ADAGSFORMER: COMPLEXITY-ADAPTIVE GAUSSIAN EVOLUTION FOR 3D SEMANTIC OCCUPANCY PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Gaussian-based representations offer an efficient alternative to dense volumetric features for 3D semantic occupancy prediction. Existing methods adapt the Gaussian population mainly through progressive densification or allocation under a predefined final budget, limiting their ability to increase capacity in structurally complex regions while removing redundancy elsewhere. To this end, we propose **AdaGSFormer**, a complexity-adaptive Gaussian evolution framework that adjusts the population size according to local scene structure without prescribing its final cardinality. AdaGSFormer predicts Gaussian-wise complexity scores using supervision derived from local semantic heterogeneity and uses scene-relative thresholds to guide population restructuring. It splits Gaussians in complex regions and removes redundant primitives through contribution-aware pruning and moment-preserving merging. To decode the resulting variable-size population, we formulate Gaussian contributions and Gaussian-to-voxel splatting in additive optical-density space. On the SurroundOcc dataset, AdaGSFormer achieves state-of-the-art performance in semantic occupancy prediction, while offering approximately $2.5\times$ faster inference and 57.8% lower GPU memory usage compared to representative Gaussian methods. These results demonstrate the effectiveness of complexity-adaptive Gaussian evolution for accurate and efficient 3D semantic occupancy prediction.

1 INTRODUCTION

3D semantic occupancy prediction reconstructs geometric occupancy and voxel-level semantics from multi-view images, providing a scene representation that extends beyond predefined object instances. Existing methods use dense voxels, tri-perspective features, sparse latent queries, compact transformers, and low-rank representations to balance accuracy and computational cost Huang et al. (2023); Wei et al. (2023); Zhang et al. (2023); Wang et al. (2023); Tang et al. (2024); Ma et al. (2024); Zhao et al. (2024); Huang et al. (2024a); Oh et al. (2025). Nevertheless, interaction among high-resolution 3D features remains expensive, motivating more compact and flexible scene representations.

3D Gaussian Splatting (3DGS) Kerbl et al. (2023) offers an explicit representation of 3D space. GaussianFormer Huang et al. (2024b) introduces semantic Gaussians for vision-based occupancy prediction, and GaussianFormer-2 Huang et al. (2025) develops a probabilistic Gaussian formulation. Subsequent work extends Gaussian occupancy prediction to self-supervision, temporal modeling, streaming perception, and geometry-aware initialization Gan et al. (2025); Chambon et al. (2025); Boeder et al. (2025); Zuo et al. (2025); Park et al. (2025); Yan et al. (2026); Lin et al. (2026); Li et al. (2026). More recently, methods have adapted Gaussian populations through progressive densification or explicit primitive allocation Yan & Xu (2026); Ning et al. (2026). Although these methods move beyond static populations, their adaptation is constrained by progressive expansion or a prescribed final Gaussian budget. They therefore leave open how to redistribute representation capacity through both expansion and compression without fixing the final population size.

This question is particularly relevant to driving scenes, where representation demand varies substantially across space. Broad, semantically homogeneous surfaces may need fewer primitives, whereas

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.

object boundaries, semantically mixed regions, and crowded foreground areas may require finer representation. Effective adaptation must therefore identify where capacity is insufficient and where it may be redundant. Yet structural simplicity alone cannot justify removing a primitive: even in a homogeneous region, a Gaussian may provide essential occupancy evidence. Local structure should guide where capacity is reconsidered, while primitive contribution and redundancy should determine whether a candidate is split, pruned, merged, or retained.

To address this, we propose **AdaGSFormer**, a complexity-adaptive Gaussian evolution framework for 3D semantic occupancy prediction. AdaGSFormer predicts Gaussian-wise structural complexity using supervision derived from local semantic heterogeneity. It then derives decision boundaries from the complexity distribution of each scene, allowing population restructuring to respond to scene-specific variation. High-complexity Gaussians undergo learnable splitting to increase local capacity. Low-complexity Gaussians are evaluated for compression: contribution-aware pruning removes dispensable primitives, moment-preserving merging consolidates redundant ones, and primitives that remain informative are retained. The final Gaussian cardinality thus results from local decisions rather than a prescribed budget. To decode the resulting variable-cardinality representation into dense occupancy, we further formulate Gaussian contribution and Gaussian-to-Voxel splatting in additive optical-density space.

Our main contributions are summarized as follows:

- We propose **AdaGSFormer**, a Gaussian evolution framework that adapts representation capacity to each scene without prescribing its final Gaussian cardinality.
- We introduce Gaussian-wise structural complexity estimation and scene-relative decision boundaries to guide local population restructuring. Learnable splitting, contribution-aware pruning, and moment-preserving merging adjust capacity while retaining informative primitives.
- We formulate optical-density Gaussian-to-Voxel splatting (OD-G2V) to decode variable-cardinality Gaussian populations through additive contributions.
- Extensive experiments on SurroundOcc demonstrate strong semantic occupancy performance and substantially improved Gaussian representation efficiency. Controlled fixed-capacity comparisons and mechanism analyses further verify that the gains arise from complexity-guided population restructuring rather than simply increasing the primitive budget.

2 Related work

2.1 Gaussian-based 3D Semantic Occupancy Prediction

GaussianFormer Huang et al. (2024b) introduces sparse semantic Gaussians into vision-based occupancy prediction, while GaussianFormer-2 Huang et al. (2025) formulates Gaussian superposition probabilistically. Subsequent works extend this paradigm with weak or self-supervision, rendering consistency, temporal modeling, streaming perception, and geometry-aware initialization Sun et al. (2024); Gan et al. (2025); Chambon et al. (2025); Ye et al. (2025); Boeder et al. (2025); Zhu et al. (2025); Zuo et al. (2025); Park et al. (2025); Yan et al. (2026); Zhou et al. (2026); Lin et al. (2026); Li et al. (2026). Recent studies further investigate adaptive Gaussian populations. PG-Occ Yan & Xu (2026) employs progressive online densification to incrementally augment Gaussian representations, while SAGFormer Ning et al. (2026) formulates semantic Gaussian allocation through candidate generation and global selection under a fixed final budget. These methods demonstrate the benefit of adapting Gaussian populations beyond attribute refinement. AdaGSFormer explores a complementary direction by using Gaussian-wise structural complexity and scene-relative criteria to guide local population restructuring, allowing the resulting cardinality to emerge from split, prune, merge, and retain decisions rather than from progressive densification or an explicitly enforced final budget.

