VGSSC: SURFACE-AWARE VOXEL–GAUSSIAN REPRESENTATIONS FOR MONOCULAR SEMANTIC SCENE COMPLETION

Anonymous authors
Paper under double-blind review

ABSTRACT

Monocular semantic scene completion seeks to infer a complete semantic occupancy field from a single RGB observation, requiring precise reconstruction of observed surfaces and reliable completion of unseen scene regions. This challenge becomes particularly pronounced in cluttered indoor scenes, where local geometry is complex and large portions of the scene remain unobserved. Existing methods typically represent scenes using either voxels or Gaussian primitives. Voxels preserve location-specific information throughout the volume but often produce spatially fragmented predictions, while Gaussians promote local geometric continuity yet may blur semantic boundaries when primitive supports overlap. We introduce VGSSC, a surface-aware Voxel–Gaussian framework that combines geometry-guided feature sharing with location-specific volumetric reasoning. Specifically, we introduce surface-relative distance tokens to encode the signed depth offset of each voxel from the predicted surface, providing explicit geometric cues that distinguish free space, near-surface regions, and occluded space. These surface-relative cues are incorporated into voxel features and also guide the initialization of surface-anchored Gaussian primitives. A Gaussian-assisted voxel decoder subsequently transfers locally aggregated geometric and semantic information from the Gaussian representation back to the dense volumetric, enabling volumetric completion beyond directly observed regions. Experiments on Occ-ScanNet and EmbodiedOcc-ScanNet demonstrate that VGSSC achieves state-of-the-art performance, including a 5.28% mIoU improvement over the previous method.

1 INTRODUCTION

Monocular semantic scene completion (SSC) constructs a dense 3D scene representation, jointly encoding occupancy, semantics, and free space, from a single RGB view despite the substantial geometric ambiguity inherent in partial 2D observations. Such a representation can support downstream tasks including spatial reasoning Xu et al. (2025); Wang et al. (2025b); Qi et al. (2026), navigation Li et al. (2025b); Zheng et al. (2024), and embodied interaction Liu et al. (2024); Li et al. (2024). The central difficulty is that a single image directly observes only the visible surfaces of a scene, whereas SSC requires predicting the geometry and semantics of both observed and occluded regions.

Much of the recent progress in SSC has been driven by outdoor driving scenarios, where scene layouts are relatively structured and objects are often spatially separated. Indoor environments present a different set of challenges: objects are densely arranged, mutual occlusions are frequent, and neighboring instances often exhibit weak visual boundaries. Accurate indoor completion therefore requires both spatially coherent geometry and sufficiently discriminative features to preserve semantic boundaries between adjacent objects. Existing methods mainly build on voxel- or Gaussian-based scene representations. Voxel-based approaches Yu et al. (2024); Wang et al. (2026) maintain dense features at fixed spatial locations, providing direct support for volumetric completion. However, under severe occlusion, interactions among independently represented voxels can be difficult to establish reliably, which may lead to locally fragmented predictions. Gaussian-based approaches Wu et al. (2025); Zhou et al. (2026) instead represent local geometry with continuous primitives whose positions, scales, and orientations adapt to scene structure. Their spatial extent naturally propagates information across nearby regions and encourages local geometric coherence, but overlap-

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.

ping Gaussians may also mix features from different semantic regions and blur boundaries between neighboring objects. These complementary properties motivate a representation that combines the dense spatial coverage of voxels with the adaptive local geometry of Gaussians while controlling how information is exchanged between them.

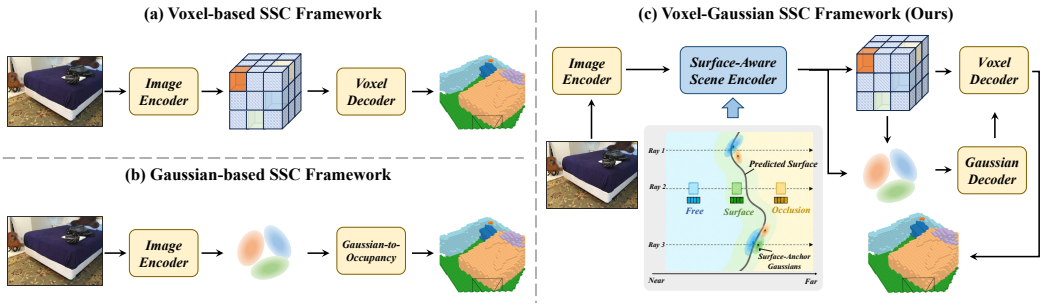

Figure 1: **Comparison of scene representation paradigms for semantic scene completion.** (a) Voxel-based methods provide dense volumetric support at fixed spatial locations. (b) Gaussian-based methods capture local geometry with adaptive spatial support. (c) Our VGSSC organizes both representations around a shared predicted surface: surface-relative distance tokens encode voxel positions relative to the surface, while surface-anchored Gaussians capture local geometry and transfer the refined information back to the dense voxel field for completion.

These limitations motivate a representation in which local geometric information can be shared without discarding the spatial specificity required for volumetric completion. We use the predicted surface as a common geometric reference for both voxels and Gaussians. The surface determines where Gaussian primitives are initialized and, more importantly, provides each voxel with an explicit description of its position relative to the observed geometry. This distinction is critical during image-to-voxel lifting: voxels along the same camera ray sample the same image feature even though they may lie in front of, near, or behind the visible surface. Appearance features alone therefore provide no direct cue for distinguishing these locations.

Based on this observation, we propose VGSSC, a monocular semantic scene completion framework that couples surface-aware voxel reasoning with geometry-guided Gaussian modeling. During image-to-voxel lifting, we encode the signed optical-axis offset between each voxel and the predicted surface using surface-relative distance tokens. Their depth intervals are controlled by an image-adaptive learnable scale, providing finer positional resolution near the surface while retaining coverage farther away. The resulting voxel features distinguish locations that share the same projected appearance but occupy different geometric positions. We then initialize Gaussians at predicted surface locations using both image features and the corresponding voxel context, allowing each primitive to encode local appearance, geometry, and surrounding volumetric information. After contextual refinement, the Gaussians transfer their features and semantic predictions back to the voxel field through geometry- and opacity-weighted aggregation. The fused voxel features are subsequently refined across multiple stages, propagating local surface information into occluded regions while preserving dense volumetric support for scene completion.

Our contributions are summarized as follows.

- We propose VGSSC, a **surface-aware Voxel–Gaussian framework** that couples adaptive surface modeling with dense volumetric reasoning through explicit interactions between surface-anchored Gaussians and voxel representations.
- We introduce **surface-relative distance tokens** that encode signed depth offsets from the predicted surface, enabling voxels along the same camera ray to retain distinct surface-relative geometric cues throughout the volume.
- Experiments on Occ-ScanNet and EmbodiedOcc-ScanNet datasets demonstrate that VGSSC achieves state-of-the-art performance, improving mIoU by 5.28% over the previous method.

2 Related Work

2.1 Scene Representations for Semantic Completion

Voxel-based methods such as MonoScene, ISO, Symphonies, and AdaSFormer (Cao & De Charette, 2022; Yu et al., 2024; Jiang et al., 2024; Wang et al., 2026) lift image features into dense voxel grids and model spatial context through volumetric reasoning, instance-level interactions, or attention. Gaussian-based methods instead represent scenes with adaptive spatial primitives. GaussianFormer and EmbodiedOcc (Huang et al., 2024; Wu et al., 2025) use semantic Gaussians for occupancy prediction in driving and indoor environments, while SplatSSC and GPOcc (Qian et al., 2026; Zhou et al., 2026) further incorporate predicted depth or visual geometry priors to guide Gaussian construction.

