BIOCC: BIDIRECTIONAL CONTEXT MODELING FOR SEMANTIC OCCUPANCY COMPRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

3D semantic occupancy has emerged as a unified representation of 3D scenes, with broad applications in scene generation, autonomous navigation, and collaborative perception. Nevertheless, transmitting dense 3D occupancy representations requires substantial bandwidth overhead, motivating efficient compression of both geometry and semantic labels. A natural approach is to treat semantic labels as point attributes and adapt existing point cloud codecs accordingly. Unlike conventional appearance attributes, semantic labels are spatially coherent and closely coupled with geometry, making them potentially informative for geometry coding. However, leveraging semantic information as context for geometry coding remains largely underexplored. To address this, we propose BiOcc, a bidirectional semantic occupancy compression framework that exploits early-decoded semantics for subsequent geometry coding and reconstructed geometry for finer-scale semantic coding. For semantics to geometry, label relative spatial context captures the arrangement of the same label neighbors, while the semantic category context encodes the current node's label. The semantic representations are combined with the geometry context to improve the prediction of occupancy patterns. For geometry to semantics, the local geometry context is combined with the available semantic context to predict label probabilities on occupied voxels. We further analyze the potential coding gain from semantic conditioning and show how semantic coarsening can reduce the geometry relevant information available for occupancy prediction. Extensive experiments on the outdoor nuScenes-Occ and indoor ScanNet-Occ datasets demonstrate that BiOcc achieves state-of-the-art performance in geometry and semantic compression.

1 INTRODUCTION

3D semantic occupancy has become a scene-centric representation for 3D scene understanding, providing a unified description of geometry and semantics over the surrounding space (Wang et al., 2023; Zheng et al., 2024; Li et al., 2025; Zhu et al., 2026a; Yang et al., 2026). Unlike point clouds, which primarily represent observed surfaces as a sparse set of 3D points, semantic occupancy models the surrounding 3D space on a structured grid, enabling a unified representation of occupied space, and semantic categories (Wei et al., 2023; Tian et al., 2023; Shi et al., 2026). However, the fine-grained volumetric representation of semantic occupancy introduces considerable storage and transmission overhead (Li et al., 2026; Guo et al., 2026). By discretizing the surrounding 3D space into dense spatial elements, its data volume grows rapidly with increasing spatial resolution, perception range, and temporal duration. As semantic occupancy is increasingly adopted for scene modeling, mapping, and collaborative perception, efficient storage and transmission become essential for practical deployment (Chen et al., 2026). Therefore, efficiently compressing semantic occupancy is important for preserving detailed scene representations under limited storage and bandwidth.

However, dedicated compression methods for 3D semantic occupancy remain largely unexplored. By treating semantic labels as categorical point attributes, semantic occupancy can also be formulated as a joint geometry-attribute compression problem, making point cloud compression methods a natural choice. Point cloud compression pipelines typically recover the geometry first, then decode attributes on the reconstructed spatial base. While this geometry-first design provides a natural causal decoding order, attribute information is rarely exploited as context for subsequent geometry coding. This limitation is particularly relevant to semantic occupancy, where semantic labels differ

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.

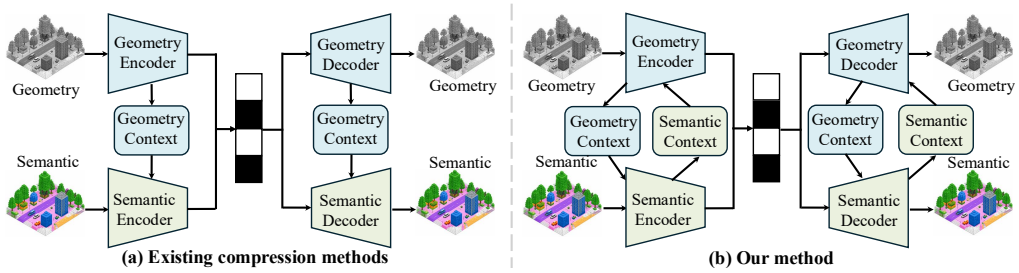

Figure 1: Comparison of existing joint compression methods and our bidirectional compression framework. Existing methods mainly follow a geometry-to-semantic decoding pipeline, whereas our method enables bidirectional context interaction: semantic context assists geometry decoding, and reconstructed geometry context facilitates semantic decoding.

fundamentally from appearance attributes. Unlike appearance attributes such as RGB or reflection, whose statistics are sensitive to local texture, illumination, and sensor noise, semantic labels encode more abstract structural information, remain consistent within regions and objects, and are less affected by these appearance variations (Schönberger et al., 2018; Stenborg et al., 2018; Xue et al., 2023). Such semantic regularities are closely correlated with the underlying occupancy structure: voxels belonging to the same object or semantic region often exhibit coherent spatial layouts, while semantic boundaries tend to align with geometric boundaries. Therefore, semantics can serve as an informative context for geometry coding, providing complementary information beyond what geometry alone can capture, thereby improving occupancy prediction and compression efficiency.

Semantic information has been explored in point cloud geometry compression from several perspectives. One line of work exploits semantics to construct compact, semantically aware geometry representations, selectively represent different scene regions, and guide content-aware visual decomposition for point cloud compression (Zhao et al., 2022; Xie et al., 2024). More recently, R^{2}D-LPCC (Nan et al., 2026) combines semantic information with predictive uncertainty to estimate region relevance, which subsequently guides adaptive bit allocation and region-aware feature transformation. Another line of work incorporates semantic labels into geometry representation learning by conditioning features or latent representations on semantic labels and scene-level structures (Stathoulopoulos et al., 2025). Recent studies have also explored joint geometry-attribute modeling to capture cross-modal dependencies. Deep-JGAC (Zhang et al., 2026) exploits correlations between attributes and geometry, while SEDD-PCC (Hsieh et al., 2025) and MEGA-PCC (Hsieh et al., 2026) learn shared latent representations for joint reconstruction. Despite this progress, incorporating semantic information as contextual input into the geometry codec to explicitly exploit the dependency between geometry and semantics remains largely unexplored.

To bridge this gap, we propose BiOcc, a bidirectional semantic occupancy compression framework that exploits semantics for geometry coding and geometry for semantic coding, as shown in Fig. 1. BiOcc adopts a coarse-to-fine interleaved schedule to accommodate these dependencies. After an initial geometry-only stage, decoded semantics guide geometry coding at the next scale, while the reconstructed geometry supports subsequent semantic coding. At the core of this framework, a semantic conditional geometry model combines label relative spatial context and semantic category information to refine occupancy probabilities. The former captures the arrangement of same-label neighbors, while the latter identifies the category associated with that arrangement. Together, they provide complementary conditioning information beyond the geometry context, reducing uncertainty in occupancy prediction. We further provide an information-theoretic analysis that characterizes the potential geometry coding gain from semantic context via conditional mutual information. The analysis shows that the gain depends on how much geometry-relevant information is preserved in the semantic context. Evaluated on nuScenes-Occ and ScanNet-Occ, BiOcc establishes new state-of-the-art results in geometry and semantic compression across outdoor and indoor scenes. Code and data will be available for research purposes. Our main contributions are as follows:

- We introduce BiOcc, a bidirectional compression framework for semantic occupancy. Its interleaved coding schedule enables decoded semantics to assist geometry coding and decoded geometry to assist semantic coding, while maintaining a causal decoding process.

