPUs in the extraction setup. Calls are serialized (concurrency one), with temperature 0 and at most 1,024 output tokens; the configured top-p is 0.001 but is inactive under greedy decoding. The deployed Chinese prompt is translated below; the runtime also appends the detected entity list:

> Inspect the image itself. A numeric label on each detection box maps to `obj_N` in the entity list; boxes identify objects but do not establish occlusion. Report only visible partial contour occlusions between detected objects. The nearer occluder is `subject_id`, and the farther occluded object is `object_id`. Do not infer direction from box size or containment. Relative image position, apparent size, and detail are only weak depth cues;

> actual contour occlusion determines the output. Adjacent, nonoccluding objects must not be reported. Return only a JSON array of directed `occludes` relations with confidence in $[0, 1]$; return `[]` when none is visible.

The required response is a JSON array such as

```
[{"subject_id":"obj_0",
"predicate":"occludes",
"object_id":"obj_1",
"confidence":0.95}]
```

The VLM receives neither a depth map nor metric-depth values.

3. Relation validation and first JSON record. The parser retains directed `occludes` pairs with existing entity identifiers, finite confidence $c_r \geq 0.85$, and positive-area overlap of the *original* boxes. The per-image JSON contains `sequence`, `frame_id`, `image_path`, `entities`, and `relationships`; `caption_summary` is a human-readable diagnostic field. A diagnostic overlay is saved separately. These records represent candidate relations, not depth-verified training edges.

4. Metric-depth verification. A separate pass reads the precomputed metric-depth map aligned to the same image; it does not rerun the VLM or estimate depth from its response. Each endpoint depth is the median of at least 20 finite, positive pixels inside its box. For subject and object depths d_s, d_o, an edge is consistent if $d_o - d_s > 0.5\,\mathrm{m}$, inverted if $d_s - d_o > 0.5\,\mathrm{m}$, and tied otherwise; insufficient valid pixels form a fourth, rejected state. Only deduplicated consistent edges remain in `relationships`. The depth-checked JSON also retains all checked candidates in `relationships_all_depth_checked`, with `subject_depth_m`, `object_depth_m`, `depth_delta_m`, and `depth_consistency`, plus `depth_filter` metadata. The stored depth delta is $d_s - d_o$ (negative for an accepted edge), whereas the method equations use $d_o - d_s$. The original-box overlap check is distinct from the later 8-pixel dilation used for voxel projection.

5. SSC input conversion. The dataset loader retrieves the depth-checked JSON for the current image and valid history images, and serializes their graphs as `scene_graphs_json` in sample metadata. Missing or malformed records yield an empty graph rather than invented relations. At training and inference, accepted endpoints and directed edges are projected with camera calibration and poses to the current coarse voxel grid; the resulting relation prior and endpoint supports feed Graph-Prior Fusion and ORA as specified in Sections A.3 and A.4. Audit-only fields and the diagnostic caption do not supply model features. Caching avoids repeated detector/VLM calls across training epochs; a new scene still needs graph generation and depth verification before prediction.

Figure 4 pairs representative SemanticKITTI RGB inputs with their automatically produced scene-graph diagnostics. Numbered boxes and relation markings are not manually verified occlusion ground truth; training uses only edges that pass the filters described above.



Figure 4: Nine image-derived scene-graph examples. Each row pairs the original RGB image (left) with its automatic diagnostic overlay (right). Only depth-consistent accepted relations enter subsequent training.

A.3 Projection and Accepted Voxel Supports

Voxel centers in the current grid are projected with the calibration and poses of the frame that supplied a relation. Current and historical edges therefore reach the same current-coordinate grid. We retain only positive camera depth and projections inside endpoint boxes dilated by 8 pixels. The subject depth band is $d_s \pm 1.0\,\mathrm{m}$, and the object-side relation region may extend to $d_o + 3.0\,\mathrm{m}$. If the two projected endpoint boxes share support on the discrete grid, their intersection is used. If not, the object box is used with half spatial weight. These cases receive spatial quality 1 and 0.5, respectively. A relation requires at least 32 nonzero coarse-grid relation voxels; ORA additionally requires at least four supported voxels at *each* endpoint.

Only historical frames with nonzero accepted graph support enter the mean of historical priors; that mean is multiplied by 0.5. Empty historical frames are excluded from the denominator. Graph-Prior Fusion may use current and historical relations, whereas ORA and semantic modulation use current-frame endpoint supports. All these supports approximate object regions from boxes, depth, and projection. They are not precise 3D instance masks.

A.4 Relation Refinement and Ground-Truth Calibration

Graph-Prior Fusion projects the fixed ordered class-pair encoding and its confidence to voxel features, then adds a gated residual to the memory-enhanced feature. Its learned residual scale starts at zero. ORA pools features over normalized subject and object supports, forms a directed subject-to-object message, and writes a bounded residual only to object support. The maximal residual scale is 0.2; the edge depth gap is clipped at $50\,\mathrm{m}$. When no relation passes the ORA filters, it leaves the feature unchanged. Thus, memory supplies node content, while the graph supplies the sparse directed structure used to refine that content.

The auxiliary $\mathcal{L}_{\mathrm{graph}}$ in the main-paper objective reweights cross-entropy against the *existing* voxel ground-truth label; it is not a relation-classification loss, and no VLM label replaces a ground-truth target. Additional semantic modulation is allowed only for a current-frame relation with shared box support (spatial quality 1) and at least 16 object-support voxels whose ground truth agrees with the relation's object class. The maximum eligible support per voxel bounds its multiplier to $[1, 1+\mu]$, with $\mu = 0.2$ in the main run and $\lambda_g = 1$. Ground truth is used for this calibration only during training, never for graph generation at inference. The prototype-class auxiliary $\mathcal{L}_{\mathrm{pro}}$ is a separate branch.

B Implementation Details

Table 5 summarizes the main configuration. Temporal aggregation uses the current frame t and two historical frames, $t-3$ and $t-6$.

B.1 Datasets

The SemanticKITTI dataset includes dense semantic scene completion annotations and labels a voxelized scene with 20 semantic classes. It consists of 10 training sequences, 1 validation sequence, and 11 testing sequences. RGB images are resized to 1280×384 for input processing. The SSCBench-KITTI-360 dataset contains 7 training sequences, 1 validation sequence, and 1 testing sequence, covering 19 semantic classes in total. The RGB images are resized to 1408×384 for input processing.

B.2 Metrics

We use intersection over union (IoU) to evaluate the scene completion performance. To assess the effectiveness of our 3D Semantic Scene Completion method, we focus on the mean IoU (mIoU). A higher IoU value reflects accurate geometric predictions, while a higher mIoU value indicates more precise semantic segmentation.

Table 5: Network architecture and training hyperparameters for the SemanticKITTI main experiment.