Hybrid Voxel–Gaussian representations remain less explored. Most closely related to our work, GaussianSSC (Xian et al., 2026) introduces Gaussian support into a grid-based framework, using Gaussian-weighted image aggregation for image-to-voxel association and a learned per-voxel Gaussian field for triplane refinement. Its design is developed for outdoor driving scenes and does not explicitly encode the signed geometric relation between a voxel and the visible surface. This relation is particularly informative in cluttered indoor scenes, where voxels projected to the same image location may lie in free space, near the observed surface, or within occluded regions. Our VGSSC therefore uses the predicted surface as a shared geometric reference: surface-anchored Gaussians model adaptive local geometry, while signed voxel-to-surface depth offsets provide location-specific geometric cues for dense volumetric reasoning.

2.2 Surface-Aware Geometric Encoding

Early depth-based semantic completion methods explicitly encode geometry relative to observed surfaces. SSCNet (Song et al., 2017) converts a single depth image into a flipped truncated signed distance field, exposing surface proximity to subsequent volumetric reasoning. For vision-based SSC, SOAP and VisHall3D (Lee et al., 2025; Lu et al., 2025) incorporate occlusion structure through visibility-aware projection and explicit separation of visible reconstruction from invisible completion, respectively. SurfOcc and MonoMRN (Ye et al., 2024; Wang et al., 2025d) further exploit surface geometry by propagating surface features into occluded regions or weighting image-to-voxel projection according to surface distance.

Related depth representations have also been studied for image-to-3D lifting and monocular depth estimation. LSS and CaDDN (Philion & Fidler, 2020; Reading et al., 2021) represent depth using categorical distributions during feature lifting, while AdaBins and LocalBins (Bhat et al., 2021; 2022) learn adaptive depth intervals to allocate representation capacity according to scene depth structure. Building on these ideas, VGSSC encodes the signed optical-axis depth offset between each voxel and the predicted surface using learnable surface-relative distance tokens. This encoding preserves whether a voxel lies in front of or behind the surface while providing finer positional resolution near it, allowing voxels with identical projected appearance to retain distinct geometric cues.

3 Method

3.1 Problem Formulation and Overview

Problem Formulation. Given a monocular RGB image $I \in \mathbb{R}^{H\times W\times 3}$, camera intrinsics $\mathbf{K}_{\text{cam}}$, monocular SSC predicts a semantic occupancy volume $\widehat{Y} \in \{0,\dots,N_{\text{cls}}-1\}^{N_x\times N_y\times N_z}$, where label 0 denotes free space and the remaining labels represent semantic classes.

Method Overview. As shown in Fig. 2, we introduce VGSSC, a surface-aware Voxel–Gaussian framework that couples adaptive local surface modeling with dense volumetric reasoning through a shared predicted surface. Given a monocular image I, we first extract image features $\mathbf{F}_{2D}$ and predict a depth map D. The predicted depth defines a geometric reference for both representations: it provides signed surface-relative offsets for voxel lifting and 3D surface anchors for Gaussian initialization. We augment the lifted image features with learnable surface-relative distance tokens to

obtain surface-aware voxel features $\mathbf{F}_{3D}^{V}$. In parallel, back-projected surface anchors, image features, and voxel context initialize a set of surface-anchored Gaussians $\mathcal{G}$, which are further refined through scene-level context aggregation to produce $\mathcal{G}'$. The refined Gaussians are then converted into voxel-aligned features $\mathbf{F}_{3D}^{G}$ and semantic occupancy cues $\mathbf{M}_{3D}^{G}$. Finally, the Gaussian-derived information is fused with $\mathbf{F}_{3D}^{V}$, and a shared iterative voxel decoder propagates the fused representation throughout the target volume to produce the final semantic occupancy prediction $\widehat{Y}$.

In the following, Sec. 3.2 introduces the surface-aware scene encoder, Sec. 3.3 describes the surface-anchored Gaussian enhancer, and Sec. 3.4 presents the Gaussian-assisted voxel decoder together with the training objective. Finally, Sec. 3.5 extends the trained single-frame model to online prediction through training-free probability fusion.

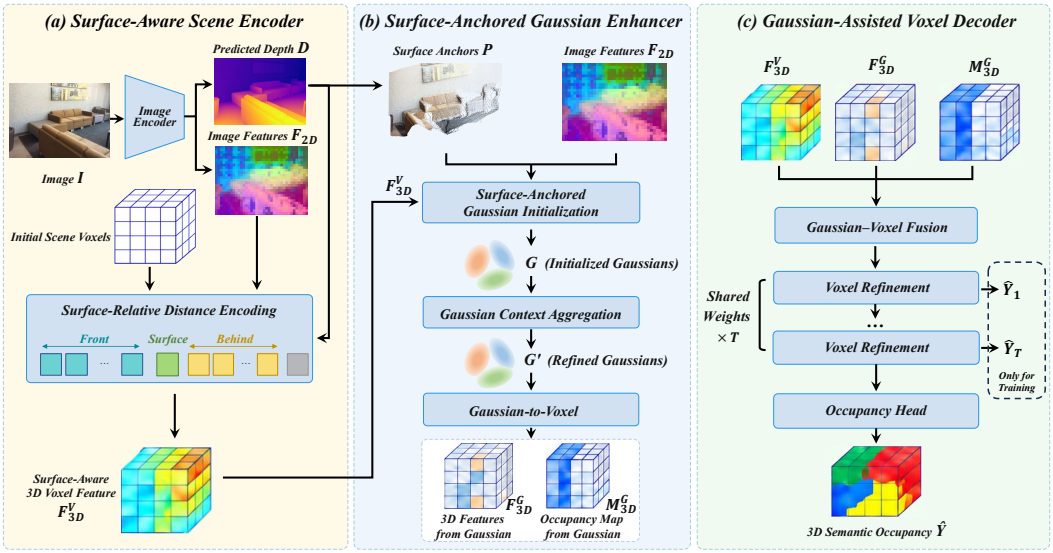

Figure 2: **Overall framework of VGSSC.** (a) The *Surface-Aware Scene Encoder* augments lifted image features with surface-relative distance tokens to construct $\mathbf{F}_{3D}^{V}$. (b) The *Surface-Anchored Gaussian Enhancer* initializes and contextually refines surface-anchored Gaussians, which are converted into voxel-aligned features $\mathbf{F}_{3D}^{G}$ and semantic occupancy cues $\mathbf{M}_{3D}^{G}$. (c) The *Gaussian-Assisted Voxel Decoder* fuses the two representations and iteratively refines the dense voxel field using shared weights. Dashed branches denote intermediate predictions used only for training.

3.2 Surface-Aware Scene Encoder

Surface-Relative Geometry. For voxels along the same viewing ray, their signed displacement from the visible surface provides the geometric information that image appearance alone cannot encode. For each voxel v, we transform its center $\mathbf{x}_v$ into the camera coordinate system, project it onto the image plane, and compute its optical-axis depth offset from the predicted surface:

$$\mathbf{x}_v^{\text{cam}} = T_{\text{cam}\leftarrow\text{grid}}(\mathbf{x}_v), \qquad \mathbf{p}_v = \text{proj}(\mathbf{K}_{\text{cam}}\mathbf{x}_v^{\text{cam}}), \qquad d_v = z_v - D_z(\mathbf{p}_v). \tag{1}$$

Here, $D_z(\mathbf{p}_v)$ denotes the predicted surface depth along the optical axis. Accordingly, $d_v < 0$, $d_v = 0$, and $d_v > 0$ identify locations in front of, on, and behind the predicted surface, respectively.

Surface-Relative Distance Encoding. Figure 3 illustrates how the surface-relative distance resolves the ray-wise feature ambiguity. As shown in Fig. 3(a), voxels projected to the same image location receive the same appearance feature despite lying in front of, on, or behind the predicted surface. We therefore encode the signed offset d_v into a learnable distance token and inject it into the lifted image feature, as illustrated in Fig. 3(b).