2.2 Adaptive Gaussian Density Management

The 3DGS Kerbl et al. (2023) adapts primitive density through cloning, splitting, and pruning during per-scene optimization. Subsequent studies improve densification, pruning, compression, and

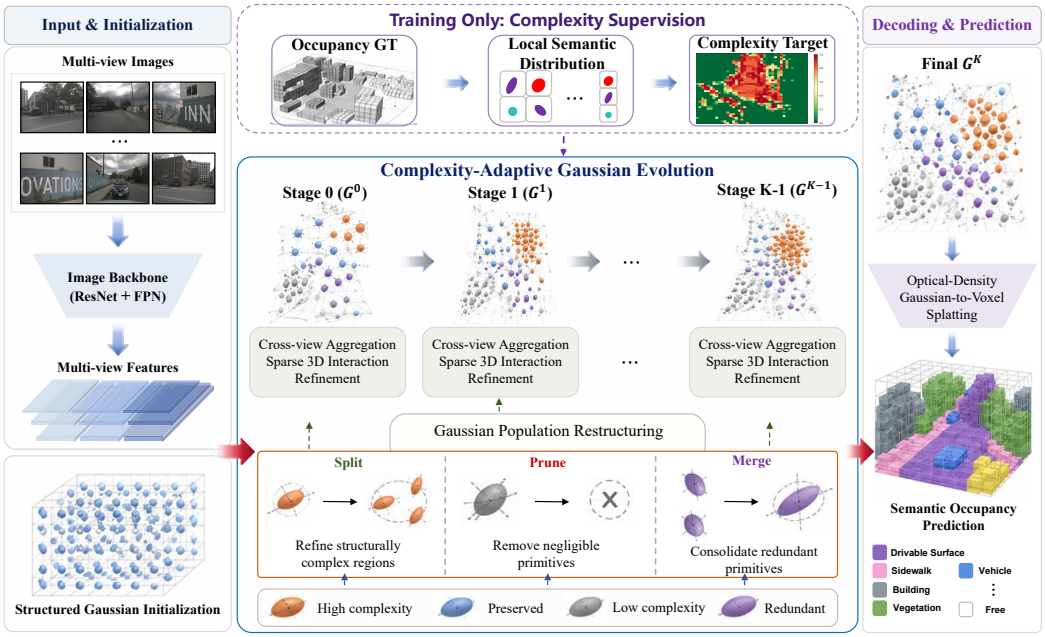

Figure 1: Overall architecture of AdaGSFormer. Starting from structured Gaussian initialization, the model progressively updates Gaussian representations through cross-view aggregation, sparse 3D interaction, and refinement, while adaptively evolving the Gaussian population via splitting, pruning, and merging. The resulting Gaussians are decoded into dense semantic occupancy through Optical-Density Gaussian-to-Voxel Splatting (OD-G2V).

Gaussian merging using structured anchors, refined gradient criteria, importance estimation, sparsification, and hierarchical representations Lu et al. (2024); Ye et al. (2024); Lee et al. (2024); Niedermayr et al. (2024); Fan et al. (2024); Liu et al. (2025); Yang et al. (2025); Wang et al. (2025); Fang & Wang (2024); Zhang et al. (2025); Hanson et al. (2025); Wang et al. (2026); Kerbl et al. (2024); Bai et al. (2025). These techniques primarily target rendering, reconstruction, or SLAM and rely on iterative optimization or rendering feedback. In contrast, AdaGSFormer performs scene-conditioned population restructuring within a single feed-forward occupancy pass, without relying on per-scene optimization or rendering feedback.

3 Methodology

3.1 Gaussian-Based Semantic Scene Representation

Given multi-view images $\mathcal{I} = \{I_n\}_{n=1}^{N_{\text{cam}}}$, 3D semantic occupancy prediction aims to estimate a voxel-wise semantic field $\mathcal{O} \in \{1, \ldots, C\}^{H \times W \times D}$, where H, W, and D denote the spatial resolution and C is the number of semantic categories including free space.

Following Gaussian-based occupancy representations (Huang et al., 2024b; 2025), we represent a scene by a sparse set of anisotropic semantic Gaussians $\mathcal{G} = \{G_i\}_{i=1}^{N}$. Each primitive is parameterized as $G_i = (\boldsymbol{\mu}_i, \mathbf{s}_i, \mathbf{r}_i, \alpha_i, \mathbf{c}_i)$, where $\boldsymbol{\mu}_i \in \mathbb{R}^3$, $\mathbf{s}_i \in \mathbb{R}^3$, $\mathbf{r}_i \in \mathbb{R}^4$, $\alpha_i \in (0, 1)$, and $\mathbf{c}_i \in \mathbb{R}^C$ denote its center, scale, rotation, opacity, and semantic logits, respectively. Let $\mathbf{R}_i$ be the rotation matrix converted from $\mathbf{r}_i$. The covariance of G_i is defined as $\boldsymbol{\Sigma}_i = \mathbf{R}_i \operatorname{diag}(\mathbf{s}_i^2)\mathbf{R}_i^{\mathrm{T}}$, and its spatial response at an arbitrary location $\mathbf{x} \in \mathbb{R}^3$ is

$$g_i(\mathbf{x}) = \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^{\top} \boldsymbol{\Sigma}_i^{-1} (\mathbf{x} - \boldsymbol{\mu}_i)\right]. \tag{1}$$

In AdaGSFormer, the population cardinality is allowed to change during decoding as a consequence of local restructuring, and the final Gaussian count is not prescribed in advance.

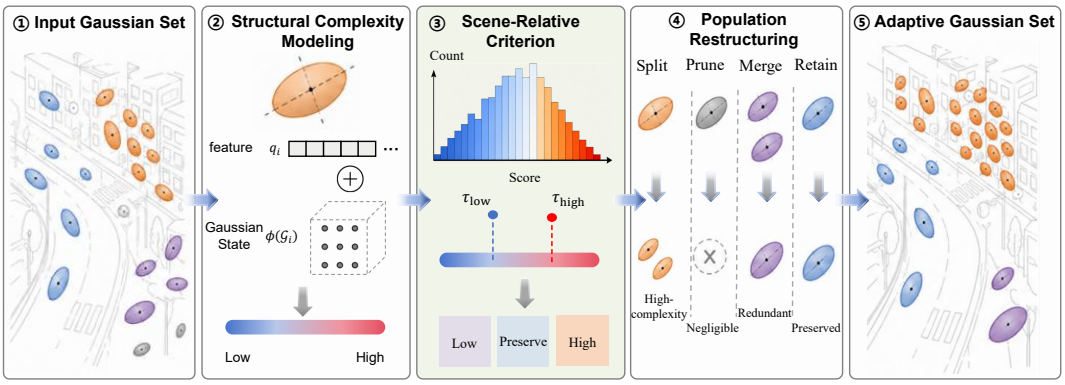

Figure 2: Complexity-adaptive Gaussian evolution in AdaGSFormer. A Gaussian-wise predictor estimates local structural complexity under semantic-heterogeneity supervision. Scene-relative thresholds partition the population into high-, preserved-, and low-complexity primitives. High-complexity Gaussians are split for finer representation, whereas low-complexity primitives are pruned when negligible, merged when redundant, or otherwise retained.

3.2 Overall Architecture

As illustrated in Fig. 1, AdaGSFormer consists of structured Gaussian initialization, complexity-adaptive Gaussian evolution, and optical-density semantic occupancy decoding. Given multi-view images, an image backbone extracts multi-scale visual features $\mathcal{F}$, while a metric-aligned Gaussian scaffold $\mathcal{G}^0$ provides complete coarse coverage of the perception space. At the k-th decoding stage, cross-view feature aggregation, sparse 3D interaction, and residual Gaussian refinement are abstracted as $\mathcal{R}_k$:

$$\widetilde{\mathcal{G}}^k = \mathcal{R}_k\left(\mathcal{G}^k; \mathcal{F}\right), \qquad \mathcal{G}^{k+1} = \begin{cases} \mathcal{E}_k\left(\widetilde{\mathcal{G}}^k\right), & k \in \mathcal{A}, \\ \widetilde{\mathcal{G}}^k, & k \notin \mathcal{A}, \end{cases} \tag{2}$$

where $\mathcal{A}$ denotes the intermediate stages equipped with adaptive Gaussian evolution, $\mathcal{E}_k$ reallocates representation capacity through splitting, pruning, and merging. The resulting variable-cardinality Gaussian set is finally converted into dense semantic occupancy through Optical-Density Gaussian-to-Voxel splatting (OD-G2V).