- We develop a semantic conditional geometry model that combines label relative spatial context and semantic category context. These representations jointly capture local semantic organization and category identity, reducing uncertainty in geometry occupancy prediction.
- We present an information-theoretic analysis that relates the geometry coding gain from semantic context to conditional mutual information. It further shows that this gain depends on semantic information relevant to geometry and weakens as semantic reliability degrades.

2 Related Work

2.1 Semantic-Aware Point Cloud Compression

Existing semantic-aware point cloud compression methods exploit semantics to improve geometry coding in various ways. Semantic priors have been used to remove task-irrelevant geometry (Zhao et al., 2022), while semantic classes have been exploited to partition point clouds into independently coded streams for content-aware compression and transmission (Mari et al., 2023). Other approaches use semantic classes or regions to guide class-wise coding, importance-aware bit allocation, and region-adaptive processing (Xie et al., 2024; Nan et al., 2026). Object-aware priors have also been incorporated into occupancy probability estimation (Luo et al., 2025), and semantic labels or scene structures have been used to condition learned geometry representations (Stathoulopoulos et al., 2025). Recent work has introduced textual semantic information to guide the reconstruction of generative geometry (Zheng & Gao, 2026). However, existing methods primarily rely on semantics to guide geometry preprocessing, representation, or rate allocation, rather than explicitly encoding and decoding semantic labels as context for subsequent geometry coding.

2.2 Joint Point Cloud Compression

Joint point cloud compression aims to exploit the dependencies between geometry and attributes to improve the coding efficiency of both modalities. DeepPCC (Zhang et al., 2024) adopts a sequential pipeline, first reconstructing lossy geometry and then encoding attributes on the reconstructed spatial support, with multiscale neighborhood aggregation tailored to both modalities. Deep-JGAC (Zhang et al., 2026) adopts a two-stream design in which geometry is reconstructed first and subsequently used to support re-colorization and attribute coding, with the geometry and attribute codecs jointly optimized under a unified rate-distortion objective. SEDD-PCC (Hsieh et al., 2025) replaces separate encoders with a shared encoder that maps geometry and attributes into a unified latent representation, followed by specialized decoders that sequentially reconstruct geometry and attributes. MEGA PCC (Hsieh et al., 2026) extends this unified design with Mamba-based representation and entropy modeling, enabling the joint end-to-end optimization of dependencies between geometry and attributes, as well as bitrate allocation. More recently, variable rate joint compression (Rudolph et al., 2026) embeds geometry and attributes into a common latent space and conditions the model on modality-specific quality targets, enabling flexible tradeoffs among geometry quality, attribute quality, and bitrate within a single model. Despite these advances, existing joint compression methods primarily couple geometry and attributes through joint rate-distortion optimization and bitrate allocation, while attribute information is rarely exploited as context for geometry coding.

3 Method

3.1 Overview

Let a semantic occupancy map be $X = (G, S)$, where $G \subseteq \Omega$ denotes the set of occupied voxels in a known integer grid Ω, and $S : G \to \{0, \dots, K-1\}$ assigns a semantic label to each occupied voxel. We construct a coarse-to-fine hierarchy $\{X^v = (G^v, S^v)\}_{v=0}^{L}$, where $X^L = X$. Each occupied voxel in G^v corresponds to a spatial block containing at least one occupied voxel at the finest level. Its semantic label is obtained by majority voting over the occupied finest-level voxels within that block, with ties resolved by the smallest class index. Thus, $S^v : G^v \to \{0, \dots, K-1\}$ defines the semantic representation at level v.

BiOcc models the joint geometry-semantic representation using a causal coarse-to-fine factorization. Let l denote the first hierarchy level at which semantic labels are transmitted, and let

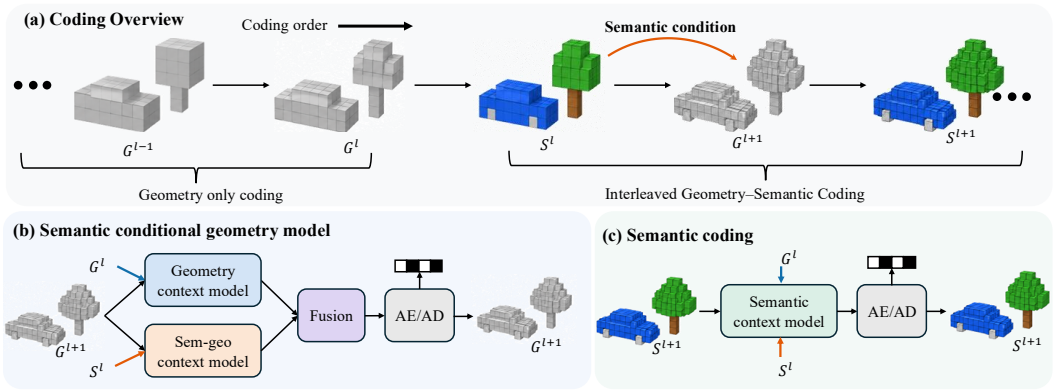

Figure 2: Overview of the proposed BiOcc framework for semantic occupancy compression. (a) Geometry is first coded without semantic conditioning, then interleaved with semantic coding across successive levels. Previously decoded semantics S^l provide additional context for coding geometry G^{l+1}. (b) For geometry coding, the geometry and sem-geo context models provide complementary features, which are fused to predict occupancy probabilities for arithmetic coding. (c) For semantic coding, the semantic context model predicts label probabilities using available geometry and previously decoded semantics. AE/AD denote arithmetic encoding and decoding, respectively.

$G^{\le v} = \{G^0, \dots, G^v\}$. The joint distribution can be factorized as

$$p(G^{1:L}, S^{l:L} \mid G^0) = \prod_{v=0}^{l-1} p_G(G^{v+1} \mid G^{\le v}) \cdot p_S(S^l \mid G^{\le l}) \cdot \prod_{v=l}^{L-1} p_G(G^{v+1} \mid G^{\le v}, S^v) p_S(S^{v+1} \mid G^{\le v+1}, S^v). \tag{1}$$

The first term corresponds to geometry-only coding at coarse levels. Once S^l has been decoded, semantic information becomes available as causal context for the subsequent geometry stage, while the reconstructed finer geometry is in turn used to predict the next semantic representation. This factorization enables semantic-to-geometry and geometry-to-semantic dependencies.

As illustrated in Fig. 2 (a), BiOcc first codes the coarse geometry levels without semantic conditioning. At these levels, semantic aggregation over large spatial regions may obscure local semantic structure, limiting the additional information provided by coarse labels. Therefore, we introduce semantic coding only from level l, after which decoded S^l is available for coding G^{l+1}, and reconstructed G^{l+1} subsequently supports the prediction of S^{l+1}. In our default configuration, we set $l = L - 1$ motivated by the information-theoretic analysis in Sec. 3.2 and result in Sec. 4.3. This schedule makes decoded semantics available for subsequent geometry coding. For semantic conditional geometry coding (Fig. 2 (b)), a geometry context model extracts features from the decoded geometric context, while a sem-geo context model extracts complementary features. These features are fused to predict the probabilities of child occupancy for arithmetic coding. For semantic coding (Fig. 2 (c)), a semantic context model uses the geometry and previously decoded semantics to predict label probabilities on occupied voxels for arithmetic coding.