To distinguish the two sides of the surface while retaining higher resolution near the surface boundary, we predict an image-adaptive scale and discretize each signed offset as

$$S = \text{softplus}\big(\Phi_{\text{scale}}(\mathbf{F}_{2D})\big), \qquad u_v = \sigma(d_v/S), \qquad k_v = \text{clip}\big(\lfloor K u_v \rfloor, 0, K-1\big). \tag{2}$$

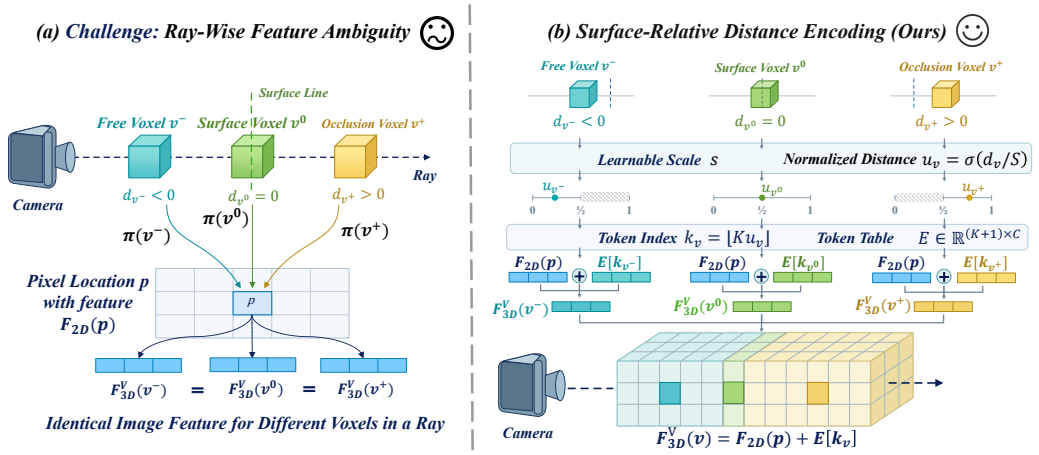

Figure 3: **Surface-relative distance encoding.** (a) Voxels projected to the same image location receive identical appearance features despite occupying different positions relative to the predicted surface. (b) We map each signed voxel-to-surface depth offset to a normalized value using a learnable scale and use it to select a learnable distance token. Adding the selected token to the image feature provides distinct surface-relative geometric cues for voxels along the same camera ray. An additional token handles invalid projections or unavailable surface depths.

Here, Φ_{scale} predicts an image-dependent scalar $S > 0$, shared across all voxels of the same image. The sigmoid centers the predicted surface at $u_v = 0.5$, with negative and positive offsets mapped to opposite sides. Uniform discretization in the normalized space therefore induces finer depth intervals near the surface and progressively coarser intervals farther away (Bhat et al., 2022).

A learnable embedding table $\mathbf{E} \in \mathbb{R}^{(K+1)\times C}$ stores K surface-relative distance tokens, together with one additional token for invalid projections or unavailable surface depths. For each valid voxel, we retrieve the corresponding token and add it to the image feature sampled at its projected location:

$$\mathbf{F}^V_{3D}(v) = \mathbf{F}_{2D}(\mathbf{p}_v) + \mathbf{E}[k_v]. \tag{3}$$

Consequently, voxels along the same camera ray retain their shared appearance feature while receiving distinct embeddings according to their positions relative to the predicted surface. Applying this operation over the target grid produces the surface-aware voxel field $\mathbf{F}^V_{3D}$, which provides location-specific volumetric features for subsequent Gaussian initialization and voxel prediction. During training, we use a straight-through Gumbel–Softmax relaxation (Jang et al., 2017) for differentiable token assignment, while inference directly uses the deterministic index k_v in Eq. 2.

3.3 Surface-Anchored Gaussian Enhancer

Surface-Anchored Gaussian Initialization. Anchoring Gaussian primitives to the predicted surface concentrates their support on geometrically relevant regions and avoids allocating primitives uniformly throughout the 3D volume. We back-project selected valid image locations using the predicted depth and camera geometry to obtain surface anchors $\mathbf{P} = \{\boldsymbol{\mu}_i\}_{i=1}^{N_G}$, with each anchor initializing one Gaussian. Each primitive contains a center $\boldsymbol{\mu}_i$, anisotropic scale $\mathbf{s}_i$, rotation quaternion $\mathbf{q}_i$, opacity o_i, feature $\mathbf{h}_i$, and semantic scores $\mathbf{c}_i$. Its feature is initialized jointly from image appearance and the surface-aware voxel field:

$$\mathbf{h}_i = \mathbf{W}_f \mathbf{F}_{2D}(\mathbf{p}_i) + \text{Sample3D}\left(\mathbf{F}^V_{3D}, \boldsymbol{\mu}_i\right). \tag{4}$$

Here, $\mathbf{p}_i$ denotes the image location of the anchor, $\mathbf{W}_f$ aligns the feature dimensions, and Sample3D denotes trilinear interpolation. A prediction head maps $\mathbf{h}_i$ to the remaining non-positional Gaussian attributes.

Gaussian Context Aggregation. Each initialized Gaussian describes only a local neighborhood around its anchor, whereas semantic completion benefits from interactions across distant scene re-

gions. We therefore refine the Gaussian set with scene-level context:

$$\mathcal{G}' = \Phi_{\mathrm{G}}(\mathcal{G}). \tag{5}$$

Specifically, attribute-augmented Gaussian features are first aggregated into learned latent tokens through attention, after which the Gaussians attend back to these tokens to receive global context. Residual prediction heads then update the Gaussian attributes, including their centers, while enforcing positive scales and unit-norm rotation quaternions.

Gaussian-to-Voxel Conversion. The refined Gaussians encode local surface geometry and semantics with adaptive spatial support, but semantic scene completion ultimately requires predictions over the full voxel grid. We therefore rasterize the Gaussian representation into voxel-aligned feature and semantic fields:

$$\left(\mathbf{F}_{3D}^{G}, \mathbf{M}_{3D}^{G}\right) = \Phi_{\mathrm{G2V}}(\mathcal{G}'). \tag{6}$$

Here, $\mathbf{F}_{3D}^{G}$ and $\mathbf{M}_{3D}^{G}$ denote the Gaussian-derived voxel features and semantic occupancy cues, respectively. Each Gaussian contributes to the neighborhood around its nearest voxel, with contributions weighted by opacity and the Gaussian kernel evaluated from the corresponding squared spatial distance. This conversion preserves the local geometric support of the Gaussians while making their information directly compatible with dense voxel reasoning.

3.4 Gaussian-Assisted Voxel Decoder

Gaussian–Voxel Fusion. The two representations provide complementary information at each voxel: $\mathbf{F}_{3D}^{V}$ preserves dense, location-specific context throughout the volume, while $\mathbf{F}_{3D}^{G}$ and $\mathbf{M}_{3D}^{G}$ inject locally aggregated geometric and semantic information around Gaussian-supported regions. We combine them with a voxel-wise fusion operator:

$$\mathbf{V} = \Phi_{\mathrm{fuse}}\left(\mathbf{F}_{3D}^{V}, \mathbf{F}_{3D}^{G}, \mathbf{M}_{3D}^{G}\right). \tag{7}$$

The resulting feature field $\mathbf{V}$ retains dense spatial support while incorporating the adaptive local structure encoded by the Gaussian branch.

Voxel Refinement and Prediction. Gaussian information is concentrated near regions supported by the predicted surface, whereas completion also requires propagating context into unobserved parts of the volume. After the Gaussian–voxel fusion, we therefore apply a shared refinement operator recurrently over the dense voxel field and decode a semantic occupancy prediction at each stage:

$$\mathbf{V}_t = \Phi_{\mathrm{ref}}(\mathbf{V}_{t-1}), \qquad \widehat{Y}_t = \Phi_{\mathrm{dec}}(\mathbf{V}_t), \qquad t = 1, \ldots, T. \tag{8}$$

Here, $\mathbf{V}_0 = \mathbf{V}$, Φ_{ref} shares parameters across all refinement stages, and Φ_{dec} denotes the semantic occupancy head. Intermediate predictions provide auxiliary supervision during training, while inference uses only the final prediction $\widehat{Y} = \widehat{Y}_T$.