Configurations	Values
Network architecture	
Backbone	FoundationSSC; FoundationStereo-derived stereo features
Temporal input	$t, t-3, t-6$; left/right stereo images
Image / output voxel grid	384×1280 / $256 \times 256 \times 32$
Coarse graph grid / channels	$128 \times 128 \times 16$ / 128
Episodic region / token budget / top-K	$4 \times 4 \times 2$ / 256 per history / 3
Semantic prototypes	Four per class; 20 classes including empty
Episodic / prototype mixing gates	0.25 / 0.05
Training hyperparameters	
Optimizer / weight decay	AdamW / 0.01
Maximum learning rate / schedule	3×10^{-4} / OneCycleLR
Optimizer steps / seed	35,000 / 7,240
GPUs / per-GPU batch / accumulation	3 / 1 / none
Train / validation workers per process	8 / 6
Distributed normalization	DDP with SyncBatchNorm
Graph construction and refinement	
Detector / relation model	Grounding DINO base / Qwen3-VL-8B-Instruct
Box/text thresholds; duplicate NMS IoU	0.25/0.25; 0.8
Relation confidence / depth gap	≥ 0.85 / 0.5 m
Valid depth pixels / box dilation	≥ 20 / 8 pixels
Relation / ORA endpoint support	≥ 32 / ≥ 4 coarse voxels
Subject / object depth extent	± 1.0 m / up to $d_o + 3.0$ m
Historical graph-prior factor	0.5 over active history frames
Graph-local loss λ_g / semantic μ	1.0 / 0.2
Semantic ground-truth match threshold	16 coarse voxels
ORA maximum residual scale	0.2
Prototype auxiliary-loss coefficient λ_p	0.1

C Additional Quantitative Results

C.1 Results in SemanticKITTI Validation

Methods	IoU	**road** (15.30%)	**sidewalk** (11.13%)	**parking** (1.12%)	**other-grnd** (0.56%)	**building** (14.10%)	**car** (3.92%)	**truck** (0.16%)	**bicycle** (0.03%)	**motocycle** (0.03%)	**other-vehicle** (0.20%)	**vegetation** (39.3%)	**trunk** (0.51%)	**terrain** (9.17%)	**person** (0.07%)	**bicylist** (0.07%)	**motorcyclist** (0.05%)	**fence** (3.90%)	**pole** (0.29%)	**traf.-sign** (0.08%)	**mIoU**
MonoScene	36.86	56.5	26.7	14.3	0.5	14.1	23.3	7.0	0.6	0.5	1.5	17.9	2.8	29.6	1.9	1.2	**0.0**	5.8	4.1	2.3	11.08
TPVFormer	35.61	56.5	25.9	20.6	0.9	13.9	23.8	8.1	0.4	0.0	4.4	16.9	2.3	30.4	0.5	0.9	**0.0**	5.9	3.1	1.5	11.36
OccFormer	36.50	58.9	26.9	19.6	0.3	14.4	25.1	25.5	0.8	1.2	8.5	19.6	3.9	32.6	2.8	2.8	**0.0**	5.6	4.3	2.9	13.46
VoxFormer	44.15	53.6	26.5	19.7	0.4	19.5	26.5	7.3	1.3	0.6	7.8	26.1	6.1	33.0	1.9	2.0	**0.0**	7.3	9.2	4.9	13.35
H2GFormer	44.57	56.1	29.1	17.8	0.5	19.7	28.2	10.0	0.5	0.5	7.4	26.3	6.8	34.4	1.5	2.9	**0.0**	7.2	7.9	4.7	13.73
Symphonize	41.92	56.4	27.6	15.3	0.9	21.6	28.7	20.4	2.5	2.8	13.9	25.7	6.6	30.9	3.5	2.2	**0.0**	8.4	9.6	5.8	14.89
HTCL-S	45.51	63.7	32.5	23.3	0.1	24.1	34.3	20.7	4.0	2.8	12.0	27.0	8.8	37.7	2.6	2.3	**0.0**	11.2	11.5	7.0	17.13
Ours	**50.63**	**68.3**	**36.2**	**30.2**	**2.0**	**32.4**	**39.8**	**43.7**	**7.4**	**8.2**	**24.0**	**31.3**	**13.6**	**42.0**	**5.3**	**3.6**	**0.0**	**14.1**	**16.6**	**10.3**	**22.58**

Table 6: Quantitative results on the SemanticKITTI validation set. **Bold** denotes the best performance.

Table 6 shows that our method achieves the highest IoU and mIoU among the compared methods, reaching 50.63% and 22.58%, respectively. Compared with HTCL-S, these results represent gains of 5.12 and 5.45 percentage points. Our method also achieves the highest class-wise IoU in 18 of the 19 semantic categories, while all methods score zero for motorcyclist. Particularly large improvements over the strongest competing result are observed for truck (+18.2 points), other-vehicle

Table 7: Class-wise observation-state distributions on the SemanticKITTI training split.

Class	CS (%)	HS (%)	OU (%)
car	20.51	8.98	70.51
bicycle	13.17	9.58	77.25
motorcycle	20.63	9.70	69.66
truck	18.49	8.96	72.55
other-vehicle	14.96	8.19	76.84
person	11.82	6.78	81.40
bicyclist	4.25	2.20	93.55
motorcyclist	6.43	2.03	91.54
road	40.40	15.21	44.39
parking	31.32	12.57	56.11
sidewalk	27.00	10.74	62.26
other-ground	24.01	13.34	62.65
building	12.31	7.47	80.21
fence	26.94	12.22	60.84
vegetation	16.76	9.90	73.34
trunk	21.22	12.06	66.72
terrain	27.84	13.34	58.82
pole	20.53	12.44	67.03
traffic-sign	19.90	13.45	66.65
Voxel-weighted	22.67	10.89	66.44
Class-macro	19.92	9.96	70.12

Only valid occupied GT voxels (classes 1–19) are included. Each class row is normalized by its own voxel count. Voxel-weighted pools counts across classes and frames; class-macro averages the 19 class-wise percentages equally. CS, HS, and OU indicate current-surface coverage, history-only surface coverage, and coverage by neither, respectively. Rows sum to 100% up to rounding. These are depth-derived states, not ground-truth occlusion labels.

(+10.1 points), and building (+8.3 points), demonstrating improvements across both vehicle and static scene categories.

C.2 Observation-State Statistics

Protocol and definitions. We analyse the epoch-15 checkpoint on all 3,834 frames of the SemanticKITTI training split (sequences 00–07, 09, and 10) in evaluation mode, with random image and voxel augmentations disabled. The current frame t and valid historical frames at offsets $t-3$ and $t-6$ are used. Let $C(v)$ denote coverage by a current-frame stereo-depth surface and $H(v)$ denote coverage by at least one valid historical stereo-depth surface after pose alignment. The three mutually exclusive states are

$$\mathrm{CS} = C, \qquad \mathrm{HS} = \neg C \wedge H, \qquad \mathrm{OU} = \neg C \wedge \neg H. \tag{15}$$

These masks mark depth-surface support, rather than all free-space voxels traversed by camera rays. No mask dilation is applied. State masks are computed on a $128 \times 128 \times 16$ grid and aligned to the $256 \times 256 \times 32$ GT grid by nearest-neighbour replication. Only GT classes 1–19 are counted; empty voxels (0) and invalid labels (255) are excluded from every table below. This yields 403,113,668 frame–voxel instances. Counts are accumulated in each frame's voxel grid without world-space deduplication. This is a count of frame–voxel instances, not independent samples or unique world-space voxels.