3.2 Semantic Context for Geometry Coding

To understand when semantic context can benefit geometry coding, we consider how the occupancy distribution varies across semantic states. Let M denote a child occupancy at a given scale, C_G the available geometric context, and C_S the semantic context derived from the previously decoded semantic representation. For a fixed geometric context $C_G = c_g$, the potential reduction in the ideal geometry coding rate ΔR_G is

$$\begin{aligned} \Delta R_G(c_g) &= H(M \mid C_G = c_g) - H(M \mid C_G = c_g, C_S) = I(M; C_S \mid C_G = c_g) \\ &= \mathbb{E}_{C_S \mid C_G = c_g} \left[D_{\mathrm{KL}} \left(p(M \mid C_G = c_g, C_S) \,\|\, p(M \mid C_G = c_g) \right) \right] \ge 0, \end{aligned} \tag{2}$$

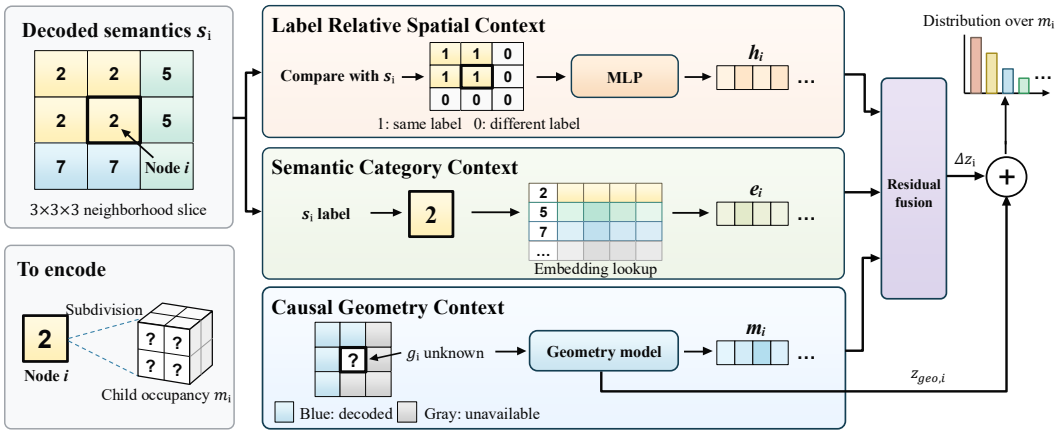

Figure 3: Semantic conditional geometry context model. For the current node i, the model predicts its child occupancy code m_i using decoded semantics and causal geometry context. Label relative spatial context converts neighboring semantic labels into a binary same-label pattern and encodes it into h_i, capturing local semantic structure. Semantic category context maps the node's known label s_i to a category-specific representation e_i. The causal geometry context produces geometry features g_i and baseline occupancy logits $z_{\mathrm{geo},i}$. Residual fusion combines these complementary representations to predict a logit correction Δz_i. The corrected logits are normalized by softmax to obtain a distribution over 255 non-empty child occupancy patterns. Black outlines indicate the current node; blue and gray cells denote decoded and unavailable geometry context, respectively.

where $H(\cdot \mid \cdot)$ denotes conditional Shannon entropy, $I(\cdot;\cdot \mid \cdot)$ denotes conditional mutual information, and $D_{\mathrm{KL}}(\cdot\|\cdot)$ denotes the Kullback-Leibler divergence. The inequality follows from the non-negativity of the KL divergence. The gain is positive when different semantic contexts yield distinct occupancy distributions for the same geometric context. If semantic context does not affect the occupancy distribution, the gain is zero.

Intuitively, the geometric context C_G available to the entropy model provides only a partial description of the underlying structure. Hence, similar coarse contexts may still correspond to different fine-scale occupancy patterns. For example, bicycle and motorcycle regions can exhibit similar coarse occupancy configurations while differing in their finer structures. Their semantic labels can therefore help resolve this ambiguity and make the child occupancy M more predictable.

Note that this gain applies only to the geometry term. Under ideal entropy coding, changing the causal factorization does not alter the joint entropy, since $H(M, S \mid C) = H(M \mid C) + H(S \mid M, C) = H(S \mid C) + H(M \mid S, C)$. In practice, learned codecs with finite-capacity entropy models and restricted causal contexts may approximate different factorizations with varying accuracy; BiOcc exposes semantic structure to reduce the modeling difficulty of the geometric distribution.

The hierarchy further explains why semantic conditioning is more useful at finer scales. Let C_S^{v+1} denote a fine-scale semantic context and $C_S^v = A(C_S^{v+1})$ its coarsened representation obtained by semantic aggregation. Since C_S^v is a deterministic function of C_S^{v+1}, the conditional data-processing inequality gives

$$I(M; C_S^v \mid C_G) \le I(M; C_S^{v+1} \mid C_G). \tag{3}$$

The inequality becomes strict when semantic aggregation discards fine-scale information that remains useful for predicting M. Thus, semantic aggregation cannot increase the geometry-relevant information available for occupancy prediction. This motivates the introduction of semantic conditioning at finer levels of hierarchy. The proof is provided in Appendix A.

3.3 Semantic Conditional Geometry Model

Although geometry context models already capture strong occupancy dependencies, semantics provide complementary information about the current node's position within a semantic region and the region's category. Same-label neighbors reveal whether the current node lies inside a semantic re-

gion or close to its boundary, while the specific label provides category-dependent information for interpreting the corresponding local geometry. We capture these complementary signals through label relative spatial context and semantic category context to refine the geometry-based occupancy distribution. As illustrated in Fig. 3, these representations are combined with causal geometry features to refine the occupancy distribution through residual logit prediction.

For a current occupied node i, let $m_i \in \{1, \dots, 255\}$ denote the joint occupancy code of its eight children. Let $\mathcal{G}_{<i}$ denote the geometry context available under the coding order, and let $\mathcal{S}_i = \{s_j \mid j \in \mathcal{N}(i)\}$ denote decoded semantic labels in a $3 \times 3 \times 3$ neighborhood centered at i. The current node's semantic label s_i is known before coding m_i. Our objective is to model $p_\theta(m_i \mid \mathcal{G}_{<i}, \mathcal{S}_i)$, where θ denotes the learnable model parameters. Only semantic labels available to both the encoder and decoder are used, while occupancy context remains causal.

Label relative spatial context. We represent the local semantic context through the relationship between each neighboring label and the current node's label, while preserving their relative spatial configuration. For each neighborhood position j, we construct $r_{ij} = \mathbf{1}[s_j = s_i \ \wedge \ s_j \neq \mathrm{unk}]$, where $\mathbf{1}[\cdot]$ denotes the indicator function, $\wedge$ denotes logical conjunction, and unk denotes an unavailable semantic label. The indicators are arranged in a fixed spatial order and encoded as $h_i = f_{\mathrm{rel}}\big(\mathrm{vec}\big(\{r_{ij}\}_{j \in \mathcal{N}(i)}\big)\big)$, where $\mathrm{vec}(\cdot)$ vectorizes the relation indicators according to a fixed spatial order, f_{rel} is a multilayer perceptron (MLP), and h_i denotes the resulting label-relative spatial feature. Preserving the spatial ordering allows h_i to represent the arrangement of same-label neighbors around the current node. This representation is invariant to a consistent permutation of semantic category identifiers, enabling the same local relation pattern to be shared across categories.