3.3 Structured Gaussian Initialization

Dynamic population evolution requires a scene-agnostic initialization with sufficient spatial coverage. We therefore partition the perception region Ω into N_0 non-overlapping metric cells $\{\Omega_i\}_{i=1}^{N_0}$ and associate each cell with an initial Gaussian:

$$\boldsymbol{\mu}_i^0 = \mathrm{ctr}(\Omega_i), \qquad \mathbf{s}_i^0 = \gamma\ \mathrm{extent}(\Omega_i), \qquad \mathcal{G}^0 = \{\mathcal{G}_i^0\}_{i=1}^{N_0}. \tag{3}$$

Here γ controls support overlap. All primitives share weak opacity and semantic priors, providing uniform coarse coverage while leaving scene-specific capacity allocation to subsequent Gaussian evolution.

3.4 Complexity-Adaptive Gaussian Evolution

Effective Gaussian occupancy modeling requires not only adding primitives where finer representation is needed, but also recovering redundant capacity elsewhere. We therefore estimate Gaussian-wise structural complexity during progressive decoding and use it to condition local population restructuring, as illustrated in Fig. 2.

3.4.1 Structural Complexity Modeling

Semantic heterogeneity as a complexity proxy. Local representation becomes more challenging when a Gaussian support spans semantically heterogeneous content, as commonly occurs around

semantic boundaries and mixed-content regions. We therefore use local semantic heterogeneity as a supervision proxy for Gaussian-wise structural complexity. For a refined Gaussian $\widetilde{\mathcal{G}}_i^k$, support-aligned semantic observations form a Gaussian-weighted category distribution $\boldsymbol{\pi}_i^k$, from which we compute the normalized semantic entropy

$$H_i^k = -\frac{1}{\log C} \sum_{c=1}^{C} \pi_{i,c}^k \log(\pi_{i,c}^k + \epsilon). \tag{4}$$

A larger H_i^k indicates stronger local semantic heterogeneity. Rather than requiring semantic annotations at inference time, H_i^k is used only as a training target, while Gaussian-wise complexity is predicted directly from the refined representation:

$$z_i^k = \Psi_\theta \left(\mathbf{q}_i^k, \phi(\widetilde{\mathcal{G}}_i^k) \right), \qquad p_i^k = \sigma(z_i^k). \tag{5}$$

where $p_i^k \in (0, 1)$ denotes the predicted complexity score. The ground-truth-derived (GT-derived) proxy H_i^k is used only for training supervision; all population restructuring decisions at inference are made solely from the predicted score p_i^k.

Scene-relative complexity criterion. An absolute complexity score alone does not determine whether a Gaussian requires population adjustment, since score distributions vary with scene composition and decoding stage. We therefore interpret each Gaussian relative to the complexity distribution of the current population and derive scene-adaptive decision boundaries:

$$\begin{aligned} \tau_{\text{low}}^k &= \Pi_{[0,1]} \left[\min \left(\tau_{\text{low}}^0, \mu_p^k - \lambda_l \sigma_p^k \right) \right], \\ \tau_{\text{high}}^k &= \Pi_{[0,1]} \left[\max \left(\tau_{\text{high}}^0, \mu_p^k + \lambda_h \sigma_p^k \right) \right], \end{aligned} \tag{6}$$

where μ_p^k and σ_p^k denote the mean and standard deviation of the predicted complexity scores at stage k. The resulting criterion identifies Gaussians that are relatively complex or simple under the current scene context, providing candidates for subsequent population restructuring.

3.4.2 Gaussian Population Restructuring

The estimated complexity score indicates where Gaussian capacity may need adjustment, while the specific restructuring operation further depends on each primitive's physical contribution and redundancy. For contribution-aware compression, we define the optical-density mass

$$\rho_i = -\log(1 - \alpha_i), \qquad m_i = \rho_i \sqrt{\det(\boldsymbol{\Sigma}_i)} = \rho_i \prod_{d=1}^{3} s_{i,d}, \tag{7}$$

which provides a physically meaningful measure for pruning and moment-preserving merging.

Learnable splitting. A high-complexity Gaussian may span multiple local structures that cannot be sufficiently represented by a single primitive. For each selected parent $i \in \mathcal{S}^k$, we generate two specialized child Gaussians conditioned on its refined state and instance representation:

$$\left\{ \mathcal{G}_{i,r}^{k+1} \right\}_{r=1}^{2} = \mathcal{S}_\theta \left(\widetilde{\mathcal{G}}_i^k, \mathbf{q}_i^k \right). \tag{8}$$

The children inherit the parent orientation while learning differentiated local positions, scales, opacities, semantic attributes, and latent representations. This allows structurally complex regions to acquire finer representation capacity without simply duplicating the parent primitive.

Contribution-aware pruning. Low structural complexity alone does not imply that a Gaussian is dispensable, since semantically homogeneous regions may still provide essential occupancy support. We therefore prune a low-complexity primitive only when both its opacity and integrated optical-density contribution are negligible:

$$i \in \mathcal{D}^k \iff p_i^k \le \tau_{\text{low}}^k \wedge \alpha_i^k \le \epsilon_\alpha \wedge m_i^k \le \epsilon_m. \tag{9}$$

Moment-preserving merging. Low-complexity Gaussians with non-negligible contribution are retained unless they are redundant with nearby primitives. For a candidate pair (i, j), geometric agreement is measured by the Bhattacharyya distance

$$D_{\mathrm{B}}(i,j) = \frac{1}{8}\boldsymbol{\delta}_{ij}^{\top}\bar{\boldsymbol{\Sigma}}_{ij}^{-1}\boldsymbol{\delta}_{ij} + \frac{1}{2}\log\frac{\det\bar{\boldsymbol{\Sigma}}_{ij}}{\sqrt{\det\boldsymbol{\Sigma}_i\det\boldsymbol{\Sigma}_j}}, \\ \bar{\boldsymbol{\Sigma}}_{ij} = \frac{1}{2}\left(\boldsymbol{\Sigma}_i + \boldsymbol{\Sigma}_j\right), \qquad \boldsymbol{\delta}_{ij} = \boldsymbol{\mu}_i - \boldsymbol{\mu}_j. \tag{10}$$

Together with semantic agreement, this identifies geometrically and semantically redundant pairs. Let $\beta_i = m_i/(m_i + m_j)$ and $\beta_j = 1 - \beta_i$; the merged Gaussian is obtained through density-mass-weighted moment matching:

$$m_{ij} = m_i + m_j, \qquad \boldsymbol{\mu}_{ij} = \beta_i\boldsymbol{\mu}_i + \beta_j\boldsymbol{\mu}_j, \\ \boldsymbol{\Sigma}_{ij} = \beta_i\left(\boldsymbol{\Sigma}_i + \boldsymbol{\Delta}_i\boldsymbol{\Delta}_i^{\top}\right) + \beta_j\left(\boldsymbol{\Sigma}_j + \boldsymbol{\Delta}_j\boldsymbol{\Delta}_j^{\top}\right), \tag{11}$$

where $\boldsymbol{\Delta}_i = \boldsymbol{\mu}_i - \boldsymbol{\mu}_{ij}$ and $\boldsymbol{\Delta}_j = \boldsymbol{\mu}_j - \boldsymbol{\mu}_{ij}$. Semantic attributes and latent representations are aggregated using the same mass weights. The resulting Gaussian approximately preserves the dominant spatial moments and semantic statistics of the merged pair while reducing representation redundancy.