Let $N_{c,s}$ be the pooled count for class c and state s, and $N_c = \sum_s N_{c,s}$. Class-wise percentages are $100N_{c,s}/N_c$; voxel-weighted percentages pool counts before normalization, whereas class-macro percentages average the 19 class-wise percentages equally. All results are descriptive statistics from one checkpoint and one split.

Class-wise observation coverage. Table 7 shows that 22.67% of valid occupied voxel instances receive current-frame surface support, while 10.89% receive support only from the available history. Their union covers 33.56%, leaving 66.44% in OU. Historical observations therefore add 10.89

Table 8: Class frequency, the composition of OU, and the contribution of historical observations.

Class	Voxel instances	Class share (%)	OU share (%)	History share (%)
car	15,783,539	3.915	4.155	30.45
bicycle	125,136	0.031	0.036	42.10
motorcycle	118,809	0.029	0.031	31.99
truck	646,799	0.160	0.175	32.63
other-vehicle	821,951	0.204	0.236	35.38
person	262,978	0.065	0.080	36.46
bicyclist	283,696	0.070	0.099	34.05
motorcyclist	204,750	0.051	0.070	23.98
road	61,688,703	15.303	10.224	27.35
parking	4,502,961	1.117	0.943	28.65
sidewalk	44,883,650	11.134	10.434	28.46
other-ground	2,269,923	0.563	0.531	35.71
building	56,840,218	14.100	17.024	37.77
fence	15,719,652	3.900	3.571	31.20
vegetation	158,442,623	39.305	43.386	37.15
trunk	2,061,623	0.511	0.514	36.23
terrain	36,970,523	9.171	8.119	32.40
pole	1,151,988	0.286	0.288	37.74
traffic-sign	334,146	0.083	0.083	40.33
All classes	403,113,668	100.000	100.000	32.45

Class share is $N_c / \sum_j N_j$; OU share is $N_{c,\mathrm{OU}} / \sum_j N_{j,\mathrm{OU}}$; history share is $N_{c,\mathrm{HS}}/(N_{c,\mathrm{CS}} + N_{c,\mathrm{HS}})$. All ratios are multiplied by 100. OU share measures class composition within OU, whereas Table 7 measures OU within each class. Counts are frame–voxel instances; repeated physical surfaces across frames are not deduplicated. Empty and invalid GT voxels are excluded.

Table 9: History availability and occupied-voxel observation states.

Valid history frames	Frames	Frame share (%)	Voxel instances	CS (%)	HS (%)	OU (%)
0	10	0.26	1,083,812	24.82	0.00	75.18
1	10	0.26	1,091,248	24.84	7.28	67.88
2	3,814	99.48	400,938,608	22.66	10.93	66.41

History validity is read from the masks saved during inference for offsets $t-3$ and $t-6$. State percentages pool occupied-voxel counts within each group. The groups contain different frames; this table is a data-availability audit, not a controlled history-length ablation.

percentage points of coverage under this state definition; this is not a measured improvement in completion accuracy. Coverage varies substantially by class: road has 40.40% CS and 15.21% HS, whereas bicyclist and motorcyclist have OU fractions of 93.55% and 91.54%, respectively. The class-macro OU fraction is 70.12%, compared with the voxel-weighted value of 66.44%, a difference of 3.68 percentage points.

Class imbalance and the role of history. Table 8 separates within-class coverage from the class composition of the dataset. Vegetation and building account for 53.40% of occupied voxel instances and 60.41% of OU instances. Bicyclist has the highest within-class OU fraction (93.55%) and contributes 0.099% of all OU instances. Vegetation contributes the largest fraction of all OU instances (43.386%). Among instances supported by either the current frame or history, 32.45% receive support exclusively from history. This fraction reaches 42.10% for bicycle and 40.33% for traffic-sign. Their corresponding CS+HS coverage fractions are 22.75% and 33.35%, respectively.

History availability. Table 9 records 3,814 frames (99.48%) with both historical inputs valid, 10 frames with one valid historical input, and 10 frames with neither valid. The two-valid-history subset has CS, HS, and OU fractions of 22.66%, 10.93%, and 66.41%, respectively. Its OU fraction differs from the all-frame value by 0.03 percentage points. HS is zero for every frame with no valid history. Across all frames, HS constitutes 14.08% of the occupied instances without current-frame support, as measured by $N_{\mathrm{HS}}/(N_{\mathrm{HS}} + N_{\mathrm{OU}})$.

Interpretation and scope. OU denotes absence of predicted depth-surface support from the selected frames. The statistics describe this model-defined partition on the training split. They do not measure ground-truth physical occlusion, test-set accuracy, or the causal effect of temporal memory on completion performance. Frame quantiles are descriptive distribution summaries; no significance tests or independence assumptions at the voxel level are used.

C.3 Scene-Graph Reliability and Support Coverage

VLM-predicted direction is not geometric ground truth. We therefore audit the depth ordering of all 16,915 candidates with valid endpoint depths from 23,201 SemanticKITTI train/val graphs; 9,801 graphs contain at least one candidate. Using the $0.5\,\mathrm{m}$ margin defined in Section A.2, "consistent" means the proposed occluder is nearer, "inverted" means it is farther, and "tie" means the two median depths differ by at most the margin. Table 10 reports these mutually exclusive states; only consistent relations are retained for subsequent voxel projection.

Table 10: Depth-consistency audit of raw SemanticKITTI train/val relations.

Depth ordering	Relations	Fraction
Consistent	13,337	78.85%
Inverted	2,860	16.91%
Tie	718	4.24%
Total	16,915	100%

Projection must retain enough support to affect supervised voxels. Table 11 counts candidates on the labeled frames after confidence and depth screening. A 32-voxel minimum retains 1,886 of 1,955 training candidates at confidence 0.85 while covering 1,231 labeled frames. Raising confidence to 0.90 reduces coverage; the main run retains 0.85. Image-level graph availability is therefore not the same as usable voxel-level support.

Table 11: Projected support retention on labeled SemanticKITTI frames.

Split / policy	Eligible	Zero support	≥ 32	Covered frames
Train, $c_r \geq 0.85$	1,955	60	1,886	1,231 / 3,834
Val, $c_r \geq 0.85$	710	5	700	428 / 815
Train, $c_r \geq 0.90$	1,784	56	1,720	not measured

Overall consistency can conceal long-tail failures. Table 12 breaks candidates down by the VLM object class. Car dominates, whereas several rare classes have lower estimated consistency from small samples. This motivates depth checking and cautions against naively upweighting inverse class frequency.

Table 12: Depth consistency by candidate object class, before filtering.