Semantic category context. Since the label relative spatial context captures only relative label agreement, we also encode the current node's semantic category using a learned embedding, $e_i = E(s_i)$. where E is a learnable embedding table, and e_i denotes the semantic category feature of node i. Whereas h_i describes the spatial arrangement of label agreement, e_i identifies the semantic category associated with that arrangement. Together, they enable the predictor to learn both shared spatial dependencies and category-dependent occupancy statistics.

Residual geometry prediction. The causal geometry encoder extracts a feature vector g_i from $\mathcal{G}_{<i}$ and produces base occupancy logits $z_{\mathrm{geo},i}$. A learnable fusion module combines the geometry and semantic representations to predict a residual correction:

$$\Delta z_i = f_{\mathrm{fuse}}(g_i, h_i, e_i), \qquad z_i = z_{\mathrm{geo},i} + \Delta z_i, \tag{4}$$

where Δz_i denotes the predicted residual correction to the base occupancy logits, f_{fuse} denotes a gated fusion network that maps the geometry and semantic representations to a 255-dimensional residual logit correction, and z_i denotes the final occupancy logits. The resulting probabilities are

$$p_\theta(m_i = k \mid \mathcal{G}_{<i}, \mathcal{S}_i) = \frac{\exp(z_{i,k})}{\sum_{q=1}^{255} \exp(z_{i,q})}, \tag{5}$$

where $z_{i,k}$ is the logit associated with occupancy code $k \in \{1, \dots, 255\}$. This residual formulation preserves the geometry prediction as a reference and allows the semantic context to adjust the relative likelihoods of candidate occupancy patterns. The resulting distribution is used to code m_i.

3.4 Geometry and Semantic Context Model

The geometry context model predicts each node's child occupancy code from the available geometry context. To avoid costly node-by-node autoregressive coding, we partition nodes into ordered groups and perform causal modeling at the group level. Nodes within the same group are processed in parallel, while their predictions depend only on the geometry available from previously decoded groups. Local occupancy embeddings and ancestor-path information are processed by learned projections and a causal Transformer backbone. The resulting representations are fused by a prediction head to produce logits over the 255 non-empty occupancy patterns. A softmax converts these logits into categorical distributions, which are used to construct the cumulative distributions.

The semantic context model follows the same group-wise coding strategy to estimate the label distribution for each node based on its local semantic context. Nodes within each group are processed

in parallel using semantic information available from previously decoded groups. For context nodes whose fine-level labels are not yet available, we substitute the corresponding known coarse-level semantic labels. The resulting labels are mapped to learned embeddings, combined with positional embeddings, and processed by an MLP to extract semantic context features. A separate MLP encodes the available local geometry. These features, together with coarse-level semantic and coordinate embeddings, are combined by MLP prediction heads to produce semantic label logits. A softmax converts the logits into categorical distributions for arithmetic coding into the semantic bitstream.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Dataset. We evaluate our method on nuScenes-Occ (Tian et al., 2023) and ScanNet-Occ (Yu et al., 2024), covering outdoor driving scenes and indoor environments, respectively. nuScenes-Occ provides voxel-wise occupancy annotations on a $200 \times 200 \times 16$ grid with a voxel size of 0.4 m, distinguishing 17 semantic categories. ScanNet-Occ is constructed from ScanNet and represents indoor scenes using $60 \times 60 \times 36$ voxel grids with 11 semantic categories. Together, these datasets cover different spatial scales, scene structures, and semantic categories, allowing us to assess the rate-distortion performance of semantic occupancy compression in both outdoor and indoor settings.

Baselines. For semantic occupancy compression, we compare BiOcc with representative joint geometry-attribute codecs, including MPEG G-PCC (Schwarz et al., 2018), Unicorn (Wang et al., 2025a), and UPCC (Rudolph et al., 2026). These baselines represent different strategies for exploiting dependencies between geometry and attributes, ranging from geometry-first coding to joint rate-distortion optimization. For geometry compression, we further include specialized learned codecs, TopNet (Wang et al., 2025b) and AnyPCC (Wang et al., 2026), to provide a direct comparison of geometry coding efficiency. To enable evaluation on semantic occupancy, we adapt each baseline to voxelized occupancy inputs and categorical semantic labels while preserving its original coding architecture and core entropy-modeling mechanism. Please refer to Appendix C.1 for more details.

Metrics. Lossy semantic occupancy compression introduces distortions in both geometry and semantics. Following prior work on semantic occupancy (Wang et al., 2023; Pan et al., 2024; Xu & Xu, 2025), we evaluate geometric fidelity using occupancy Intersection-over-Union (IoU) and semantic fidelity using mean Intersection-over-Union (mIoU). The coding rate is measured in bits per point (bpp). For each quality metric, we further report the Bjøntegaard Delta Bit Rate (BD-Rate) to quantify the average bitrate difference, where a negative BD-Rate indicates bitrate savings.

Implementation details. For each dataset, we train two variants: BiOcc and BiOcc-G. BiOcc includes all proposed components, while BiOcc-G excludes the semantic conditional geometry model and serves as the baseline for the ablation studies. Since semantic occupancy is defined on a fixed voxel grid, geometry errors also introduce semantic errors; we therefore control semantic fidelity through geometry reconstruction. We vary λ_G to control the fraction of explicitly transmitted child-occupancy patterns. The remaining occupancy is reconstructed by constrained Top-K selection using a transmitted occupied-child count K, retaining at least one child per parent. For reconstructed occupied voxels without corresponding semantic labels in the original occupancy map, we assign the label of the nearest labeled voxel. Please refer to Appendix C.2 for more details.

4.2 PERFORMANCE EVALUATION

Geometry coding. As shown in Fig. 4 and Table 1, BiOcc consistently achieves the lowest geometry-stream BD-Rate on both datasets. It achieves BD-Rates of -55.64% on nuScenes-Occ and -93.01% on ScanNet-Occ, corresponding to additional reductions of 15.68 and 14.54 percentage points over the best competing method on each dataset. These gains are observed against both conventional geometry-attribute codecs and specialized learned geometry codecs, demonstrating the effectiveness of semantic conditioning in reducing the geometry-stream bitrate across diverse scenes.

Joint compression. When jointly accounting for geometry and semantic bitstreams, BiOcc also achieves the best rate–distortion performance, with BD-Rates of -70.32% and -87.61% on nuScenes-

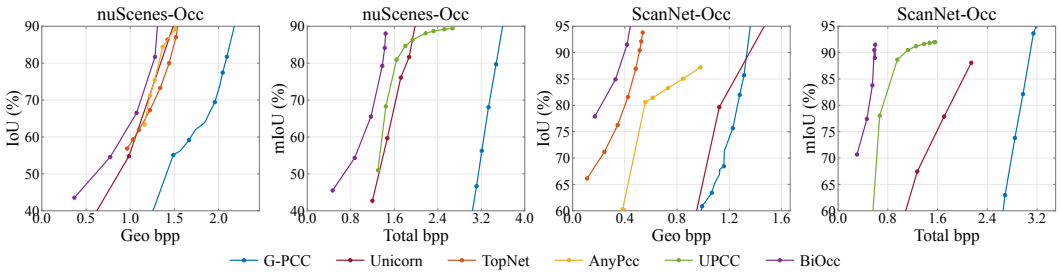

Figure 4: Comparison of rate-distortion (R-D) performance across nuScenes-Occ and ScanNet-Occ.