3.5 Optical-Density Gaussian-to-Voxel Splatting

Gaussian-to-Voxel decoding maps sparse semantic Gaussians to dense occupancy predictions. Following the Gaussian-to-Voxel splatting paradigm introduced by GaussianFormer Huang et al. (2024b) and adopted by subsequent Gaussian occupancy methods Park et al. (2025); Qian et al. (2026), voxel evidence is commonly accumulated from opacity-modulated spatial responses. However, opacity is bounded and does not provide an additive measure under changes in primitive decomposition. Since AdaGSFormer explicitly modifies Gaussian cardinality through splitting, pruning, and merging, we instead formulate voxel aggregation in optical-density space:

$$\rho_i = -\log(1 - \alpha_i), \qquad w_i(\mathbf{x}) = \rho_i g_i(\mathbf{x}). \tag{12}$$

The cumulative optical density and occupancy probability are

$$D(\mathbf{x}) = \sum_{i=1}^{N} w_i(\mathbf{x}), \qquad P_{\mathrm{occ}}(\mathbf{x}) = 1 - \exp[-D(\mathbf{x})]. \tag{13}$$

The same optical-density weights are used for semantic aggregation, and the resulting semantic distribution is combined with the complementary free-space probability. Thus, OD-G2V provides an additive decoding interface for variable-cardinality Gaussian representations.

3.6 Training Objectives

AdaGSFormer is optimized with standard semantic occupancy supervision and auxiliary Gaussian-complexity supervision:

$$\mathcal{L} = \mathcal{L}_{\mathrm{occ}} + \lambda_{\mathrm{cmp}}\mathcal{L}_{\mathrm{cmp}}, \tag{14}$$

where $\mathcal{L}_{\mathrm{occ}}$ follows standard occupancy objectives, combining binary occupancy supervision, class-balanced cross-entropy, and Lovasz-Softmax loss Berman et al. (2018).

For each controlled stage, the predicted complexity logit is supervised by the semantic-heterogeneity proxy defined in Eq. 4:

$$\mathcal{L}_{\mathrm{cmp}} = \frac{1}{|\mathcal{A}|}\sum_{k\in\mathcal{A}}\frac{\sum_i \omega_i^k\,\ell_{\mathrm{BCE}}\left(z_i^k, H_i^k\right)}{\sum_i \omega_i^k + \epsilon}, \tag{15}$$

where ℓ_{BCE} denotes binary cross-entropy applied to the predicted logits, and ω_i^k downweights Gaussian supports dominated by free space.

Table 1: Comparison on the SurroundOcc validation set. We report scene completion IoU, semantic scene completion mIoU, and per-class IoU. The best and second-best results are highlighted in **bold** and underlined, respectively.

Method	Overall		Per-class IoU															
	IoU	mIoU	Bar.	Bic.	Bus	Car	C.V.	Mot.	Ped.	T.C.	Tra.	Tru.	D.S.	O.F.	Side.	Terr.	Man.	Veg.
TPVFormer	30.86	17.10	15.96	5.31	23.86	27.32	9.79	8.74	7.09	5.20	10.97	19.22	38.87	21.25	24.26	23.15	11.73	20.81
OccFormer	31.39	19.03	18.65	10.41	23.92	30.29	10.31	14.19	13.59	10.13	12.49	20.77	38.78	19.79	24.19	22.21	13.48	21.35
SurroundOcc	31.49	20.30	20.59	11.68	28.06	30.86	10.70	15.14	14.09	**12.06**	**14.38**	22.26	37.29	23.70	24.49	22.77	14.89	21.86
GaussianFormer	29.83	19.10	19.52	11.26	26.11	29.78	10.47	13.83	12.58	8.67	12.74	21.57	39.63	23.28	24.46	22.99	9.59	19.12
GaussianFormer-2	30.56	20.02	20.15	12.99	27.61	30.23	11.19	15.31	12.64	9.63	13.31	22.26	39.68	23.47	25.62	23.20	12.25	20.73
QuadricFormer	31.22	20.12	19.58	**13.11**	27.27	29.64	11.25	**16.26**	12.65	9.15	12.51	21.24	40.20	24.34	25.69	24.24	12.95	21.86
SplatSSC	**32.19**	19.05	16.20	11.10	25.50	25.50	**12.20**	13.60	11.40	8.50	12.80	19.20	37.30	22.50	25.20	24.80	**15.30**	**23.80**
AdaGSFormer	31.79	**21.13**	**21.72**	12.23	**28.85**	**32.04**	9.57	14.49	**14.21**	11.57	13.42	**24.31**	**42.53**	**27.73**	**27.53**	**25.06**	11.23	21.57

Bar.: barrier; Bic.: bicycle; C.V.: construction vehicle; Mot.: motorcycle; Ped.: pedestrian; T.C.: traffic cone; Tra.: trailer; Tru.: truck; D.S.: driveable surface; O.F.: other flat; Side.: sidewalk; Terr.: terrain; Man.: manmade; Veg.: vegetation.

4 Experiments

4.1 Experimental Setup

Dataset & metrics. We evaluate on SurroundOcc Wei et al. (2023), built upon nuScenes Caesar et al. (2020), using its standard 700/150/150 train/validation/test split. The perception range is $[-50, 50] \times [-50, 50] \times [-5, 3]\,\mathrm{m}^3$ and is discretized into a $200 \times 200 \times 16$ occupancy grid. We report semantic scene completion mIoU and class-agnostic scene completion IoU following prior work Huang et al. (2024b; 2025).

Implementation. We use a ResNet-101-DCN He et al. (2016); Wang et al. (2021) backbone with FPN Lin et al. (2017), a 128-dimensional Gaussian feature, and a four-stage decoder. Images are resized to 1600×900, and each scene is initialized with 10k structured Gaussians. We train for 35 epochs using AdamW with learning rate 2×10^{-4}, weight decay 0.01, and batch size 8 on eight NVIDIA A800 GPUs.

4.2 Main Results

Semantic Occupancy Prediction. As shown in Tab. 1, AdaGSFormer achieves the highest semantic occupancy accuracy with 21.13 mIoU. It improves over SurroundOcc Wei et al. (2023), GaussianFormer-2 Huang et al. (2025), and the recent QuadricFormer Zuo et al. (2026) by 0.83, 1.11, and 1.01 mIoU points, respectively, and surpasses SplatSSC Qian et al. (2026) by 2.08 points. Meanwhile, AdaGSFormer attains 31.79 IoU, ranking second in class-agnostic scene completion and remaining close to SplatSSC (32.19 IoU). This indicates that the main advantage of AdaGSFormer lies in semantic scene representation rather than simply improving occupancy coverage. At the category level, AdaGSFormer achieves the best performance across multiple semantic classes, including barrier, bus, car, pedestrian, truck, driveable surface, other flat, sidewalk, and terrain. The gains span both localized foreground objects and large-scale structural regions, supporting the effectiveness of complexity-adaptive Gaussian evolution for heterogeneous scene structures.

Gaussian Representation Efficiency. Table 2 evaluates efficiency from two complementary perspectives.

Comparison with existing Gaussian representations. AdaGSFormer substantially reduces the computational cost of Gaussian-based occupancy prediction. Compared with GaussianFormer-2, it improves mIoU from 20.02 to 21.13 while reducing latency from 511.41 ms to 205.90 ms and measured GPU memory from 6,037.81 MB to 2,550.17 MB, corresponding to approximately $2.5\times$ faster inference and 57.8% lower memory usage. The comparison with GaussianFormer further highlights the benefit of avoiding a uniformly large primitive population.

Fixed versus adaptive population capacity. We further isolate the effect of population restructuring within the same framework. Increasing the fixed budget from 10.0k to 16.3k Gaussians improves

Table 2: Gaussian representation efficiency. Final #G denotes the average Gaussian cardinality after decoding. Latency and memory are measured with batch size 1 on the same RTX PRO 5000 GPU.