Object class	Valid-depth edges	Consistent rate
car	15,363	79.71%
truck	395	84.05%
other-vehicle	473	80.55%
motorcycle	86	77.91%
person	425	53.65%
bicycle	127	47.24%
bicyclist	38	52.63%
motorcyclist	8	37.50%

C.4 SemanticKITTI Validation by Class

The main text reports aggregate metrics; Tables 13 and 14 give per-class IoU at each run's mIoU-best checkpoint. Relative to Temporal+Memory, the complete model gains 3.966 points on truck,

6.004 on parking, and 1.295 on bicyclist, while bicycle, motorcycle, and fence lose 2.528, 1.308, and 1.150 points. Thus, the aggregate gain has class trade-offs. Global class IoU alone cannot prove that improvements occur specifically on occluded voxels. No matching per-class log exists for the Temporal reference, so its class results are not reconstructed. In the tables, "OFR prior" contains Graph-Prior Fusion without ORA, while "OFR" includes both stages.

Table 13: Thing-class IoU (%) on SemanticKITTI validation set

Configuration	car	bicycle	motorcycle	truck	other-vehicle	person	bicyclist	motorcyclist
Temporal+Memory	39.055	9.893	9.526	39.777	23.626	4.792	2.304	0.000
+OFR prior	39.839	9.381	8.859	40.389	24.549	5.357	2.820	0.000
+OFR	39.793	7.365	8.218	43.743	24.026	5.314	3.599	0.000
Temporal+OFR	39.958	8.105	9.965	38.652	24.678	4.180	2.758	0.000

Table 14: Stuff-class IoU (%) on SemanticKITTI validation set

Configuration	road	parking	sidewalk	other-ground	building	fence	vegetation	trunk	terrain	pole	traffic-sign
Temporal+Memory	67.997	24.227	36.202	1.351	31.978	15.298	31.694	13.399	42.350	16.194	10.496
+OFR prior	68.851	29.982	37.455	2.212	31.514	14.895	30.996	12.252	39.833	16.154	10.143
+OFR	68.326	30.231	36.219	1.971	32.386	14.148	31.292	13.574	42.008	16.601	10.258
Temporal+OFR	68.857	26.690	36.741	1.481	31.942	15.149	31.460	12.970	44.496	15.684	9.425

D Additional Ablation Studies

D.1 Semantic Modulation Strength

We compare the graph-local semantic-modulation coefficient μ to assess whether stronger weighting improves both occupancy and semantic completion. Table 15 reports each run's mIoU-best checkpoint and the IoU measured at that *same* step; it does not compare separately selected IoU-best checkpoints. The highest mIoU is 22.583 at $\mu = 0.2$, while the highest paired IoU is 50.792 at $\mu = 0.1$. Relative to $\mu = 0.2$, $\mu = 0.3$ lowers mIoU by 0.698 points and paired IoU by 0.272 points. These validation results do not show a consistent gain from stronger modulation; they do not by themselves identify the cause of the trade-off.

Table 15: Graph-local semantic modulation on SemanticKITTI validation set (%). Each row is selected by mIoU; IoU is measured at the same step.

μ	Step	mIoU	IoU
0.1	21,725	22.320	**50.792**
0.2	20,447	**22.583**	50.633
0.3	26,837	21.885	50.361

D.2 Semantic Prototype Number

More prototypes may improve capacity but can also add redundant categories. Table 16 reports the three configurations with complete records. Four prototypes per class perform best among these settings.

Table 16: Effect of prototypes per class on SemanticKITTI validation (%).

Prototypes number M	IoU	mIoU
3	50.34	21.93
4	**50.63**	**22.58**
5	50.59	22.22
6	49.92	21.86

D.3 Relation Filtering and Model Parameters

We measure how depth verification and the minimum projected-voxel support affect relation availability without retraining the SSC model. Table 17 reports two successive populations: the first row audits 16,915 train/validation candidate relations with valid depths; the remaining rows use training relations that already pass depth ordering, class mapping, and the indicated relation-confidence threshold c_r. Here c_r is the VLM's confidence in the directed relation, and "support voxels" counts nonzero cells in its projected coarse-grid relation region. The 1,955 candidates at $c_r \geq 0.85$ are therefore not the full raw set. Raising the minimum support from 1 to 32 voxels retains nine fewer of these 1,955 candidates. These counts describe filter retention, not accuracy after training with alternative thresholds.

Table 17: Relation retention at successive depth and projected-support filters. Denominators follow the eligible population described in the text.

Filter	Threshold	Retained / eligible	Fraction
Depth ordering	$d_o - d_s > 0.5\,\mathrm{m}$	13,337 / 16,915	78.85%
Projected support	$c_r \geq 0.85$; ≥ 32 voxels	1,886 / 1,955	96.47%
Projected support	$c_r \geq 0.90$; ≥ 32 voxels	1,720 / 1,784	96.41%
Projected support	$c_r \geq 0.85$; min. 1, 4, or 8 voxels	1,895 / 1,955	96.93%
Projected support	$c_r \geq 0.85$; ≥ 16 voxels	1,894 / 1,955	96.88%
Projected support	$c_r \geq 0.85$; ≥ 32 voxels	1,886 / 1,955	96.47%

Table 18 accounts for the four successive additions in the main-paper ablation: Temporal aggregation, Observation-Conditioned Dual Memory, Graph-Prior Fusion, and ORA. Counts are relative to the same FoundationSSC backbone and distinguish trainable from all instantiated parameters. Memory contains 16,578 instantiated but frozen routing parameters; the other three additions have no such difference. Fixed class-pair encoding and graph-local loss reweighting add no parameters. These counts do not measure graph-extraction time or inference latency.

Table 18: Parameter increments across the SemanticKITTI ablation sequence. Cumulative counts are relative to FoundationSSC; "all" includes frozen parameters in the added modules.

Configuration	Added trainable	Added all	Cumulative all
FoundationSSC	–	–	0
+Temporal	12,690	12,690	12,690
+Memory	44,000	60,578	73,268
+Graph-Prior Fusion (OFR prior)	33,025	33,025	106,293
+ORA (full OFR)	104,578	104,578	210,871
Total added	194,293	210,871	210,871

E Additional Qualitative Results

The panels compare ground truth, Ours (MemGraphSSC), ConSSC, and HTCL-S in a fixed order. They show overall SSC outputs, not verified graph edges or an occlusion-only mask. Figure 5 adds five aligned scenes to the main-paper examples; these visual differences cannot establish a relation-specific gain or replace per-class IoU.