Table 1: Comparison of BD-Rate relative to G-PCC (%) and encoding/decoding runtimes (in seconds). For **Geometry**, BD-Rate is computed from geometry bitrate and IoU. For **Joint**, BD-Rate is computed from the total geometry-and-semantic bitrate and mIoU, and the runtime includes both geometry and semantic coding. The best results are marked in **bold**.

Method	nuScenes-Occ			ScanNet-Occ		
	BD-Rate↓	**Encoding↓**	**Decoding↓**	**BD-Rate↓**	**Encoding↓**	**Decoding↓**
Geometry Compression						
G-PCC	0	0.177	0.142	0	**0.028**	0.024
Unicorn	-31.08	0.604	0.552	-9.43	0.253	0.249
TopNet	-39.75	0.538	90.36	-78.47	0.108	5.420
AnyPCC	-39.96	**0.114**	**0.054**	-62.04	0.055	**0.021**
BiOcc	**-55.64**	0.336	0.644	**-93.01**	0.056	0.114
Joint Compression						
G-PCC	0	0.236	0.177	0	0.032	0.024
Unicorn	-52.74	1.160	1.077	-43.42	0.510	0.501
UPCC	-57.43	**0.050**	**0.059**	-69.28	**0.031**	**0.032**
BiOcc	**-70.32**	0.539	0.837	**-87.61**	0.146	0.209

Occ and ScanNet-Occ, respectively. Compared with UPCC, BiOcc further reduces BD-Rate by 12.89 and 18.33 percentage points on the two datasets. These results demonstrate the benefit of bidirectional context modeling, where decoded semantics assist geometry coding and reconstructed geometry in turn facilitates semantic coding.

Runtime. Table 1 also reports encoding and decoding runtimes. For geometry coding, BiOcc substantially reduces the decoding time compared with TopNet while achieving a lower geometry-stream BD-Rate. For joint compression, BiOcc is consistently faster than Unicorn on both datasets, whereas UPCC achieves a lower runtime. Overall, BiOcc provides substantial rate savings with a moderate increase in computational cost relative to the fastest baseline.

4.3 Ablation Studies

Contributions of sem-geo context. We evaluate the two semantic context branches by adding each to BiOcc-G individually. As shown in Table 2, the label relative spatial context reduces BD-Rate by 9.93% and 7.44% on nuScenes-Occ and ScanNet-Occ, respectively, while the semantic category context achieves reductions of 7.38% and 3.66%. Consistent gains across both datasets show that semantic information is effective across domains. Combining both further improves the reductions to 12.54% and 8.36%. Fig. 5 shows consistent qualitative improvements, with the combined model recovering local structures more faithfully at lower bitrates. The stronger gain from the label relative spatial context suggests that local semantic continuity and boundaries provide more direct cues for occupancy prediction, while category identity supplies complementary information.

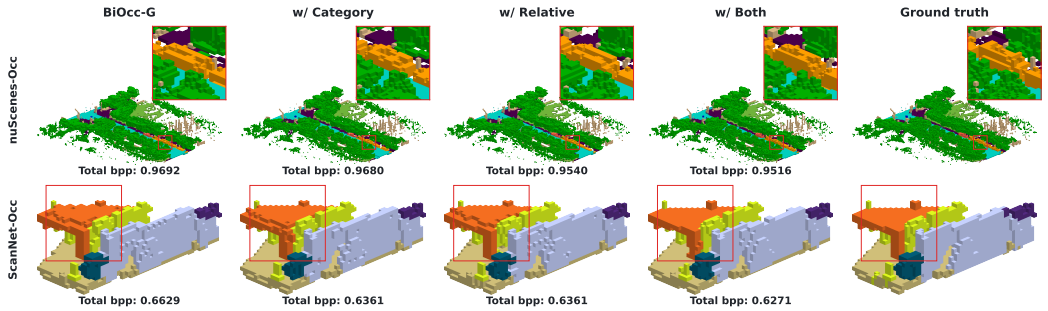

Figure 5: Qualitative comparison of the sem-geo context ablation on nuScenes-Occ and ScanNet-Occ. We compare BiOcc-G with semantic category context, label relative spatial context, and their combination. The numbers below each reconstruction denote the total bitrate in bpp, and the enlarged regions highlight local differences in reconstruction.

Table 2: Ablation on sem-geo context.

Model	BD-Rate to BiOcc-G	
	nuScenes-Occ	ScanNet-Occ
w/ Relative	-9.93%	-7.44%
w/ Category	-7.38%	-3.66%
w/ Both	-12.54%	-8.36%

Table 3: Ablation on conditioning level.

Level	BD-Rate to BiOcc-G	
	nuScenes-Occ	ScanNet-Occ
$L-1$	-12.54%	-8.36%
$L-2$	-3.31%	-1.69%
$L-3$	-2.74%	-0.64%

Semantic conditioning at different hierarchy levels. We activate the sem-geo context model at only one hierarchy level per ablation, while all other levels use geometry only. Specifically, the $L-k$ setting applies semantic context exclusively at level $L-k$, allowing us to isolate its geometry coding benefit at different hierarchy levels. As shown in Table 3, conditioning at $L-1$ achieves the largest BD-Rate reductions of 12.54% and 8.36% on nuScenes and ScanNet, respectively, whereas the gains decrease substantially at $L-2$ and $L-3$. This reduced benefit at coarser levels may arise from semantic aggregation, which suppresses local semantic distinctions and leaves less information relevant to geometry for occupancy prediction, consistent with the analysis in Sec. 3.2.

Contributions of geo-sem context. To assess contextual semantic modeling, we remove all geometry contexts from the semantic entropy model. Removing the geometry context also eliminates the spatial support required to construct local semantic contexts. Therefore, the semantics reduce to a sequence of discrete labels, which are coded independently using a global categorical prior estimated from the training set. Removing contextual modeling increases the semantic bitrate by $9.39\times$ and $4.85\times$ on nuScenes-Occ and ScanNet-Occ, respectively.

5 Conclusion

We present BiOcc, a bidirectional context modeling framework that leverages semantic information to improve geometry coding while using reconstructed geometry to facilitate semantic coding. Its coarse-to-fine interleaved schedule enables both semantic-to-geometry and geometry-to-semantic dependencies during causal decoding. Central to BiOcc is a sem-geo context model that combines label relative spatial context and semantic category information to refine geometry-based occupancy probabilities. Our information-theoretic analysis characterizes when semantic conditioning can reduce geometry coding rate and explains why semantic aggregation at coarser scales weakens this benefit, a trend further supported by our ablation results across both datasets. Experiments on nuScenes-Occ and ScanNet-Occ demonstrate state-of-the-art compression performance across both outdoor and indoor scenes. Together, these findings highlight that semantic labels are not merely attributes to be compressed, but also informative context for coding the underlying geometry. Future work will extend the bidirectional context modeling framework to temporal occupancy compression, exploiting both spatial and inter-frame dependencies between geometry and semantics.