Training Objective. Following GPOcc (Zhou et al., 2026), we jointly optimize the image encoder, depth head, and completion network with final occupancy, auxiliary occupancy, and depth supervision:

$$\mathcal{L} = \mathcal{L}_{\mathrm{occ}}^{\mathrm{final}} + \lambda_{\mathrm{aux}}\mathcal{L}_{\mathrm{occ}}^{\mathrm{aux}} + \lambda_{\mathrm{depth}}\mathcal{L}_{\mathrm{depth}}. \tag{9}$$

The final occupancy loss supervises the prediction from $\mathbf{V}_T$, the auxiliary occupancy loss is applied to predictions from the intermediate states $\mathbf{V}_t$, and the depth loss supervises the surface prediction used by both the voxel and Gaussian branches.

3.5 Online Prediction

Following the online SSC setting of EmbodiedOcc (Wu et al., 2025), we extend the trained single-frame model to posed RGB sequences without introducing additional learnable parameters.

For frame n, the single-frame model produces a local semantic occupancy prediction $\widehat{Y}^n$ in its camera-centered voxel grid. Let $\mathbf{P}^n$ denote the corresponding class-probability volume. Using the known camera pose, we align the local coordinate system of frame n with global coordinate

system and trilinearly sample $\mathbf{P}^n$ at the transformed location. This yields the world-aligned class-probability vector $\mathbf{p}^{\mathrm{world},n}(v)$, which provides the observation of voxel v from the current frame.

We progressively fuse these frame-wise observations in the global voxel grid by maintaining their weighted probability sum and accumulated observation weight:

$$\mathbf{A}_n(v) = \mathbf{A}_{n-1}(v) + \omega^n(v)\mathbf{p}^{\mathrm{world},n}(v), \quad B_n(v) = B_{n-1}(v) + \omega^n(v), \quad \mathbf{p}_n^{\mathrm{global}}(v) = \frac{\mathbf{A}_n(v)}{B_n(v)}. \tag{10}$$

Here, $\mathbf{A}_n(v)$ accumulates the weighted class probabilities from all valid observations up to frame n, while $B_n(v)$ records their total observation weight. The fused distribution $\mathbf{p}_n^{\mathrm{global}}(v)$ is evaluated only when $B_n(v) > 0$. The observation weight $\omega^n(v) \geq 0$ is determined by prediction confidence and viewing geometry, with free-space predictions and voxels behind the predicted surface receiving lower weights. Invalid observations are assigned zero weight and therefore leave the accumulated global statistics unchanged. After processing frame n, the global semantic prediction is obtained as $\widehat{C}_n^{\mathrm{global}}(v) = \arg\max_c \mathbf{p}_n^{\mathrm{global}}(v)_c$.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate single-frame monocular semantic scene completion on Occ-ScanNet (Yu et al., 2024) and multi-frame online prediction on EmbodiedOcc-ScanNet (Wu et al., 2025). The **Occ-ScanNet** dataset contains 45,755 training samples and 19,764 test samples, each of which is represented by a $60 \times 60 \times 36$ voxel grid with a resolution of 8 cm, corresponding to a 4.8m $\times$ 4.8m $\times$ 2.88m volume in front of the camera. The **EmbodiedOcc-ScanNet** dataset is derived from Occ-ScanNet and comprises 537 training scenes and 137 validation scenes. Each scene consists of 30 posed frames with their corresponding local occupancies. The resolutions of global occupancy of each scene are calculated by $l_x \times l_y \times l_z/(0.08m)^3$, where $l_x \times l_y \times l_z$ is the range of this scene in the world coordinate system. In both datasets, each voxel is annotated with 12 semantic classes, including 11 valid categories and one class for empty space. Following prior work, we report intersection over union (IoU) and mean IoU (mIoU) as evaluation metrics.

Implementation Details. We instantiate the image backbone from Depth Anything V2 (Yang et al., 2024). Training epochs are set to 10 epochs with AdamW, using a learning rate of 2×10^{-4}, a weight decay of 0.01, and a batch size of 2 per GPU. The learning rate of the backbone is scaled by 0.1. We use $K = 64$ for surface-relative distance encoding and $T = 2$ for voxel refinement modules.

4.2 Comparison with State-of-the-Art Methods

Monocular semantic scene completion. Table 1 compares VGSSC with existing methods on Occ-ScanNet. VGSSC achieves 66.26% IoU and 57.16% mIoU, outperforming all methods listed in the table on both metrics. Compared with GPOcc, it improves IoU by 9.30% and mIoU by 5.28%. It also surpasses SplatSSC, the strongest competing method in geometric IoU, by 3.43%. The semantic gains are consistent across categories, with VGSSC obtaining the highest IoU for all 11 classes. These results support the effectiveness of combining local Gaussian geometry with dense voxel reasoning for monocular scene completion. Figure 4 shows the qualitative results.

Online semantic scene completion. Table 2 compares VGSSC with existing methods on EmbodiedOcc-ScanNet. VGSSC achieves 59.73% IoU and 53.17% mIoU, yielding the highest performance among the compared methods. Relative to GPOcc, it achieves gains of +3.34% IoU and +1.95% mIoU. These results demonstrate that VGSSC can support online scene completion through prediction-level fusion. Figure 5 shows the qualitative results.

4.3 Ablation Study

Effect of VGSSC Main Components. Table 3 evaluates the main components of VGSSC. Removing surface-relative distance encoding reduces mIoU from 57.16% to 55.40%, while removing

Table 1: Monocular prediction performance on Occ-ScanNet dataset. **Bold** and Underlined denote the best and the second-best result.

Method	Venue	IoU↑	ceiling	floor	wall	window	chair	bed	sofa	table	tvs	furniture	objects	mIoU↑
Voxel-based														
MonoScene Cao & De Charette (2022)	CVPR'22	41.60	15.17	44.71	22.41	12.55	26.11	27.03	35.91	28.32	6.57	32.16	19.84	24.62
TPVFormer Huang et al. (2023)	CVPR'23	33.39	6.96	32.97	14.41	9.10	24.01	41.49	45.44	28.61	10.66	35.37	25.31	24.94
SurroundOcc Wei et al. (2023)	ICCV'23	42.52	18.90	49.30	24.80	18.00	26.80	42.00	44.10	32.90	18.60	36.80	26.90	30.83
ISO Yu et al. (2024)	ECCV'24	42.16	19.88	41.88	22.37	16.98	29.09	42.43	42.00	29.60	10.62	36.36	24.61	28.71
GA-MonoSSC Li et al. (2025a)	ICCV'25	48.59	26.01	51.89	30.35	24.58	30.34	48.44	49.92	37.59	22.48	43.13	27.41	35.65
AdaSFormer Wang et al. (2026)	CVPR'26	54.63	37.66	57.34	40.51	29.46	43.29	60.41	63.08	47.05	29.63	54.08	36.15	45.33
Gaussian-based														
GaussianFormer Huang et al. (2024)	ECCV'24	40.91	20.70	42.00	23.40	17.40	27.00	44.30	44.80	32.70	15.30	36.70	25.00	29.93
EmbodiedOcc Wu et al. (2025)	ICCV'25	53.55	39.60	50.40	41.40	31.70	40.90	55.00	61.40	44.00	36.10	53.90	42.20	45.15
EmbodiedOcc++ Wang et al. (2025a)	ACM MM'25	54.90	36.40	53.10	41.80	34.40	42.90	57.30	64.10	45.20	34.80	54.20	44.10	46.20
RoboOcc Zhang et al. (2025)	arXiv'25	56.48	45.36	53.49	44.35	34.81	43.38	56.93	63.35	46.35	36.12	55.48	44.78	47.67
SplatSSC Qian et al. (2026)	AAAI'26	62.83	49.10	59.00	48.30	38.80	47.40	62.40	67.00	49.50	42.60	60.70	45.40	51.83
GPOcc Zhou et al. (2026)	CVPR'26	56.96	51.42	50.35	46.97	41.84	46.98	60.39	66.16	50.51	47.97	58.88	49.23	51.88
Voxel-Gaussian hybrid														
VGSSC	-	**66.26**	**55.39**	**66.93**	**51.52**	**44.97**	**50.88**	**66.24**	**71.34**	**54.60**	**50.39**	**63.78**	**52.77**	**57.16**