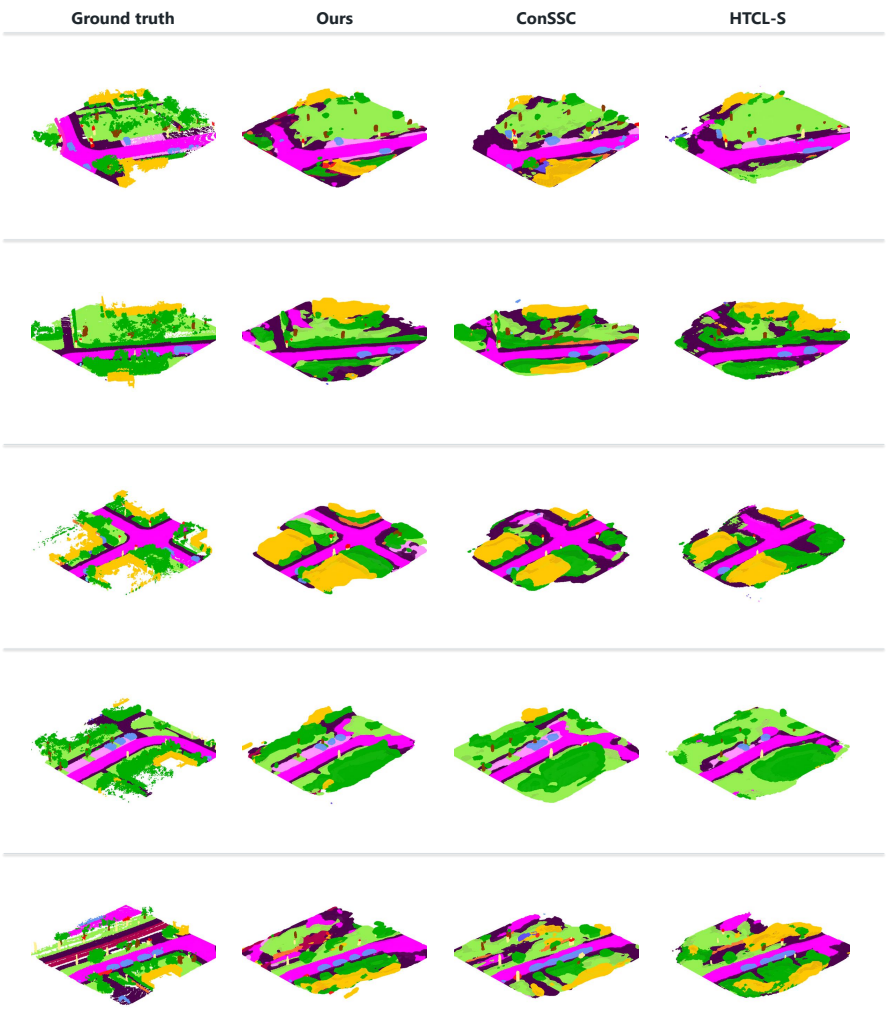

Figure 5: Additional voxel-volume comparisons on the SemanticKITTI validation set. Columns show ground truth, Ours, ConSSC, and HTCL-S; each row uses the same scene and viewpoint.

Figure 6 adds five different scenes with the same column ordering and color map. Some methods look similar in certain regions, while others show noticeable structural differences; selected examples should not be interpreted as a guarantee of universal improvement.

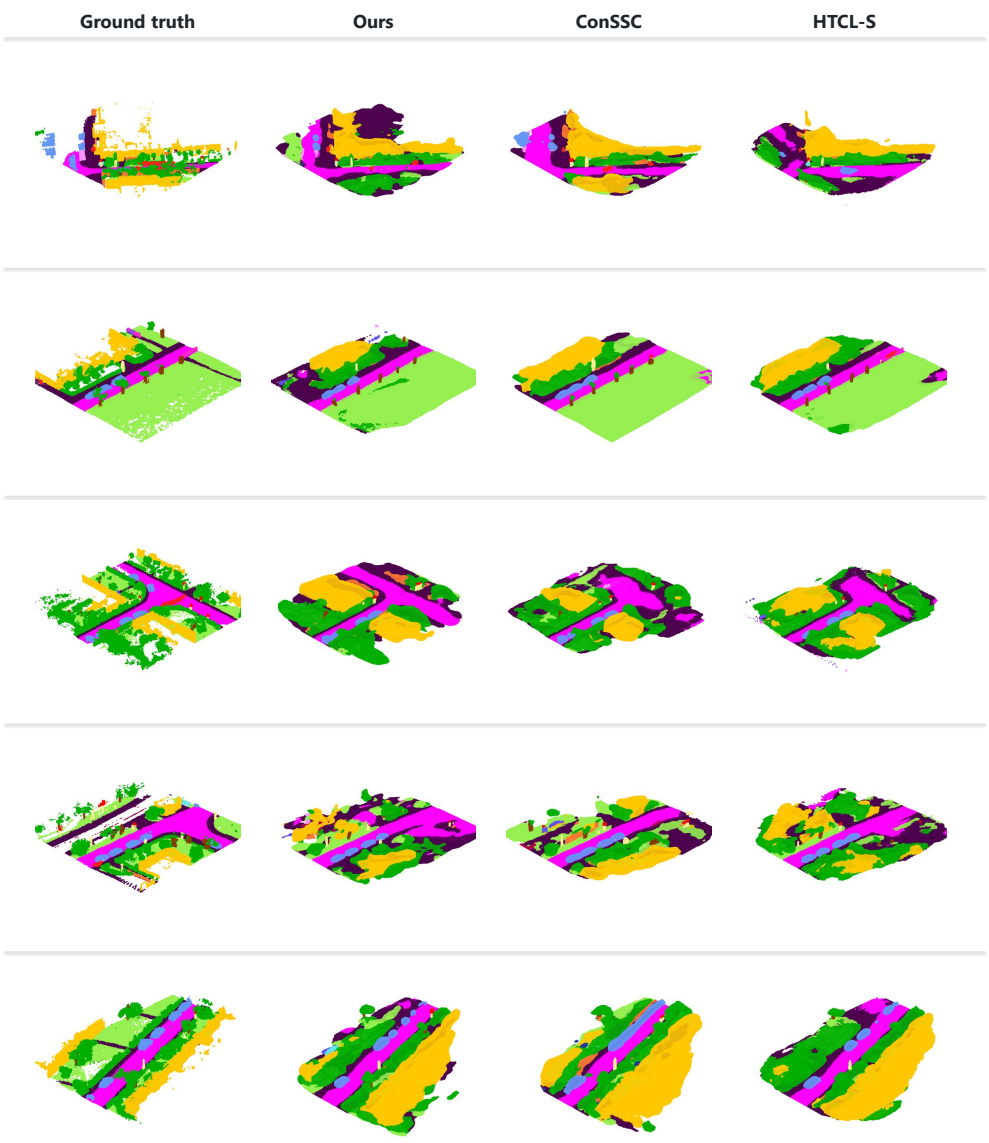

Figure 6: Further voxel-volume comparisons on the SemanticKITTI validation set. The column order and viewpoint matching follow Figure 5.

Voxel rendering alone omits camera context. Figure 7 therefore pairs the source RGB image with aligned 3D predictions for two additional scenes. It illustrates input–output correspondence, not the correctness of any particular scene-graph edge.