AI USE STATEMENT

In this work, we used generative AI tools to assist with code implementation and debugging during the development of the proposed method. We did not use generative AI tools to generate synthetic datasets, formulate mathematical claims or proofs, or perform qualitative or thematic data analysis. Generative AI tools were also used for improving the readability and clarity of the manuscript and assisting with the preparation of figures and experimental scripts.

REPRODUCIBILITY STATEMENT

We provide the implementation and experimental details necessary to reproduce our results. The model architecture and bidirectional coding pipeline are described in Sec. 3, including the causal information flow between geometry and semantics. All datasets used in our experiments are publicly available, and the corresponding preprocessing procedures, evaluation protocols, baseline adaptations, training settings, and implementation details are provided in Sec. 4 and the Appendix. We will release the source code to facilitate reproduction of the reported results.

REFERENCES

Cheng Chen, Hao Huang, and Saurabh Bagchi. Vision-only gaussian splatting for collaborative semantic occupancy prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 2796–2804, 2026.

Chunyang Fu, Ge Li, Rui Song, Wei Gao, and Shan Liu. Octattention: Octree-based large-scale contexts model for point cloud compression. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 625–633, 2022.

Qingju Guo, Shuang Li, Binhui Xie, Jing Geng, and Wei Li. Spsc: Sparse and scalable multi-modal 3d occupancy prediction for autonomous driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 4430–4438, 2026.

Kai Hsiang Hsieh, Monyneath Yim, and Jui Chiu Chiang. Sedd-pcc: A single encoder-dual decoder framework for end-to-end learned point cloud compression. *arXiv preprint arXiv:2505.16709*, 2025.

Kai-Hsiang Hsieh, Monyneath Yim, Wen-Hsiao Peng, and Jui-Chiu Chiang. Mega-pcc: A mamba-based efficient approach for joint geometry and attribute point cloud compression. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1820–1830, 2026.

Bohan Li, Jiazhe Guo, Hongsi Liu, Yingshuang Zou, Yikang Ding, Xiwu Chen, Hu Zhu, Feiyang Tan, Chi Zhang, Tiancai Wang, et al. Uniscene: Unified occupancy-centric driving scene generation. In *Proceedings of the computer vision and pattern recognition conference*, pp. 11971–11981, 2025.

Feng Li, Chaokun Zhang, and Gong Chen. Sparse-aware vector quantization for bandwidth-efficient collaborative 3d semantic occupancy prediction. *arXiv preprint arXiv:2607.01928*, 2026.

Kangrui Luo, Donghan Bu, Anhong Wang, Junhui Hou, and Yakun Yang. Osc-net: Object semantic-aware compression network for 3d point cloud. *APSIPA Transactions on Signal and Information Processing*, 14(2):e101, 2025.

Daniele Mari, Elena Camuffo, and Simone Milani. Cactus: Content-aware compression and transmission using semantics for automotive lidar data. *Sensors*, 23(12):5611, 2023.

Fangzhe Nan, Frederick WB Li, Gary KL Tam, Zhaoyi Jiang, Bailin Yang, Jingke Cui, and Changshuo Wang. R^2d-lpcc: Relevance-ranking guided region-adaptive dynamic lidar point cloud compression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 15573–15581, 2026.

Jingyi Pan, Zipeng Wang, and Lin Wang. Co-occ: Coupling explicit feature fusion with volume rendering regularization for multi-modal 3d semantic occupancy prediction. *IEEE Robotics and Automation Letters*, 9(6):5687–5694, 2024.

Michael Rudolph, Aron Riemenschneider, and Amr Rizk. Unified compression of point cloud geometry and attributes through variable-rate conditioning. In *Proceedings of the ACM Multimedia Systems Conference 2026*, pp. 1–12, 2026.

Johannes L Schönberger, Marc Pollefeys, Andreas Geiger, and Torsten Sattler. Semantic visual localization. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6896–6906. IEEE, 2018.

Sebastian Schwarz, Marius Preda, Vittorio Baroncini, Madhukar Budagavi, Pablo Cesar, Philip A Chou, Robert A Cohen, Maja Krivokuća, Sébastien Lasserre, Zhu Li, et al. Emerging mpeg standards for point cloud compression. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 9(1):133–148, 2018.

Yining Shi, Kun Jiang, Qiang Meng, Ke Wang, Jiabao Wang, Wenchao Sun, Tuopu Wen, et al. Come: Adding scene-centric forecasting control to occupancy world model. *Advances in Neural Information Processing Systems*, 38:70–96, 2026.

Nikolaos Stathoulopoulos, Christoforos Kanellakis, and George Nikolakopoulos. Have we scene it all? scene graph-aware deep point cloud compression. *IEEE Robotics and Automation Letters*, 2025.

Erik Stenborg, Carl Toft, and Lars Hammarstrand. Long-term visual localization using semantically segmented images. In *2018 IEEE international conference on robotics and automation (ICRA)*, pp. 6484–6490. IEEE, 2018.

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. *Advances in Neural Information Processing Systems*, 36:64318–64330, 2023.

Jianqiang Wang, Ruixiang Xue, Jiaxin Li, Dandan Ding, Yi Lin, and Zhan Ma. A versatile point cloud compressor using universal multiscale conditional coding – part i: Geometry. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(1):269–287, 2025a. doi: 10.1109/TPAMI.2024.3462938.

Kangli Wang, Qianxi Yi, Yuqi Ye, Shihao Li, and Wei Gao. Anypcc: Compressing any point cloud with a single universal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2972–2982, 2026.

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

Xinjie Wang, Yifan Zhang, Ting Liu, Xinpu Liu, Ke Xu, Jianwei Wan, Yulan Guo, and Hanyun Wang. Topnet: Transformer-efficient occupancy prediction network for octree-structured point cloud geometry compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27305–27314, 2025b.

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.

Liang Xie, Wei Gao, Huiming Zheng, and Hua Ye. Semantic-aware visual decomposition for point cloud geometry compression. In *2024 Data Compression Conference (DCC)*, pp. 595–595. IEEE, 2024.

Zikun Xu and Shaobing Xu. Mergeocc: Bridge the domain gap between different lidars for robust occupancy prediction. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 26539–26548. IEEE, 2025.

Fei Xue, Ignas Budvytis, and Roberto Cipolla. Sfd2: Semantic-guided feature detection and description. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5206–5216. IEEE, 2023.

Xiuyu Yang, Bohan Li, Shaocong Xu, Nan Wang, Chongjie Ye, Zhaoxi Chen, Minghan Qin, Yikang Ding, Zheng Zhu, Xin Jin, Hang Zhao, and Hao Zhao. Orv: 4d occupancy-centric robot video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1053–1066, June 2026.

Hongxiao Yu, Yuqi Wang, Yuntao Chen, and Zhaoxiang Zhang. Monocular occupancy prediction for scalable indoor scenes. In *European Conference on Computer Vision*, pp. 38–54. Springer, 2024.