Method	Population	Init. #G	Final #G ↓	mIoU ↑	Latency (ms) ↓	Memory (MB) ↓
Existing Gaussian occupancy methods						
GaussianFormer	Fixed	144.0k	144.0k	19.10	591.99	8,918.13
GaussianFormer-2	Fixed	12.8k	12.8k	20.02	511.41	6,037.81
Controlled population strategies						
	Fixed	10.0k	10.0k	20.20	**189.27**	2,826.66
AdaGSFormer	Fixed	16.3k	16.3k	20.52	245.01	2,839.82
	Adaptive	10.0k	14.3k	**21.13**	205.90	**2,550.17**

mIoU only from 20.20 to 20.52, while increasing both latency and memory. In contrast, AdaGSFormer starts from 10.0k Gaussians and evolves to only 14.3k on average, yet reaches 21.13 mIoU. It therefore outperforms the larger 16.3k fixed-capacity counterpart by 0.61 mIoU with 12.3% fewer final primitives, while reducing latency from 245.01 ms to 205.90 ms and memory from 2,839.82 MB to 2,550.17 MB. These results show that the gain arises from adaptive capacity restructuring rather than simply enlarging a fixed Gaussian budget.

4.3 Ablation and Mechanism Analysis

Gaussian population restructuring. Table 3 reveals distinct roles of the three population operations. Disabling splitting causes the largest accuracy degradation, reducing mIoU from 21.13 to 19.81, indicating that population expansion is critical for resolving structurally complex regions. In contrast, removing pruning preserves comparable accuracy but increases the average population to 17.0k Gaussians and latency to 266.46 ms, showing that pruning primarily controls redundant capacity. Removing merging likewise leaves mIoU nearly unchanged while increasing both cardinality and runtime, suggesting a complementary compression role. Replacing OD-G2V reduces mIoU to 20.14, highlighting the importance of additive optical-density decoding for the restructured variable-cardinality representation.

Table 3: Mechanism analysis of Gaussian population restructuring. S, P, and M denote splitting, pruning, and merging, respectively. The "w/o OD-G2V" variant replaces optical-density decoding with conventional Gaussian-to-Voxel aggregation.

Variant	S	P	M	OD-G2V	mIoU ↑	Latency (ms) ↓	Avg. #G ↓
Fixed population				✓	20.20	189.27	10.0k
w/o Split		✓	✓	✓	19.81	**183.29**	**8.2k**
w/o Prune	✓		✓	✓	21.14	266.46	17.0k
w/o Merge	✓	✓		✓	21.13	232.16	14.6k
w/o OD-G2V	✓	✓	✓		20.14	249.36	13.9k
AdaGSFormer	✓	✓	✓	✓	**21.13**	205.90	14.3k

4.4 Analysis of Complexity-Adaptive Evolution

Complexity-conditioned population restructuring. Fig. 3 shows that the predicted complexity scores exhibit clear spatial correspondence with population restructuring: high-score regions preferentially trigger splitting, whereas low-score primitives are pruned when negligible or merged when redundant. To quantitatively assess this behavior, Appendix A post-hoc groups Gaussians using the GT-derived complexity proxy H_i. Importantly, these GT labels are used only for analysis; all restructuring operations are produced from the predicted scores p_i during inference. The split ratio increases from 7.46% in the low-complexity group to 79.59% in the high-complexity group, confirming strong alignment between the learned complexity signal and local population restructuring.

Fine-grained evolution patterns across scene regions. A closer inspection of Fig. 3 reveals complementary restructuring behaviors across heterogeneous regions. Complex semantic transitions and object boundaries tend to trigger splitting, allocating finer local representation where needed. In contrast, low-complexity Gaussians are pruned only when their contribution is negligible, merged when redundant, and otherwise retained. This shows that low complexity does not simply imply removal, but instead determines how representation capacity should be compressed or preserved according to contribution and redundancy. This observation is consistent with Tab. 3, where splitting mainly affects prediction accuracy, while pruning and merging primarily regulate redundancy and efficiency.

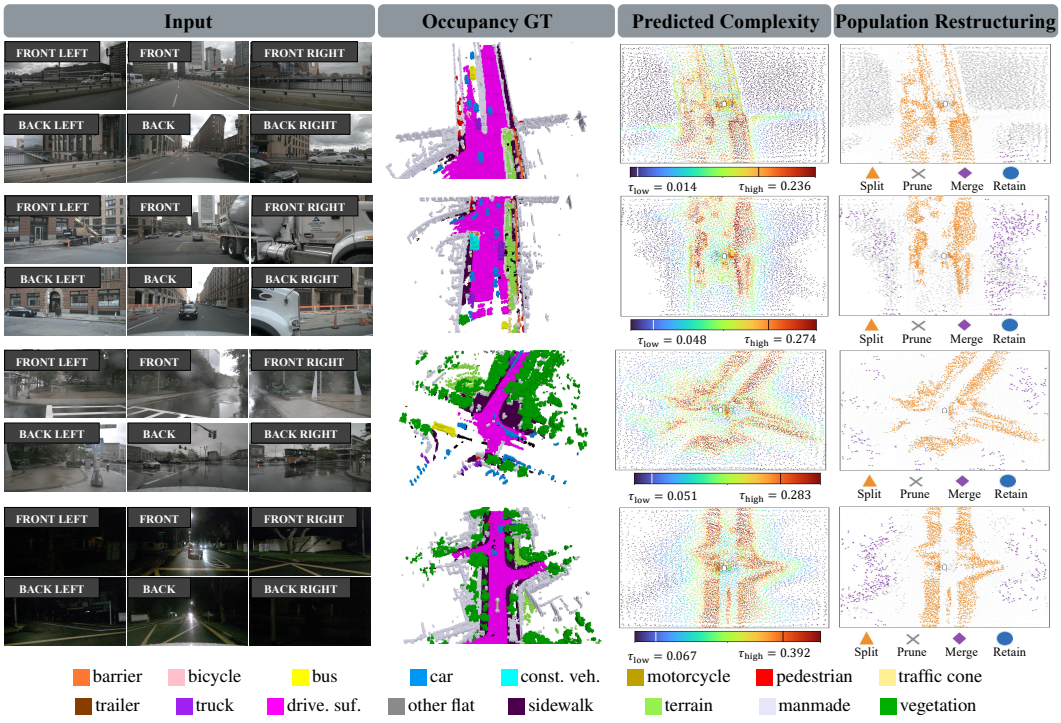

Figure 3: Qualitative analysis of complexity-adaptive Gaussian evolution. For each scene, we visualize the predicted Gaussian-wise complexity and the corresponding population restructuring operations. High-complexity regions preferentially trigger splitting for finer representation, whereas low-complexity primitives are pruned when negligible, merged when redundant, or otherwise retained. The examples illustrate that AdaGSFormer adapts the *spatial allocation* of Gaussian capacity according to scene structure, rather than uniformly increasing the global primitive count.

5 Conclusion and Limitations

We present **AdaGSFormer**, a complexity-adaptive Gaussian evolution framework in which local structural complexity drives population restructuring without enforcing a predefined final Gaussian count. By coupling complexity-guided splitting with contribution-aware pruning, moment-preserving merging, and optical-density decoding, AdaGSFormer adapts where representation capacity is expanded, compressed, or preserved. Experiments demonstrate that the proposed adaptive restructuring outperforms even a larger fixed-capacity counterpart while using fewer Gaussians on average, indicating that adaptive local capacity allocation provides benefits beyond simply increasing the global primitive count. Further analysis shows that high-complexity regions preferentially trigger local expansion, whereas redundant capacity is recovered through pruning and merging.