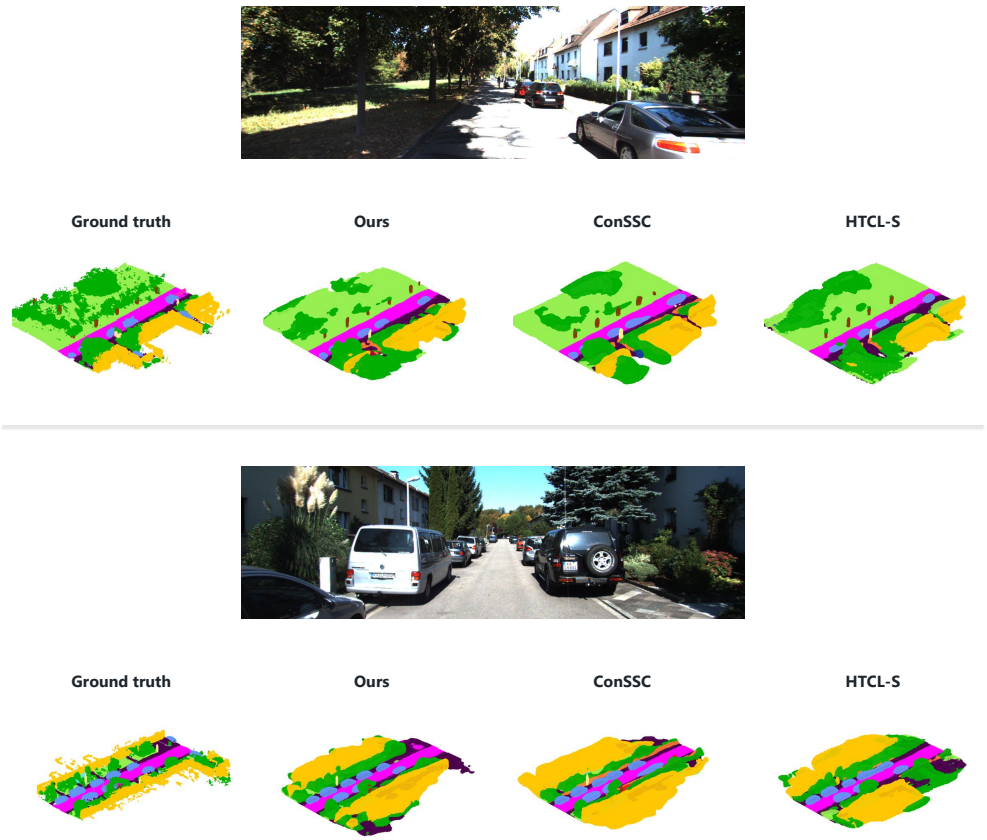

Figure 7: RGB inputs and corresponding voxel predictions on SemanticKITTI validation scenes. Each scene pairs the source image with ground truth and predictions from the compared methods.

Figure 8 broadens the sample to different road, building, and vehicle layouts. Boundary and distant-region errors vary among methods; the figure is intended as a qualitative supplement to the paired metrics.

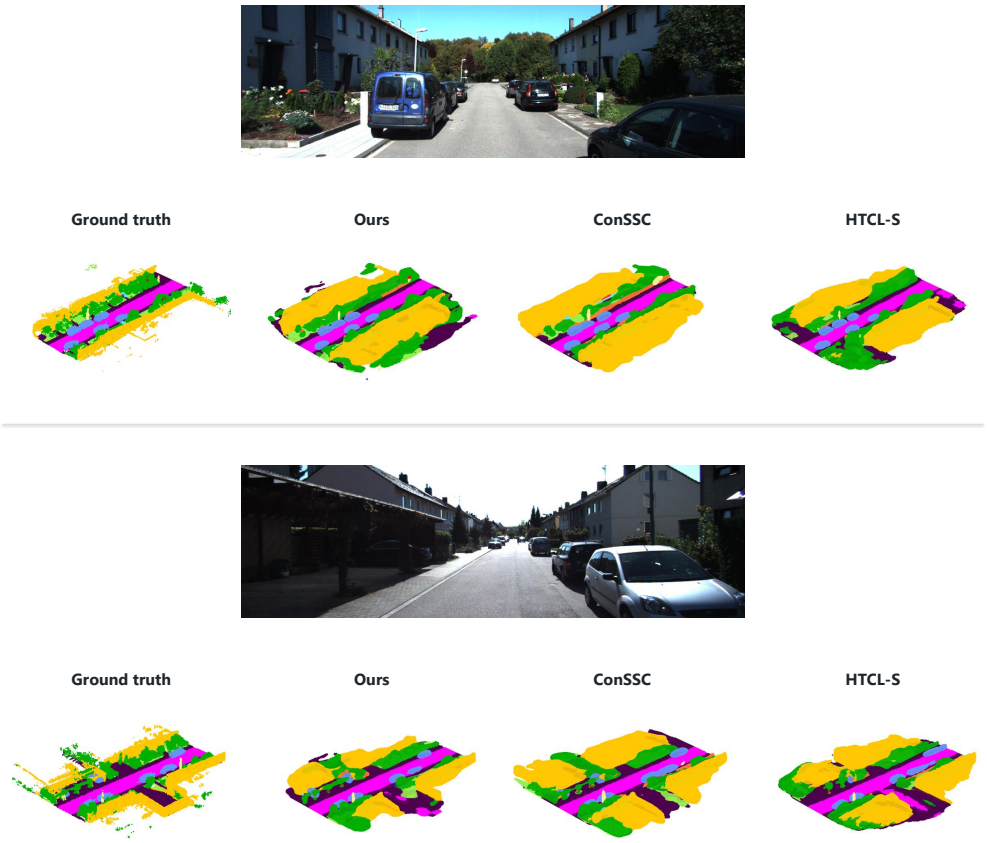

Figure 8: Additional RGB-to-voxel comparisons across different road, building, and vehicle layouts. The method ordering matches Figure 7.

Figure 9 compares two further frames with related street layouts but different vehicles. Its qualitative differences should not be turned into an unmeasured per-voxel or occlusion-specific score.