Junzhe Zhang, Gexin Liu, Junteng Zhang, Dandan Ding, and Zhan Ma. Deeppcc: Learned lossy point cloud compression. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 9(2):1897–1909, 2024.

Yun Zhang, Zhiwei Guo, Zixi Guo, Linwei Zhu, and C-C Jay Kuo. Deep-jgac: End-to-end deep joint geometry and attribute compression for dense colored point clouds. *IEEE Transactions on Circuits and Systems for Video Technology*, 2026.

Lili Zhao, Kai-Kuang Ma, Zhili Liu, Qian Yin, and Jianwen Chen. Real-time scene-aware lidar point cloud compression using semantic prior representation. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(8):5623–5637, 2022.

Huiming Zheng and Wei Gao. Diffpcgc: Efficient point cloud geometry compression via diffusion models. In *2026 IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 564–568. IEEE, 2026.

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

Hu Zhu, Bohan Li, Xianda Guo, Yanlun Peng, Zheng Zhu, Xin Jin, Wenjun Zeng, and Chang Wen Chen. Bridging 3d gaussians and semantic occupancy for comprehensive open-vocabulary scene understanding from unposed images. *arXiv preprint arXiv:2607.01633*, 2026a.

Jiahao Zhu, Kang You, Dandan Ding, and Zhan Ma. PACE: Post-causal entropy modeling for learned liDAR point cloud compression. In *Forty-third International Conference on Machine Learning*, 2026b. URL https://openreview.net/forum?id=rwBtuEvV4Q.

A Effect of Semantic Coarsening on Geometry Coding

Let M denote the child-occupancy symbol to be predicted, C_G the available geometric context, and C_S^{v+1} and C_S^v the semantic contexts at two successive hierarchy levels.

Conditioned on the corresponding geometry, the coarse semantic context C_S^v is obtained by a deterministic aggregation of the finer context:

$$C_S^v = A(C_S^{v+1}). \tag{6}$$

Therefore, conditioned on C_G, we have the Markov relation

$$M \longrightarrow C_S^{v+1} \longrightarrow C_S^v. \tag{7}$$

By the conditional data-processing inequality,

$$I(M; C_S^v \mid C_G) \leq I(M; C_S^{v+1} \mid C_G). \tag{8}$$

The condition for strict inequality follows from the chain rule for conditional mutual information. Since C_S^v is a deterministic function of C_S^{v+1},

$$\begin{aligned} &I(M; C_S^{v+1} \mid C_G) - I(M; C_S^v \mid C_G) \\ &\quad = I(M; C_S^{v+1} \mid C_S^v, C_G) \geq 0. \end{aligned} \tag{9}$$

Hence,

$$I(M; C_S^v \mid C_G) < I(M; C_S^{v+1} \mid C_G) \tag{10}$$

whenever

$$I(M; C_S^{v+1} \mid C_S^v, C_G) > 0. \tag{11}$$

This condition means that the fine-scale semantic context retains occupancy-relevant information that is not preserved by its coarse representation.

Equivalently, strict inequality occurs when, for some geometric and coarse semantic contexts with nonzero probability, different fine-scale semantic configurations mapped to the same coarse representation induce different conditional distributions of M. Therefore, semantic aggregation can discard geometry-relevant local structure, providing an information-theoretic motivation for introducing semantic conditioning at finer hierarchy levels.

B Model Details

B.1 Geometry Branch

Following previous work (Fu et al., 2022; Wang et al., 2025b; Zhu et al., 2026b), our geometry branch uses ancestor-path information and causal neighborhood occupancy as context for predicting the current node's child occupancy code. Ancestor occupancy symbols, tree levels, and octant indices are embedded and processed by a causal self-attention backbone to capture hierarchical dependencies. The resulting path representation is combined with features from the available neighboring occupancy patterns to model local geometric structure.

The geometry context includes local occupancy patterns, ancestor-path information, coordinates, tree levels, projection context, and axial occupancy histories. Discrete inputs are mapped to learned embeddings, while coordinate information is represented using positional features. A causal self-attention network captures hierarchical dependencies along the ancestor path, and lightweight MLPs encode local and directional context. The resulting features are concatenated and processed by an MLP prediction head to estimate probabilities over the 255 non-empty child occupancy patterns. These probabilities are used for arithmetic coding, with identical contexts constructed at the encoder and decoder.

For geometry-only coding, the predicted distribution is used directly for arithmetic coding. For semantic-conditioned geometry coding, the geometry branch also provides a 256-dimensional feature g_i and baseline logits $z_{\mathrm{geo},i}$ to the semantic fusion module. Semantic information refines these logits through residual prediction.

B.2 SEMANTIC BRANCH

The semantic branch estimates label probabilities using available semantic context, coarse-level labels, and reconstructed geometry. Each neighborhood label is mapped to a learned embedding and combined with an embedding of its relative spatial position. The resulting tokens are concatenated in a fixed spatial order and processed by an MLP to obtain a semantic context representation. The corresponding coarse-level label is embedded separately, while node coordinates are encoded using Fourier features. These representations are concatenated to retain local semantic dependencies, coarse-scale category information, and spatial location.

A parallel geometry encoder processes the reconstructed local occupancy pattern using an MLP. The prediction network consists of a semantic baseline head and a geometry-conditioned correction head. The baseline head maps the semantic context, coarse-label embedding, and coordinate features to class logits. The correction head additionally receives the geometry features and adjusts the logits based on the local occupancy structure. Both heads are implemented as MLPs, with one output logit for each semantic category.

The final logits are normalized by softmax to obtain a categorical distribution, which is converted into a cumulative distribution for arithmetic coding of the semantic labels. After each coding stage, the newly decoded labels are incorporated into the context for subsequent stages. The decoder follows the same order and reconstructs the same contexts and probability distributions to recover the semantic bitstream.

C IMPLEMENTATION DETAILS

C.1 BASELINES SETTINGS

G-PCC. We adapt the MPEG TMC13 G-PCC implementation to semantic occupancy grids by representing each occupied voxel as a point at its integer grid coordinate. Semantic classes are mapped to uniformly spaced 8-bit reflectance codewords and coded through the existing attribute codec. Since averaging reflectance values can mix categorical labels, we add a semantic-aware recoloring option that assigns each reconstructed point the codeword of its nearest input point, with deterministic geometric tie-breaking. Decoded reflectance values are mapped back to the nearest class codeword.

AnyPCC. We extend both the AnyPCC pipelines to accept OCC grids as sets of occupied integer coordinates with unit features. The OCC path skips point-cloud-specific input scaling, while existing geometry coding stages remain in use. We adapt the training and compression entry points to directly consume OCC samples. Semantic labels determine occupancy but are not included in the compressed representation.

TopNet. We adapt TopNet to OCC grids by constructing its octree directly from occupied integer voxels. We retain the original Transformer-based occupancy predictor and arithmetic coder. For lossy geometry coding, we transmit all octree levels before the finest level losslessly. A decoder-reproducible prediction pass ranks finest-level parent masks by confidence: the least confident fraction is coded exactly, while the remaining masks are reconstructed by constrained top-K selection from predicted child probabilities using a transmitted residual child count. Semantic labels identify occupied voxels but are not coded.