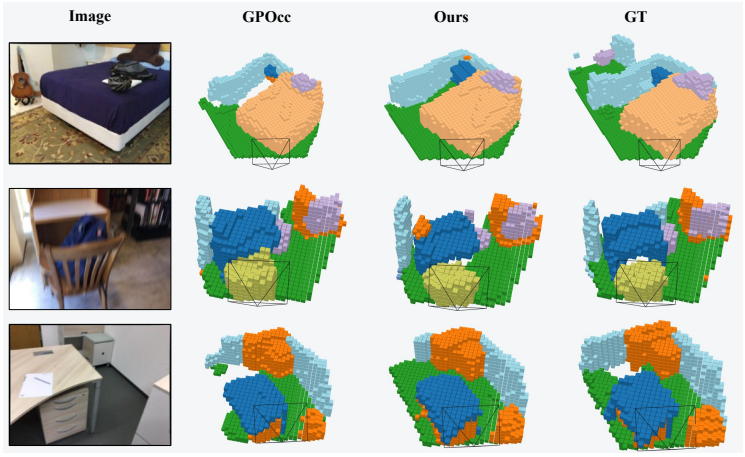

Figure 4: Qualitative comparison on Occ-ScanNet.

Table 2: Embodied prediction performance on the EmbodiedOcc-ScanNet dataset. **Bold** and Underlined denote the best and the second-best result.

Method	Venue	IoU↑	ceiling	floor	wall	window	chair	bed	sofa	table	tvs	furniture	objects	mIoU↑
Voxel-based														
TPVFormer Huang et al. (2023)	CVPR'23	35.88	1.62	30.54	12.03	13.22	35.47	51.39	49.79	25.63	3.60	43.15	16.23	25.70
SurroundOcc Wei et al. (2023)	ICCV'23	37.04	12.70	31.80	22.50	22.00	29.90	44.70	36.50	24.60	11.50	34.40	18.20	26.27
SplicingOcc Wu et al. (2025)	ICCV'25	49.01	31.60	38.80	35.50	36.30	47.10	54.50	57.20	34.40	32.50	51.20	29.10	40.74
Gaussian-based														
GaussianFormer Huang et al. (2024)	ECCV'24	38.02	17.00	33.60	21.50	21.70	29.40	47.80	37.10	24.30	15.50	36.20	16.80	27.36
EmbodiedOcc Wu et al. (2025)	ICCV'25	51.52	22.70	44.60	37.40	38.00	50.10	56.70	59.70	35.40	38.40	52.00	32.90	42.53
EmbodiedOcc++ Wang et al. (2025a)	ACM MM'25	52.20	27.90	43.90	38.70	40.60	49.00	57.90	59.20	36.80	37.80	53.50	34.10	43.60
RoboOcc Zhang et al. (2025)	arXiv'25	53.30	21.94	44.57	39.54	38.48	51.28	57.04	63.09	36.70	43.05	54.42	34.38	44.05
GPOcc Zhou et al. (2026)	CVPR'26	56.39	40.80	**48.78**	45.62	43.26	50.08	63.97	67.72	48.36	48.77	60.46	45.63	51.22
Voxel-Gaussian hybrid														
VGSSC	-	**59.73**	**42.27**	46.27	**48.44**	**45.22**	**52.83**	**66.83**	**70.05**	**52.16**	**50.97**	**62.39**	**47.45**	**53.17**

Gaussian assistance yields the largest drop to 55.16%. Removing Gaussian context aggregation gives 56.54% mIoU, showing that contextual Gaussian refinement further improves Gaussian-to-

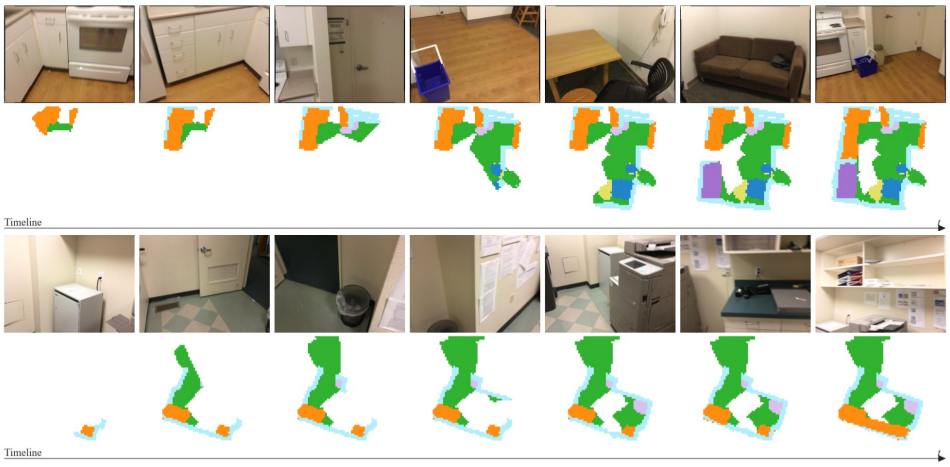

Figure 5: Qualitative results on online frames.

voxel fusion. Random Gaussian initialization reduces mIoU to 55.62%, confirming the benefit of anchoring Gaussians to the predicted surface.

Table 3: Ablation of the VGSSC main components.

Variant	Surface-Aware	Dist. Token	GS-to-Voxel	GS Ctx. Agg.	IoU↑	mIoU↑
VGSSC	✓	✓	✓	✓	**66.26**	**57.16**
w/o Dist. Encoding	✓	✗	✓	✓	64.80	55.40
w/o GS Assist	✓	✓	✗	✗	64.12	55.16
w/o GS Ctx.	✓	✓	✓	✗	65.90	56.54
Random GS	✗	✓	✓	✓	64.61	55.62

Ablation on Surface-Relative Distance Encoding. Table 4 studies the distance representation and normalization scale. Discrete distance tokens achieve 57.16% mIoU, compared with 56.20% for continuous offsets and 55.40% without distance encoding. Among different normalization strategies, image-wise scaling performs best, outperforming pixel-wise, global, and fixed variants.

Table 4: Ablation of the Surface-Relative Distance Encoding.

Variant	Distance Encoding Method	Normalization Scale Type	IoU↑	mIoU↑
VGSSC	Discrete Token	Image-wise	**66.26**	**57.16**
(a)	Discrete Token	Pixel-wise	66.12	56.85
(b)	Discrete Token	Global	65.75	56.63
(c)	Discrete Token	Fixed	65.99	56.71
(d)	Continuous Offset	-	56.54	56.20
(e)	-	-	64.80	55.40

5 Conclusion & Limitation

Conclusion. We present VGSSC, a surface-aware Voxel–Gaussian framework for monocular semantic scene completion. VGSSC uses the predicted surface as a shared geometric reference for surface-relative voxel encoding and surface-anchored Gaussian modeling, combining adaptive local geometry with dense volumetric reasoning. On Occ-ScanNet, VGSSC achieves 57.16% mIoU, surpassing the previous best result by 5.28 %, and further extends to online completion through training-free probability fusion.

Limitation. Our method currently focuses on static indoor scenes with relatively dense surface observations. Extending it to outdoor environments with sparse geometry and dynamic scenes with moving objects remains an important direction for future work.

AI USE STATEMENT

(This section is **required** and does not count toward the page limit.)

In this work, we used generative AI tools for brainstorming and refining research ideas, polishing the writing of the manuscript, and resolving LaTeX typesetting questions. We have not used generative AI tools for generating experimental results, figures. Additionally, we used generative AI tools for code assistance (drafting shell scripts for launching experiments) and language translation. We have reviewed all AI-assisted work. Specifically, we verified all AI-suggested text against our own experimental findings and prior literature, checked LLM-generated scripts line by line before execution, and confirmed that all reported results were produced by our own experiments. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

See the ICLR 2027 AI Policy for Authors for more details. This statement should not be more than 1 page.