A current limitation is that structural complexity is supervised using local semantic heterogeneity, which does not capture all forms of geometric difficulty. Moreover, dynamic population control introduces additional decision overhead. Future work may therefore explore richer geometry-aware complexity cues and more efficient population restructuring strategies.

AI USE STATEMENT

In this work, we only used generative AI tools for language refinement, LaTeX formatting, and improving the presentation of figures and tables. Generative AI tools were not used for methodological design or experimental analysis. All AI-assisted content was carefully reviewed and verified by the authors. We take full responsibility for the final content of this work, including all text, claims, and artifacts produced with the assistance of generative AI.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide the implementation, configuration files, data-preparation instructions, and training and evaluation scripts in an anonymous repository at `https://github.com/SECodeSailor/AdaGSFormer`. The main paper specifies the dataset split, input resolution, backbone, Gaussian initialization, decoder configuration, optimization settings, and training hardware used in our experiments. Additional analyses of complexity–operation alignment, scene-relative criteria, and population behavior are provided in Appendix A, while qualitative occupancy and Gaussian-evolution visualizations are included in Appendix B.

REFERENCES

Yinlong Bai, Hongxin Zhang, Sheng Zhong, Junkai Niu, Hai Li, Yijia He, and Yi Zhou. Memgs: Memory-efficient gaussian splatting for real-time slam. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 11097–11103. IEEE, 2025.

Maxim Berman, Amal Rannen Triki, and Matthew B Blaschko. The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4413–4421, 2018.

Simon Boeder, Fabian Gigengack, and Benjamin Risse. Gaussianflowocc: Sparse and weakly supervised occupancy estimation using gaussian splatting and temporal flow. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 24943–24954, 2025.

Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11621–11631, 2020.

Loick Chambon, Eloi Zablocki, Alexandre Boulch, Mickael Chen, and Matthieu Cord. Gaussrender: Learning 3d occupancy with gaussian rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 27010–27020, 2025.

Zhiwen Fan, Kevin Wang, Kairun Wen, Zehao Zhu, Dejia Xu, and Zhangyang Wang. Lightgaussian: Unbounded 3d gaussian compression with 15x reduction and 200+ fps. *Advances in neural information processing systems*, 37:140138–140158, 2024.

Guangchi Fang and Bing Wang. Mini-splatting: Representing scenes with a constrained number of gaussians. In *European conference on computer vision*, pp. 165–181. Springer, 2024.

Wanshui Gan, Fang Liu, Hongbin Xu, Ningkai Mo, and Naoto Yokoya. Gaussianocc: Fully self-supervised and efficient 3d occupancy estimation with gaussian splatting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 28980–28990, 2025.

Alex Hanson, Allen Tu, Geng Lin, Vasu Singla, Matthias Zwicker, and Tom Goldstein. Speedy-splat: Fast 3d gaussian splatting with sparse pixels and sparse primitives. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 21537–21546, 2025.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Tri-perspective view for vision-based 3d semantic occupancy prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9223–9232, 2023.

Yuanhui Huang, Wenzhao Zheng, Borui Zhang, Jie Zhou, and Jiwen Lu. Selfocc: Self-supervised vision-based 3d occupancy prediction. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19946–19956. IEEE, 2024a.

Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Gaussianformer: Scene as gaussians for vision-based 3d semantic occupancy prediction. In *European Conference on Computer Vision*, pp. 376–393. Springer, 2024b.

Yuanhui Huang, Amonnut Thammatadatrakoon, Wenzhao Zheng, Yunpeng Zhang, Dalong Du, and Jiwen Lu. Gaussianformer-2: Probabilistic gaussian superposition for efficient 3d occupancy prediction. In *Proceedings of the computer vision and pattern recognition conference*, pp. 27477–27486, 2025.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.

Bernhard Kerbl, Andreas Meuleman, Georgios Kopanas, Michael Wimmer, Alexandre Lanvin, and George Drettakis. A hierarchical 3d gaussian representation for real-time rendering of very large datasets. *ACM Transactions On Graphics (TOG)*, 43(4):1–15, 2024.

Joo Chan Lee, Daniel Rho, Xiangyu Sun, Jong Hwan Ko, and Eunbyung Park. Compact 3d gaussian representation for radiance field. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21719–21728. IEEE, 2024.

Xinzhuo Li, Xianghui Pan, Jiayuan Du, Wei Wei, Liuyi Wang, Chengju Liu, and Qijun Chen. Gaussianseed: Hierarchical gaussian seeding for high-resolution 3d occupancy prediction. *arXiv preprint arXiv:2607.20071*, 2026.

Junhong Lin, Xianda Guo, Kangli Wang, Yuqi Ye, Xiaoyu Liang, Yanlun Peng, and Wei Gao. Vgocc: Learning visual-geometric gaussians for vision-centric 3d driving occupancy prediction. *arXiv preprint arXiv:2607.18078*, 2026.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 936–944. Ieee, 2017.

Yifei Liu, Zhihang Zhong, Yifan Zhan, Sheng Xu, and Xiao Sun. Maskgaussian: Adaptive 3d gaussian representation from probabilistic masks. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 681–690. IEEE, 2025.

Tao Lu, Mulin Yu, Linning Xu, Yuanbo Xiangli, Limin Wang, Dahua Lin, and Bo Dai. Scaffold-gs: Structured 3d gaussians for view-adaptive rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 20654–20664, 2024.

Qihang Ma, Xin Tan, Yanyun Qu, Lizhuang Ma, Zhizhong Zhang, and Yuan Xie. Cotr: Compact occupancy transformer for vision-based 3d occupancy prediction. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19936–19945. IEEE, 2024.

Simon Niedermayr, Josef Stumpfegger, and Rüdiger Westermann. Compressed 3d gaussian splatting for accelerated novel view synthesis. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10349–10358. IEEE, 2024.

Kanglin Ning, Yiran Zhao, Wenrui Li, Houde Quan, Qifan Li, Xingtao Wang, and Xiaopeng Fan. Learning adaptive semantic gaussian allocation for 3d occupancy. *arXiv preprint arXiv:2607.21896*, 2026.

Gyeongrok Oh, Sungjune Kim, Heeju Ko, Hyung-gun Chi, Jinkyu Kim, Dongwook Lee, Daehyun Ji, Sungjoon Choi, Sujin Jang, and Sangpil Kim. 3d occupancy prediction with low-resolution queries via prototype-aware view transformation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17134–17144. IEEE, 2025.

Jinhyung Park, Yihan Hu, Chensheng Peng, Wenzhao Zheng, Kris Kitani, and Wei Zhan. S2go: Streaming sparse gaussian occupancy prediction. *arXiv preprint arXiv:2506.05473*, 2025.

Rui Qian, Haozhi Cao, Tianchen Deng, Shenghai Yuan, and Lihua Xie. Splatssc: Decoupled depth-guided gaussian splatting for semantic scene completion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 8520–8528, 2026.

Qianpu Sun, Changyong Shu, Sifan Zhou, Runxi Cheng, Yongxian Wei, Zichen Yu, Dawei Yang, Sirui Han, and Yuan Chun. Gsrender: Deduplicated occupancy prediction via weakly supervised 3d gaussian splatting. *arXiv preprint arXiv:2412.14579*, 2024.

Pin Tang, Zhongdao Wang, Guoqing Wang, Jilai Zheng, Xiangxuan Ren, Bailan Feng, and Chao Ma. Sparseocc: Rethinking sparse latent representation for vision-based semantic occupancy prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15035–15044, 2024.