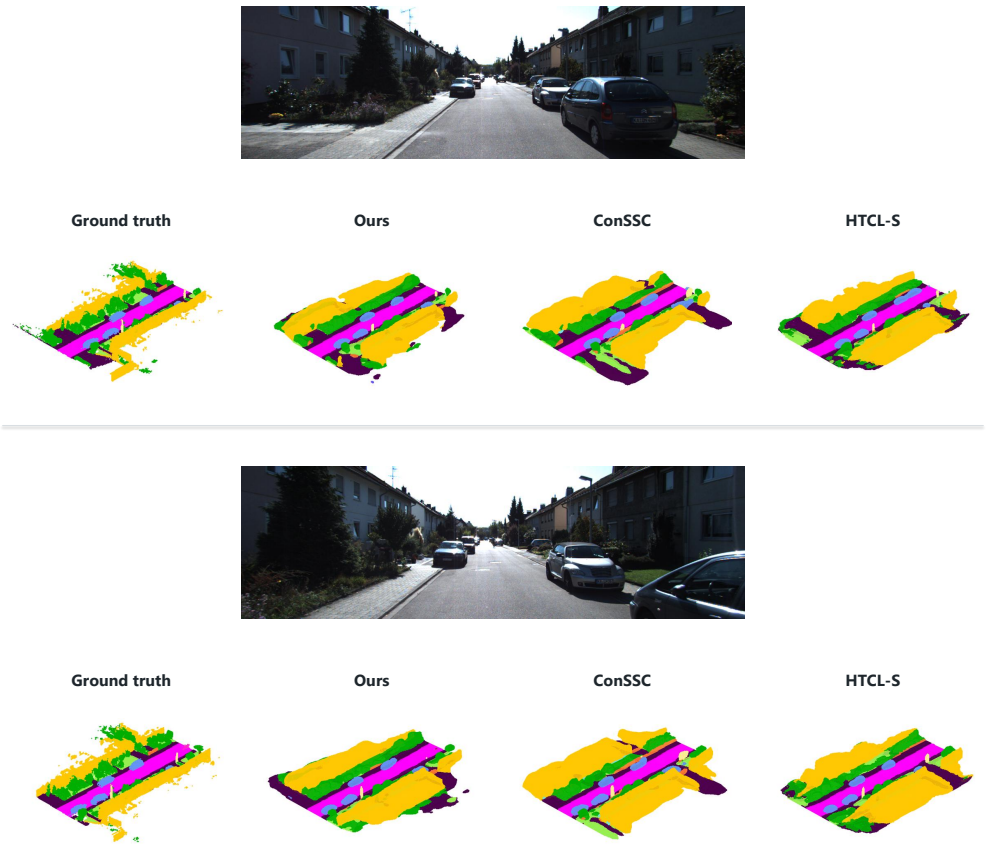

Figure 9: Further RGB-to-voxel comparisons on SemanticKITTI validation scenes with related street layouts and different vehicles.

Figure 10–15 illustrate how dual memory complements semantic scene completion. Episodic memory injects historical features into regions observed in previous frames but unseen in the current frame (cyan), while prototype memory provides class-level priors for regions unobserved in both current and historical frames (orange). Gray denotes locations without direct memory injection. The aligned views relate these injection locations to the predicted scene structure.

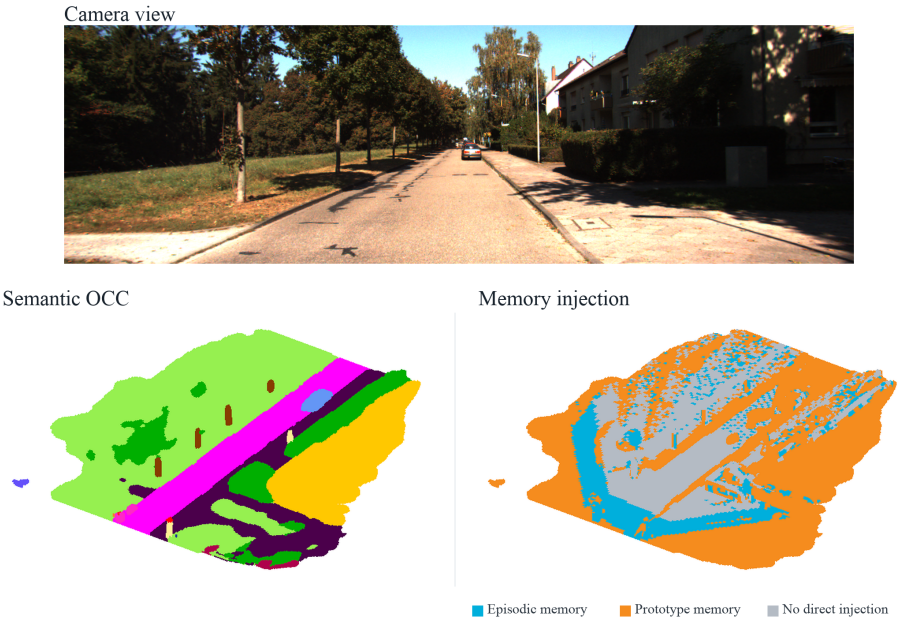

Figure 10: SemanticKITTI visualization: input image (top), semantic occupancy (bottom left), and memory injection (bottom right): episodic (cyan), prototype (orange), and none (gray).

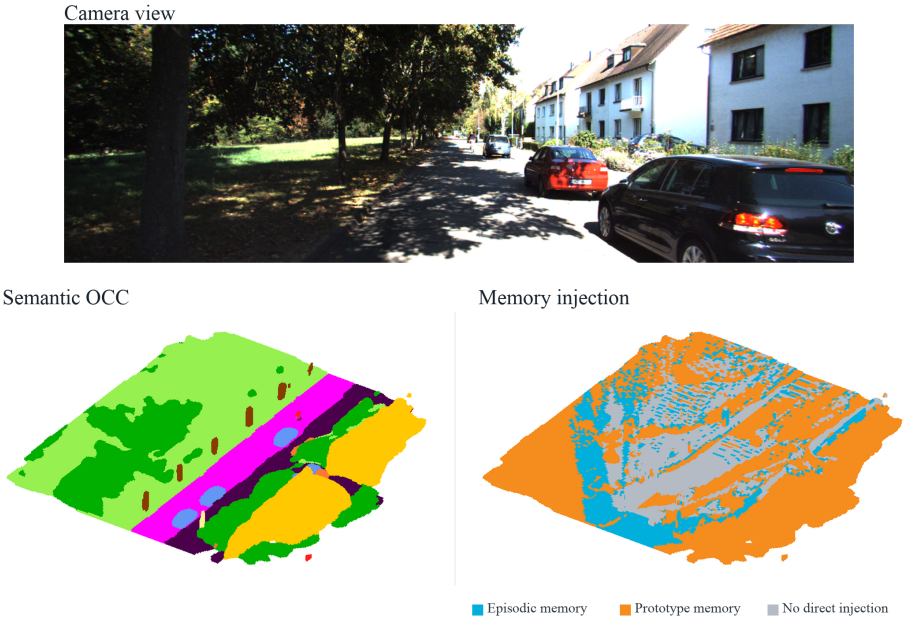

Figure 11: SemanticKITTI visualization: input image (top), semantic occupancy (bottom left), and memory injection (bottom right): episodic (cyan), prototype (orange), and none (gray).