REFERENCES

Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4009–4018, June 2021.

Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Localbins: Improving depth estimation by learning local distributions. In *European Conference on Computer Vision*, pp. 480–496. Springer, 2022.

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

Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Tri-perspective view for vision-based 3d semantic occupancy prediction. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9223–9232. IEEE, 2023.

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, 2024.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=rkE3y85ee.

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

Hyo-Jun Lee, Yeong Jun Koh, Hanul Kim, Hyunseop Kim, Yonguk Lee, and Jinu Lee. Soap: Vision-centric 3d semantic scene completion with scene-adaptive decoder and occluded region-aware view projection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17145–17154, June 2025.

Jiaman Li, Alexander Clegg, Roozbeh Mottaghi, Jiajun Wu, Xavier Puig, and C Karen Liu. Controllable human-object interaction synthesis. In *European Conference on Computer Vision*, pp. 54–72. Springer, 2024.

Shijie Li, Zhongyao Cheng, Rong Li, Shuai Li, Juergen Gall, Xun Xu, and Xulei Yang. Global-aware monocular semantic scene completion with state space models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 25550–25559. IEEE, 2025a.

Xiang Li, Pengfei Li, Yupeng Zheng, Wei Sun, Yan Wang, and yilun chen. Semi-supervised vision-centric 3d occupancy world model for autonomous driving. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=rCX9l4OTCT.

Xinpeng Liu, Haowen Hou, Yanchao Yang, Yong-Lu Li, and Cewu Lu. Revisit human-scene interaction via space occupancy. In *European Conference on Computer Vision*, pp. 1–19. Springer, 2024.

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

Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *European conference on computer vision*, pp. 194–210. Springer, 2020.

Zhangyang Qi, Zhixiong Zhang, Ye Fang, Jiaqi Wang, and Hengshuang Zhao. GPT4scene: Understand 3d scenes from videos with vision-language models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=0fib2BYc0L.

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*, pp. 8520–8528, 2026.

Cody Reading, Ali Harakeh, Julia Chae, and Steven L. Waslander. Categorical depth distribution network for monocular 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8555–8564, June 2021.

Shuran Song, Fisher Yu, Andy Zeng, Angel X. Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

Hao Wang, Xiaobao Wei, Xiaoan Zhang, Jianing Li, Chengyu Bai, Ying Li, Ming Lu, Wenzhao Zheng, and Shanghang Zhang. Embodiedocc++: Boosting embodied 3d occupancy prediction with plane regularization and uncertainty sampler. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 925–934, 2025a.

Haochen Wang, Yucheng Zhao, Tiancai Wang, Haoqiang Fan, Xiangyu Zhang, and Zhaoxiang Zhang. Ross3d: Reconstructive visual instruction tuning with 3d-awareness. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9275–9286, October 2025b.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5294–5306. IEEE, 2025c.

Xuzhi Wang, Xinran Wu, Song Wang, Lingdong Kong, and Ziping Zhao. Monocular semantic scene completion via masked recurrent networks. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 24811–24822. IEEE, 2025d.

Xuzhi Wang, Xinran Wu, Song Wang, Lingdong Kong, and Ziping Zhao. Adasformer: Adaptive serialized transformers for monocular semantic scene completion from indoor environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 34148–34159, June 2026.

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.

Yuqi Wu, Wenzhao Zheng, Sicheng Zuo, Yuanhui Huang, Jie Zhou, and Jiwen Lu. Embodiedocc: Embodied 3d occupancy prediction for vision-based online scene understanding. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 26360–26370. IEEE, 2025.

Ruiqi Xian, Jing Liang, He Yin, Xuewei Qi, and Dinesh Manocha. Gaussianssc: Triplane-guided directional gaussian fields for 3d semantic completion. *arXiv preprint arXiv:2603.21487*, 2026.

Tianshuo Xu, Hao Lu, Xu Yan, Yingjie Cai, Bingbing Liu, and Yingcong Chen. Occ-llm: Enhancing autonomous driving with occupancy-based large language models. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 8434–8441. IEEE, 2025.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in neural information processing systems*, 37:21875–21911, 2024.

Tonghui Ye, Zhi Gao, Zhipeng Lin, Xinyi Liu, and Ronghe Jin. SurfOcc: Surface-based feature lifting for vision-centric 3D occupancy prediction. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, pp. 2421–2438, December 2024.

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.

Zhang Zhang, Qiang Zhang, Wei Cui, Shuai Shi, Yijie Guo, Gang Han, Wen Zhao, Hengle Ren, Renjing Xu, and Jian Tang. Roboocc: Enhancing the geometric and semantic scene understanding for robots. *arXiv preprint arXiv:2504.14604*, 2025.

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.

Changqing Zhou, Yueru Luo, and Changhao Chen. Generalizing visual geometry priors to sparse gaussian occupancy prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28578–28587, June 2026.

A Appendix

A.1 More Ablation Studies

We complement the ablations in the main paper by examining the accuracy–efficiency trade-off and the effect of the depth-prior backbone on Occ-ScanNet.

Accuracy, Efficiency, and Model Complexity. Table 5 compares completion accuracy, parameter count, and inference throughput, with all throughput measurements performed on the same NVIDIA RTX 4090 GPU using the official baseline implementations. VGSSC achieves the highest IoU and mIoU among the compared methods with 100.56M parameters and a throughput of 17.21 FPS. Relative to GPOcc, it improves IoU and mIoU by 9.30% and 5.28 %, respectively, while adding 2.61M parameters, an increase of 2.66%. The hybrid model offers higher completion accuracy with a modest increase in parameters over GPOcc.

Table 5: Comparison on the Occ-ScanNet dataset in terms of accuracy, efficiency, and model complexity. We measure the FPS and #Params on the same machine with an NVIDIA RTX 4090 GPU, using the officially released code of each model. **Bold** and underline denote the best and second-best results, respectively.

Model	Venue	**IoU**	**mIoU**	**#Params (M)**	**FPS**
ISO Yu et al. (2024)	ECCV'24	42.16	28.71	638.36	3.89
EmbodiedOcc Wu et al. (2025)	ICCV'25	53.55	45.15	231.45	11.65
SplatSSC Qian et al. (2026)	AAAI'26	<u>62.83</u>	51.83	231.33	10.69
AdaSFormer Wang et al. (2026)	CVPR'26	54.63	45.33	139.52	8.90
GPOcc Zhou et al. (2026)	CVPR'26	56.96	<u>51.88</u>	97.95	**29.35**
VGSSC (Ours)	-	**66.26**	**57.16**	100.56	<u>17.21</u>

Effect of the Depth-Prior Backbone. The predicted surface determines both Gaussian anchor positions and voxel-to-surface depth offsets, making the depth-prior backbone relevant to both representation branches. Table 6 compares the default Depth Anything V2 (DAv2) backbone with a VGGT-based variant. This is the same Depth Anything V2 backbone used for the main experiments and described in our implementation details. Replacing DAv2 with VGGT increases IoU from 66.26% to 71.53% and mIoU from 57.16% to 62.77%, corresponding to gains of 5.27% and 5.61%, respectively. These results show that VGSSC can benefit from an alternative, higher-capacity geometry backbone. However, the total parameter count increases from 100.56M to 944.92M, approximately $9.40\times$ that of the default model. Since the backbone architecture and capacity change together, this comparison does not isolate the contribution of depth accuracy alone. The DAv2 configuration offers a substantially smaller model, while the VGGT variant illustrates the additional completion accuracy available with a larger backbone.

Table 6: Effect of different depth-prior backbone for VGSSC on Occ-ScanNet.