Haoran Wang, Guoxi Huang, Fan Zhang, David Bull, and Nantheera Anantrasirichai. Prune wisely, reconstruct sharply: Compact 3d gaussian splatting via adaptive pruning and difference-of-gaussian primitives. *arXiv preprint arXiv:2602.24136*, 2026.

Peihao Wang, Yuehao Wang, Dilin Wang, Sreyas Mohan, Zhiwen Fan, Lemeng Wu, Ruisi Cai, Yu-Ying Yeh, Zhangyang Wang, Qiang Liu, et al. Steepest descent density control for compact 3d gaussian splatting. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26663–26672. IEEE, 2025.

Tai Wang, Xinge Zhu, Jiangmiao Pang, and Dahua Lin. Fcos3d: Fully convolutional one-stage monocular 3d object detection. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 913–922. IEEE, 2021.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17850–17859, 2023.

Yi Wei, Linqing Zhao, Wenzhao Zheng, Zheng Zhu, Jie Zhou, and Jiwen Lu. Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21729–21740, 2023.

Chi Yan and Dan Xu. Progressive gaussian transformer with anisotropy-aware sampling for open vocabulary occupancy prediction. In *International Conference on Learning Representations*, volume 2026, pp. 4715–4737, 2026.

Xiaoyang Yan, Muleilan Pei, and Shaojie Shen. Vg3s: Visual geometry grounded gaussian splatting for semantic occupancy prediction. *arXiv preprint arXiv:2603.06210*, 2026.

Haosen Yang, Chenhao Zhang, Wenqing Wang, Marco Volino, Adrian Hilton, Li Zhang, and Xiatian Zhu. Improving gaussian splatting with localized points management. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21696–21705. IEEE, 2025.

Baijun Ye, Minghui Qin, Saining Zhang, Moonjun Gong, Shaoting Zhu, Hao Zhao, and Hang Zhao. Gs-occ3d: Scaling vision-only occupancy reconstruction with gaussian splatting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 25925–25937, 2025.

Zongxin Ye, Wenyu Li, Sidun Liu, Peng Qiao, and Yong Dou. Absgs: Recovering fine details in 3d gaussian splatting. In *Proceedings of the 32nd ACM international conference on multimedia*, pp. 1053–1061, 2024.

Yangming Zhang, Wenqi Jia, Wei Niu, and Miao Yin. Gaussianspa: An" optimizing-sparsifying" simplification framework for compact and high-quality 3d gaussian splatting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 26673–26682, 2025.

Yunpeng Zhang, Zheng Zhu, and Dalong Du. Occformer: Dual-path transformer for vision-based 3d semantic occupancy prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9433–9443, 2023.

Linqing Zhao, Xiuwei Xu, Ziwei Wang, Yunpeng Zhang, Borui Zhang, Wenzhao Zheng, Dalong Du, Jie Zhou, and Jiwen Lu. Lowrankocc: Tensor decomposition and low-rank recovery for vision-based 3d semantic occupancy prediction. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9806–9815. IEEE, 2024.

Changqing Zhou, Yueru Luo, and Changhao Chen. Generalizing visual geometry priors to sparse gaussian occupancy prediction. *arXiv preprint arXiv:2602.21552*, 2026.

Ziyue Zhu, Shenlong Wang, Jin Xie, Jiang-jiang Liu, Jingdong Wang, and Jian Yang. Voxelsplat: Dynamic gaussian splatting as an effective loss for occupancy and flow prediction. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 6761–6771, 2025.

Sicheng Zuo, Wenzhao Zheng, Yuanhui Huang, Jie Zhou, and Jiwen Lu. Gaussianworld: Gaussian world model for streaming 3d occupancy prediction. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 6772–6781, 2025.

Sicheng Zuo, Wenzhao Zheng, Xiaoyong Han, Longchao Yang, Jiwen Lu, et al. Quadricformer: Scene as superquadrics for 3d semantic occupancy prediction. *Advances in Neural Information Processing Systems*, 38:47779–47801, 2026.

A Additional Analysis of Complexity-Adaptive Evolution

A.1 Local Complexity–Operation Alignment

To directly examine whether the learned population behavior is consistent with local structural complexity, we post-hoc group valid Gaussian observations according to the GT-derived complexity proxy H_i. Importantly, H_i is used only for training supervision and analysis; during inference, candidate partitioning and all subsequent restructuring decisions are determined solely from the predicted score p_i.

As shown in Tab. 4, the splitting ratio increases sharply from 7.46% in low-complexity regions to 42.43% and 79.59% in the medium- and high-complexity groups, respectively. Conversely, the retention ratio decreases from 80.26% to 20.37%. Pruning and merging are predominantly concentrated in the low-complexity group, where redundant capacity is more likely to occur. These results provide direct local-level evidence that the learned complexity-guided restructuring is strongly aligned with the GT-derived structural-complexity proxy.

Table 4: Local complexity–operation alignment. Gaussians are post-hoc grouped by the GT-derived proxy H_i; operation assignments are produced using the predicted complexity score p_i during inference.

Complexity	H_i Range	Split↑	Prune	Merge	Retain
Low	$[0, 0.2)$	7.46%	9.66%	2.62%	80.26%
Medium	$[0.2, 0.4)$	42.43%	0.12%	0.10%	57.35%
High	$[0.4, 1]$	79.59%	0.03%	0.01%	20.37%

A.2 Effect of the Scene-Relative Criterion

We further analyze how the criterion used to interpret the predicted complexity scores affects population control. Fixed thresholds apply the same decision boundaries to all scenes and stages, whereas stage-global statistics adapt the boundaries across decoding stages but remain shared by all scenes. The proposed scene-relative criterion instead calibrates the boundaries using the score distribution of the current Gaussian population.

As shown in Tab. 5, fixed thresholds result in an average population of 17.1k Gaussians. Stage-global calibration reduces this to 15.7k, while the proposed scene-relative criterion further decreases it to 14.3k, a 16.4% reduction relative to fixed thresholds. Meanwhile, inference latency decreases from 224.74 ms to 205.90 ms, with mIoU slightly improving from 21.07 to 21.13. These results indicate that scene-relative calibration primarily suppresses unnecessary population growth while preserving semantic accuracy.

Table 5: Comparison of decision criteria for Gaussian population adaptation. The scene-relative criterion achieves comparable semantic accuracy while suppressing unnecessary population growth and reducing inference latency.

Criterion	mIoU ↑	Latency (ms) ↓	Avg. #G ↓
Fixed thresholds	21.07	224.74	17.1k
Stage-global statistics	21.11	223.77	15.7k
Scene-relative criterion	**21.13**	**205.90**	**14.3k**

A.3 Population Behavior across Scene Complexity Levels

For scene-level analysis, we aggregate the GT-derived local complexity targets within each validation scene and partition the validation set into three equally sized subsets. This grouping is used only for post-hoc analysis and does not affect model inference.