Unicorn. We retain Unicorn's lossless geometry model and occupancy arithmetic coder, and apply them to the sparse occupied coordinates of OCC grids. To code semantics, we add a sparse ResNet that predicts class probabilities from decoded geometry and an entropy-coded refinement stream for the class labels. Refinement layers are ordered by prediction uncertainty, allowing a decoder to recover either predicted labels or progressively more exact labels from the same geometry reconstruction. Separately, we adapt Unicorn's original lossy AE/SR geometry path to OCC coordinates while fixing point-cloud positional quantization at unit grid resolution.

UPCC. We adapt UPCC from joint geometry-color coding to joint geometry-semantic occupancy coding. Occupied voxels form a sparse coordinate set, and each class label is represented by a

learned 16-dimensional embedding concatenated with an occupancy indicator in place of RGB features. We retain the three-stage sparse analysis and synthesis transforms and the mean-scale hyperprior, but replace color reconstruction with a semantic classification head applied to the reconstructed occupied support. Separate quality controls govern occupancy and semantic fidelity.

C.2 Experimental Details

Our implementation operates on voxelized geometry and discrete semantic labels. At each modeled octree transition, an occupied parent node is represented by one of 255 non-empty child-occupancy patterns. Coarse semantic labels are constructed recursively by majority voting over occupied child nodes at the immediately finer level, with ties resolved by the smallest category index. In the default configuration, the geometry is coded without semantic conditioning up to level $L-1$.

The geometry model is trained with an occupancy code negative log-likelihood objective and an auxiliary unconditioned-geometry loss weighted by 0.25. We use AdamW with weight decay 10^{-4} and a batch size of 8,192 parent-node examples. The learning rate is 5×10^{-4} for the first three epochs and 1.5×10^{-4} for the remaining two epochs. The semantic model is trained separately from random initialization. It uses AdamW with a constant learning rate of 2×10^{-4}, weight decay 10^{-4}, and a frame batch size of one. Each training example contains the semantic nodes of one frame, with an internal processing granularity of 4,096 nodes. Training is performed for five epochs on the training split of each dataset. We evaluate experiments on a workstation with NVIDIA RTX 4090 GPU and Intel Xeon Gold 6330 CPU.

Since semantic occupancy is defined on a fixed voxel grid, correct semantic reconstruction requires both correct occupancy and the correct category at each voxel. Missing or spuriously reconstructed occupied voxels also introduce semantic reconstruction errors. We consequently control semantic fidelity primarily through the geometry reconstruction.

The parameter $\lambda_G \in [0, 1]$ controls the fraction of parent nodes whose child-occupancy patterns are explicitly transmitted, prioritizing nodes with lower prediction confidence. For the remaining nodes, occupancy is reconstructed from predicted child probabilities using a constrained Top-K selection, with the occupied-child count K transmitted as side information and at least one child retained per parent. The reconstructed geometry is then used by the semantic codec as context to encode and decode next-level semantic labels. Varying λ_G thus controls the reconstructed spatial support and produces the operating points for both the geometry-IoU and semantic-mIoU rate–distortion curves.

The geometry rate includes geometry payloads and all shared container overhead, including headers, segment lengths, model/prior/runtime fingerprints, and checksums. The semantic rate includes all transmitted semantic side-information payloads and their internal framing. The total rate is the sum of these two rates. Model weights and the fixed root prior are shared decoder assets and are not included in the per-stream rate. Encoding and decoding run in separate processes. The decoder uses only the bitstream, the shared model assets, and the fixed codec configuration to reproduce the encoder's probability distributions from decoder-available context. For every operating point, independently decoded geometry must exactly match the encoder-side reconstruction. Transmitted semantic labels are checked for exact recovery against the corresponding input side-information labels, including an independent-decoder hash comparison.

C.3 Metrics Details

Let $X = (G, S)$ and $\hat{X} = (\hat{G}, \hat{S})$ denote the original and reconstructed semantic occupancy maps, respectively. All metrics are computed on the same voxel grid and within the valid evaluation region. Here, $|\cdot|$ denotes set cardinality.

Geometry IoU. Geometry IoU measures the overlap between the original and reconstructed occupied voxel sets, irrespective of their semantic labels:

$$\mathrm{IoU} = \frac{|G \cap \hat{G}|}{|G \cup \hat{G}|} = \frac{\mathrm{TP}_G}{\mathrm{TP}_G + \mathrm{FP}_G + \mathrm{FN}_G}, \tag{12}$$

where $\mathrm{TP}_G = |G \cap \hat{G}|$ counts correctly reconstructed occupied voxels, $\mathrm{FP}_G = |\hat{G} \setminus G|$ counts falsely reconstructed occupied voxels, and $\mathrm{FN}_G = |G \setminus \hat{G}|$ counts missing occupied voxels. Empty voxels correctly reconstructed as empty do not contribute to this metric.

Semantic mIoU. For each evaluated semantic class k, define

$$G_k = \{v \in G : S(v) = k\}, \qquad \hat{G}_k = \{v \in \hat{G} : \hat{S}(v) = k\}. \tag{13}$$

The class-specific IoU and semantic mIoU are

$$\mathrm{IoU}_k = \frac{|G_k \cap \hat{G}_k|}{|G_k \cup \hat{G}_k|}, \qquad \mathrm{mIoU} = \frac{1}{|\mathcal{K}|} \sum_{k \in \mathcal{K}} \mathrm{IoU}_k, \tag{14}$$

where $\mathcal{K}$ is the fixed set of evaluated semantic classes, excluding the empty-space class. A voxel contributes to the intersection only when both its occupancy and semantic label are correct. Missing voxels, falsely reconstructed voxels, and incorrect semantic labels all affect mIoU.

Coding rate. We measure coding rate in bits per point (bpp), where a point corresponds to an occupied voxel in the original map:

$$R = \frac{B}{|G|}. \tag{15}$$

Here, B is the total number of bits in the compressed bitstream. The denominator is the original occupied voxel count, rather than the reconstructed count.

Bjøntegaard delta bit rate. BD-Rate measures the average relative bitrate difference between a tested method and a reference method at matched reconstruction quality. We compute it separately using geometry IoU and semantic mIoU as the quality variable q. Let $\ell_{\mathrm{test}}(q)$ and $\ell_{\mathrm{ref}}(q)$ denote interpolated natural-log bitrate curves. The integration interval is the common quality range:

$$q_{\min} = \max(q_{\mathrm{test}}^{\min}, q_{\mathrm{ref}}^{\min}), \qquad q_{\max} = \min(q_{\mathrm{test}}^{\max}, q_{\mathrm{ref}}^{\max}). \tag{16}$$

For $q_{\max} > q_{\min}$, we compute

$$\delta = \frac{1}{q_{\max} - q_{\min}} \int_{q_{\min}}^{q_{\max}} \left[\ell_{\mathrm{test}}(q) - \ell_{\mathrm{ref}}(q)\right] \, dq, \tag{17}$$

and

$$\text{BD-Rate}(\%) = 100\big(\exp(\delta) - 1\big). \tag{18}$$

We use piecewise-linear interpolation in the quality-log-rate domain and integrate over the common quality range without extrapolation. Negative BD-Rate indicates bitrate savings relative to the reference; for example, -10% indicates a 10% reduction in the geometric-mean rate ratio over the evaluated quality interval.