Depth-prior backbone	IoU	mIoU	#Params (M)
DAv2 Yang et al. (2024)	66.26	57.16	100.56
VGGT Wang et al. (2025c)	**71.53**	**62.77**	944.92

A.2 More Implementation Details

Image and voxel features. We use the ViT-B variant of Depth Anything V2, initialized from its ScanNet-finetuned checkpoint, and process RGB images at 392×518 resolution. The target volume is discretized into a $60 \times 60 \times 36$ grid with 8 cm voxels. Image features are projected to 256 channels, while voxel and Gaussian features use 128 channels. For voxel lifting, we bilinearly sample image features and use nearest-neighbor sampling for surface depth. Following Eq. (3) in

the main paper, the sampled image feature and the selected surface-relative distance token are first added to form the surface-aware lifted voxel field $\mathbf{F}_{3D}^{V}$. For implementation, a $1 \times 1 \times 1$ projection followed by two $3 \times 3 \times 3$ convolutions maps this field to the 128-channel voxel representation used by Gaussian initialization and the Gaussian-assisted voxel decoder. We retain the notation $\mathbf{F}_{3D}^{V}$ for this refined representation. These layers use eight-group normalization and GELU activations. Locations outside the image receive a learned appearance vector.

Distance-token assignment. Following Eq. (2) in the main paper, the scale predictor Φ_{scale} takes the image feature $\mathbf{F}_{2D}$ and predicts an image-adaptive positive scalar $S = \text{softplus}(\Phi_{\text{scale}}(\mathbf{F}_{2D}))$. This scale is shared across all voxels in that image. For voxel v, let d_v denote its optical-axis depth minus the predicted surface depth sampled at its image projection. We map this signed offset to $u_v = \text{sigmoid}(d_v/S)$. The distance table contains 64 valid entries indexed from 0 to 63 and one additional entry at index 64 for invalid projections or unavailable depths, each with 256 channels. We initialize its entries from a zero-mean Gaussian distribution with standard deviation 0.02. During training, token k receives the score $a_{vk} = -(64u_v - k - \frac{1}{2})^2$. We restrict negative offsets to the first 32 entries and nonnegative offsets to the remaining 32 entries. Gumbel perturbations and a temperature of 1 yield a hard token lookup in the forward pass, with a soft relaxation providing gradients through the discrete assignment to the predicted scale S and its image-adaptive predictor Φ_{scale}. Only the selected table entry receives the embedding gradient. At inference, valid offsets use the deterministic index $k_v = \text{clip}(\lfloor 64u_v \rfloor, 0, 63)$, consistent with Eq. (2) in the main paper.

Gaussian initialization and refinement. We initialize Gaussian centers by back-projecting image locations using the predicted depth and camera geometry. We initialize each Gaussian feature by combining a linear projection of its image feature with a trilinearly sampled voxel feature. A two-layer MLP predicts the remaining Gaussian attributes. Initial scales are mapped to $[0.02, 0.12]$ m using a sigmoid, rotations are represented by normalized quaternions, and opacities use a sigmoid. Context aggregation uses 64 learnable latent tokens of dimension 128 and two four-head cross-attention operations: the latent tokens first aggregate Gaussian features augmented with projected attributes, and the Gaussians then retrieve the resulting context. Residual heads update centers, features, scales, rotations, opacities, and semantic scores once. Their final output layer is initialized to zero, with a learnable residual coefficient initialized to 10^{-3}. Updated scales are clamped to the same range, rotations are renormalized, and opacities are updated in logit space.

Gaussian-to-voxel conversion. Each Gaussian contributes to a $3 \times 3 \times 3$ neighborhood around its nearest voxel center. We compute opacity-weighted Gaussian kernels using the rotated voxel-to-center displacement and the three axis scales. Features and semantic logits are accumulated with the same weights and divided by their sum, with a denominator floor of 10^{-6}. The normalized feature aggregation defines the Gaussian-derived voxel feature field $\mathbf{F}_{3D}^{G}$, while the normalized semantic logits define the voxel-aligned semantic occupancy cues $\mathbf{M}_{3D}^{G}$. We additionally compute an internal scalar support weight $r^{G}(v) = 1 - \exp(-\sum_i w_{iv})$, where w_{iv} denotes the contribution of Gaussian i to voxel v, to measure the total Gaussian support at each voxel. This support weight gates the Gaussian-derived signals without changing the definitions of $\mathbf{F}_{3D}^{G}$ and $\mathbf{M}_{3D}^{G}$. The resulting $\mathbf{F}_{3D}^{G}$ and $\mathbf{M}_{3D}^{G}$ are then fused with the surface-aware voxel field $\mathbf{F}_{3D}^{V}$ by the Gaussian–Voxel fusion module.

Voxel refinement and prediction. The fused volume undergoes two applications of the same residual operator, each containing two $3 \times 3 \times 3$ convolutions with 128 channels, eight-group normalization, and GELU activations. Updated voxel features connect the two stages. Both stages share an occupancy head with separate semantic and free-space branches. Each branch is a three-layer MLP with two 128-dimensional hidden layers and sine activations. The first-stage prediction is used for auxiliary supervision during training, while inference takes the second-stage output as the final semantic occupancy prediction, matching $T = 2$ in the main paper.

Losses and optimization. We jointly supervise the final occupancy prediction, the first-stage auxiliary prediction, and the predicted depth. The total training objective is

$$\mathcal{L} = \mathcal{L}_{\text{occ}}^{\text{final}} + 0.25\mathcal{L}_{\text{occ}}^{\text{aux}} + 0.2\mathcal{L}_{\text{depth}},$$

with occupancy losses

$$\mathcal{L}_{\text{occ}}^{\text{final}} = 100\mathcal{L}_{\text{focal}}^{\text{final}} + \mathcal{L}_{\text{Lovasz}}^{\text{final}} + \mathcal{L}_{\text{sem}}^{\text{final}} + \mathcal{L}_{\text{geo}}^{\text{final}},$$
$$\mathcal{L}_{\text{occ}}^{\text{aux}} = 100\mathcal{L}_{\text{focal}}^{\text{aux}} + \mathcal{L}_{\text{Lovasz}}^{\text{aux}} + \mathcal{L}_{\text{geo}}^{\text{aux}},$$

where $\mathcal{L}_{\text{focal}}$, $\mathcal{L}_{\text{Lovasz}}$, $\mathcal{L}_{\text{sem}}$, and $\mathcal{L}_{\text{geo}}$ denote focal, Lovasz, semantic scaling, and geometric scaling losses, respectively. The focal-loss parameters are $\alpha = 0.25$ and $\gamma = 2$. The depth term $\mathcal{L}_{\text{depth}}$ is a Huber loss with a transition threshold of 1 m. Predicted depth is bilinearly resized to the ground-truth resolution, and the loss is averaged over pixels with positive ground-truth depth.

We jointly train the image backbone, depth head, and completion network for 10 epochs using AdamW, with an initial learning rate of 2×10^{-4}, weight decay of 0.01, and a batch size of two per GPU. The pretrained image backbone uses a learning-rate multiplier of 0.1. We use an iteration-based cosine learning-rate schedule with a minimum learning rate of 10^{-6} and a 1,000-iteration warmup starting from 10^{-6}. The gradient norm is clipped to 10.

A.3 More Visualization Results

We provide additional qualitative results for single-frame completion on Occ-ScanNet and online semantic scene completion on EmbodiedOcc-ScanNet.

Single-Frame Semantic Scene Completion. Figure 6 compares VGSSC with GPOcc and ground truth across indoor scenes with different object layouts. VGSSC produces more continuous room structures in these examples while recovering the spatial arrangement of furniture. In the meeting-room example in the third row, VGSSC reconstructs more complete chair regions around the central table and a more continuous surrounding floor. In the last row, its floor prediction avoids the gap present in GPOcc's result and more closely matches the ground truth. These observations are consistent with the decoder design in the main paper: surface-anchored Gaussians encode local geometry and semantics near the predicted surface, Gaussian-to-voxel conversion transfers this information to the dense voxel field, and shared voxel refinement propagates the fused context into occluded regions. Nevertheless, differences from ground truth remain at object boundaries and in some semantic assignments, reflecting the remaining ambiguity of monocular completion.