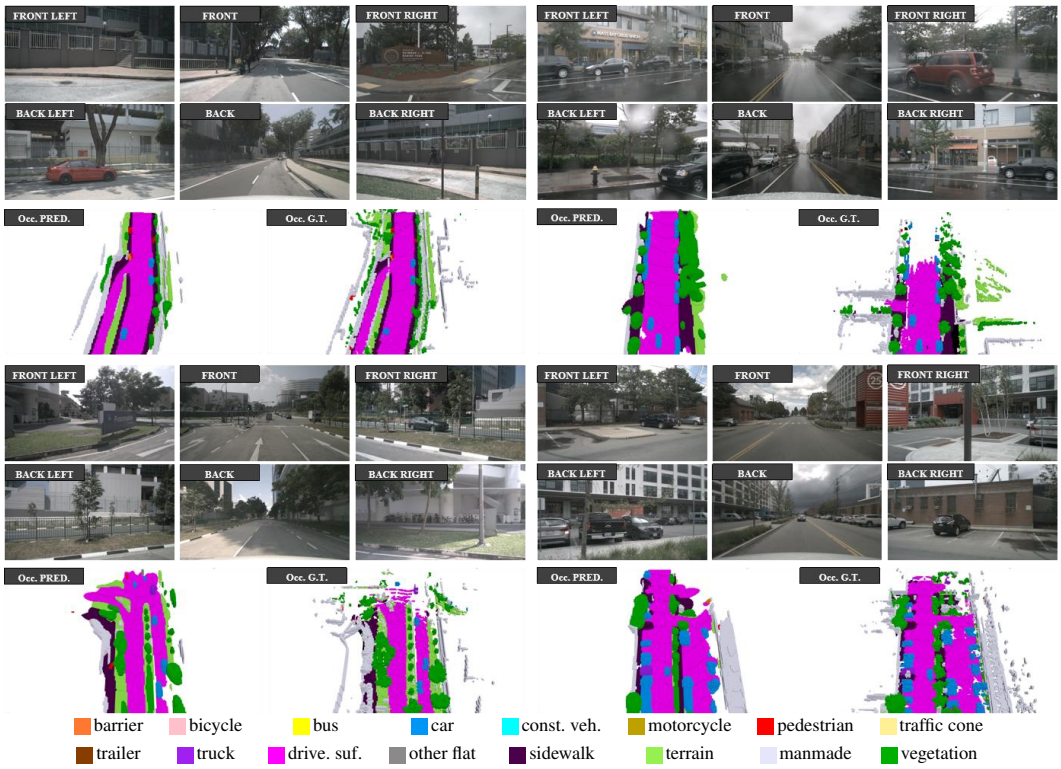

Figure 4: Qualitative semantic occupancy prediction on SurroundOcc. Each example shows six input camera views, the AdaGSFormer prediction, and the corresponding occupancy ground truth under the same visualization setting.

Table 6 reveals that higher-complexity scenes exhibit a modest increase in splitting operations, from 6.75k in the low-complexity subset to 6.99k in the high-complexity subset. Nevertheless, the final Gaussian cardinality remains within a narrow range of 14.19k–14.44k across all groups. This observation indicates that AdaGSFormer does not adapt by indiscriminately assigning more primitives to an entire complex scene; instead, local expansion is counterbalanced by pruning and merging elsewhere.

The accuracy trend further supports this interpretation. Compared with the larger fixed-capacity baseline, AdaGSFormer improves mIoU by 0.52, 0.38, and 0.80 points on the low-, medium-, and high-complexity subsets, respectively. The largest gain on high-complexity scenes suggests that adaptive local restructuring becomes particularly valuable when the scene contains stronger structural heterogeneity.

Table 6: Population restructuring across scene-complexity levels. Validation scenes are partitioned into low-, medium-, and high-complexity subsets using an analysis-only GT-derived semantic-complexity measure.

Scene Complexity	Gaussian Evolution Dynamics				Semantic Occupancy (mIoU↑)		
	Split	Prune	Merge	Final	Fixed		AdaGSFormer
					10k	16.3k	
Low	6.75k	2.11k	0.35k	14.30k	17.54	17.93	**18.45**
Medium	6.89k	2.41k	0.29k	14.19k	21.42	21.62	**22.00**
High	6.99k	2.26k	0.29k	14.44k	20.56	21.05	**21.85**
Overall	6.88k	2.26k	0.31k	14.31k	20.20	20.52	**21.13**

B Additional Qualitative Results

B.1 Semantic Occupancy Prediction

Fig. 4 presents additional qualitative results on SurroundOcc across diverse road layouts, including straight roads, intersections, and scenes with dense roadside structures. AdaGSFormer recovers the dominant drivable regions and large-scale semantic layout consistently, while retaining localized structures such as sidewalks, vegetation, and foreground objects. The remaining errors are primarily observed around thin structures and distant regions, where visual evidence from the input views is inherently limited.

B.2 Progressive Gaussian Evolution

Fig. 5 visualizes how the Gaussian representation evolves through successive decoding stages. The structured initialization $\mathcal{G}^0$ provides uniform coarse coverage without encoding scene-specific density. As visual evidence is progressively integrated, Gaussian primitives become increasingly concentrated around semantically informative scene structures, while redundant support is compressed through population restructuring. The final population $\mathcal{G}^K$ therefore exhibits substantially clearer spatial organization than the initialization and qualitatively agrees with the occupancy layout, illustrating how adaptive evolution transforms a scene-agnostic scaffold into a scene-conditioned sparse representation.

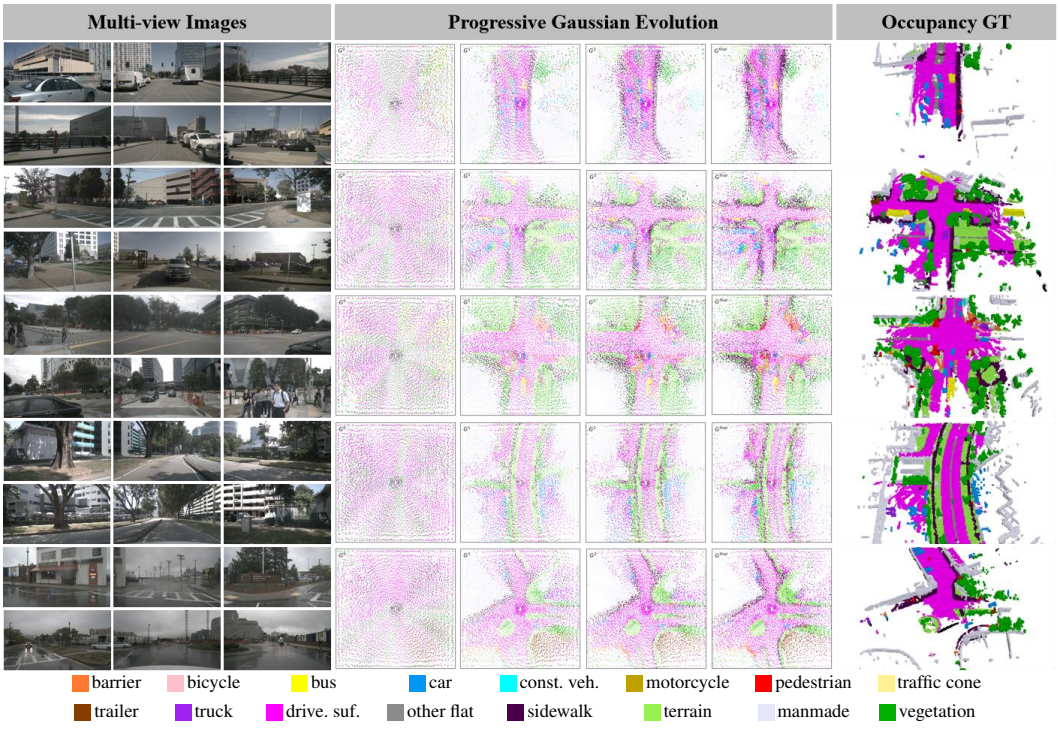

Figure 5: Visualization of progressive Gaussian evolution. Each row shows the multi-view input, Gaussian representations from $\mathcal{G}^0$ to $\mathcal{G}^K$, and the corresponding occupancy ground truth. Starting from a scene-agnostic structured scaffold, the Gaussian population progressively develops scene-aligned spatial and semantic organization through complexity-adaptive restructuring.