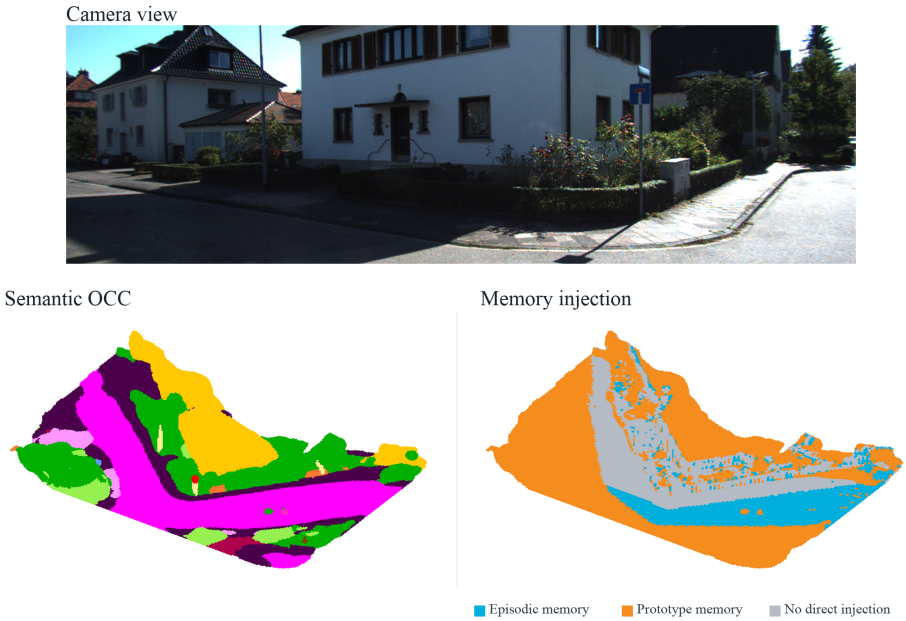

Figure 12: SemanticKITTI visualization: input image (top), semantic occupancy (bottom left), and memory injection (bottom right): episodic (cyan), prototype (orange), and none (gray).

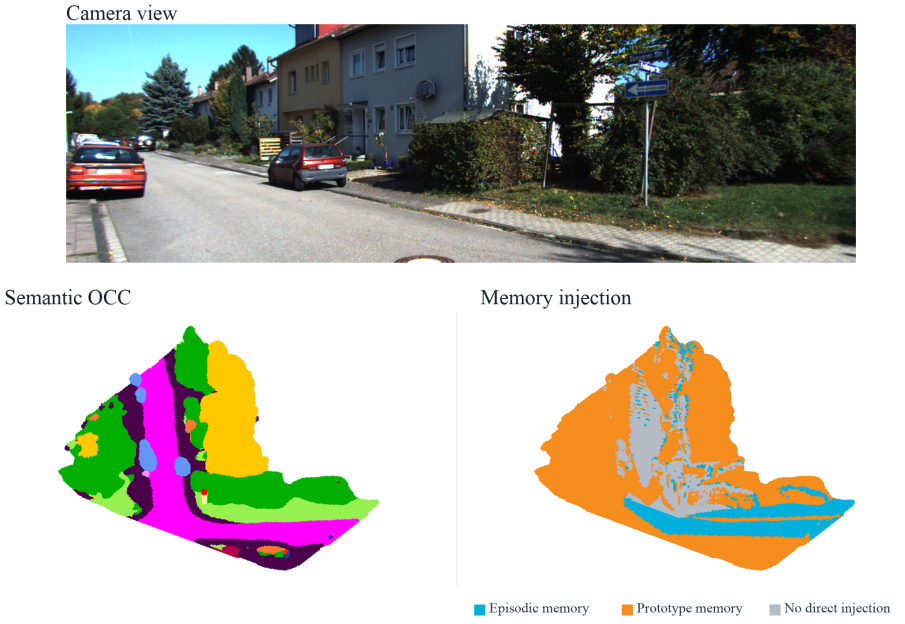

Figure 13: SemanticKITTI visualization: input image (top), semantic occupancy (bottom left), and memory injection (bottom right): episodic (cyan), prototype (orange), and none (gray).

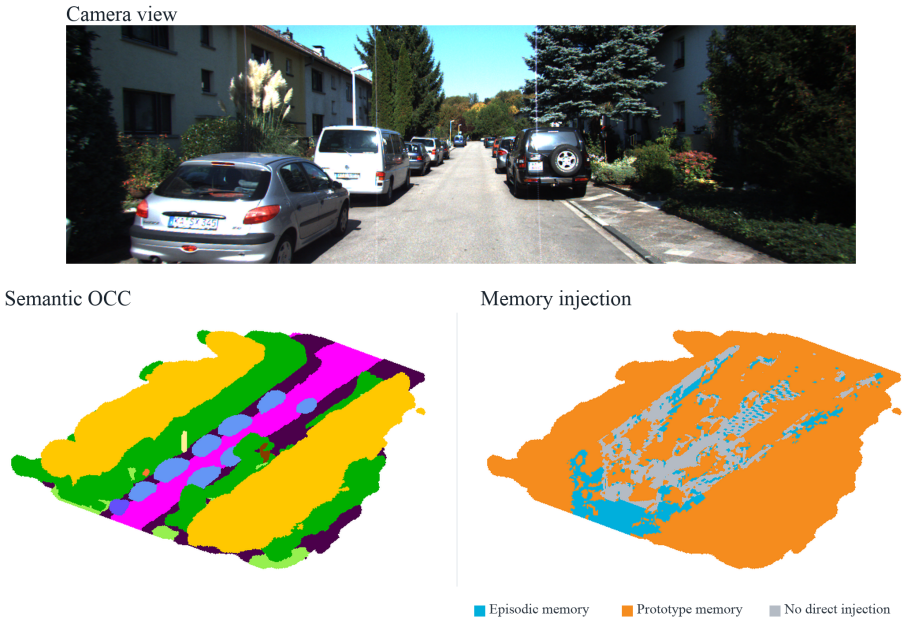

Figure 14: SemanticKITTI visualization: input image (top), semantic occupancy (bottom left), and memory injection (bottom right): episodic (cyan), prototype (orange), and none (gray).



Figure 15: SemanticKITTI visualization: input image (top), semantic occupancy (bottom left), and memory injection (bottom right): episodic (cyan), prototype (orange), and none (gray).

F LIMITATIONS AND FUTURE WORK

External graph-generation cost. New scenes require detection, VLM relation prediction, and depth verification before the SSC forward pass. Caching avoids repeated training-time work but does not make deployment preprocessing free. Distillation or lightweight relation proposal, with separate and end-to-end latency reporting, is a natural future direction.

Approximate object support. Two-dimensional boxes, median depth, and discrete projection do not recover exact 3D instance boundaries. Depth checking rejects clear order inversions but can miss mixed-background boxes and confusions between adjacent same-class instances. Instance masks or learned supports could be compared under the same training protocol.

Class imbalance. Reliable relations concentrate on cars. The completed modulation-strength comparison reveals a semantic/occupancy trade-off but cannot isolate its cause. Future class- and occlusion-stratified evaluation should precede any class-aware calibration.

One-way memory–graph connection. The graph reads memory-enhanced features but does not control memory retrieval or write relation results back into memory. Relation-conditioned retrieval is possible future work, subject to equal-compute ablation.

Occlusion-specific evaluation. Global IoU and mIoU include large visible/background regions. A fixed, prediction-independent occlusion mask and object-level grouping would make the hypothesized local effect testable. Until then, global gains alone should not be described as quantitative proof of improvement specifically at occluded voxels.