Online Semantic Scene Completion. Figures 7, 8, and 9 show how the global semantic volume evolves as additional posed frames become available. Early predictions cover a limited portion of each scene, whereas subsequent viewpoints extend the reconstruction to additional furniture, floor regions, and room boundaries. For example, the meeting-room sequence in Fig. 7 progressively accumulates the table-and-chair layout, while the kitchen sequence in Fig. 9 expands from a local cabinet region to a broader reconstruction of the surrounding workspace. For each frame n, the fixed single-frame model first produces a local class-probability volume $\mathbf{P}_n$ in its camera-centered voxel grid. Using the known camera pose, we align this local grid with the global coordinate system and trilinearly sample $\mathbf{P}_n$ to obtain a world-aligned probability observation for each global voxel. The online prediction then updates a weighted probability sum $A_n(v)$ and an accumulated observation weight $B_n(v)$, so previously observed regions remain represented while new viewpoints contribute additional evidence. Prediction confidence and viewing geometry determine the observation weight, with invalid observations assigned zero weight and free-space or behind-surface predictions down-weighted as described in Sec. 3.5. The fused class distribution is obtained from the normalized accumulated statistics and converted to the global semantic prediction by an $\arg\max$ over classes. The single-frame network remains fixed throughout the sequence, and the global memory stores only accumulated class-probability statistics and observation weights rather than Gaussian primitives or learned voxel features. The resulting sequences demonstrate incremental scene coverage without introducing a separately trained temporal model.

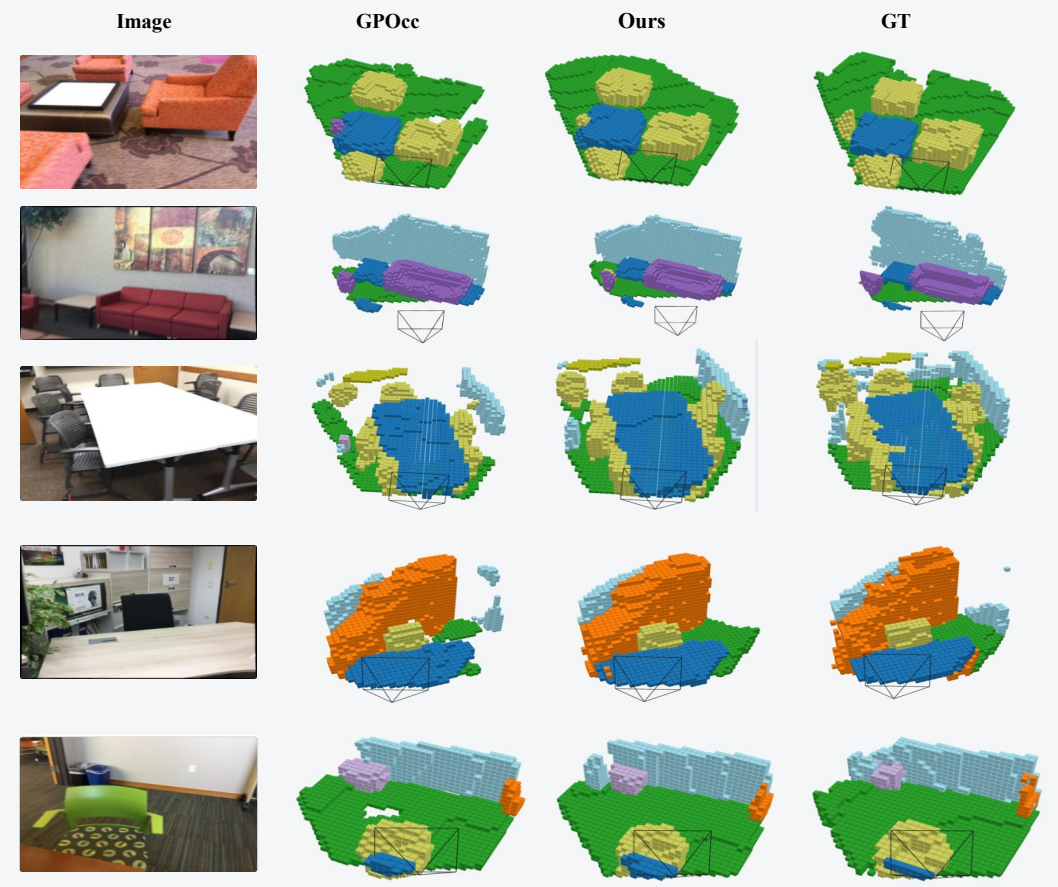

Figure 6: Qualitative comparison on Occ-ScanNet.

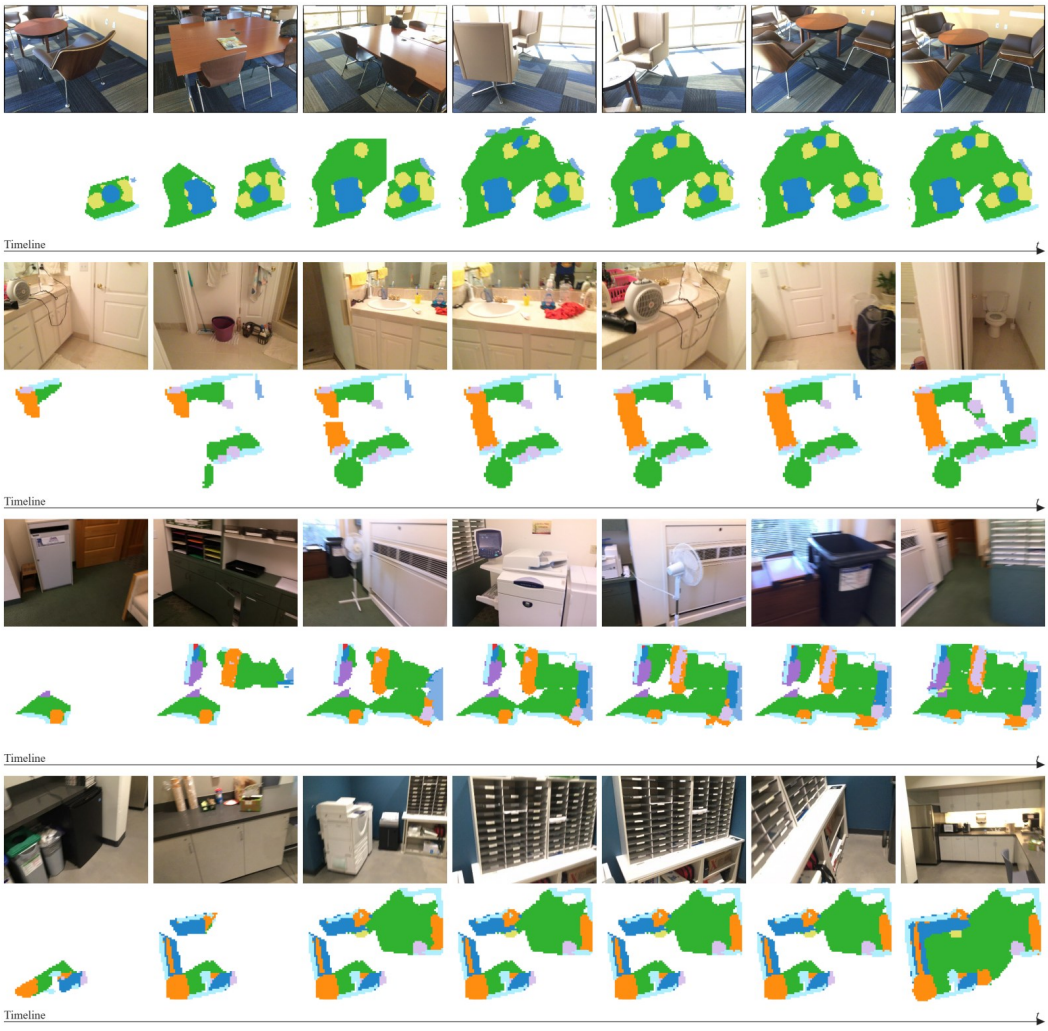

Figure 7: Qualitative results on online frames.

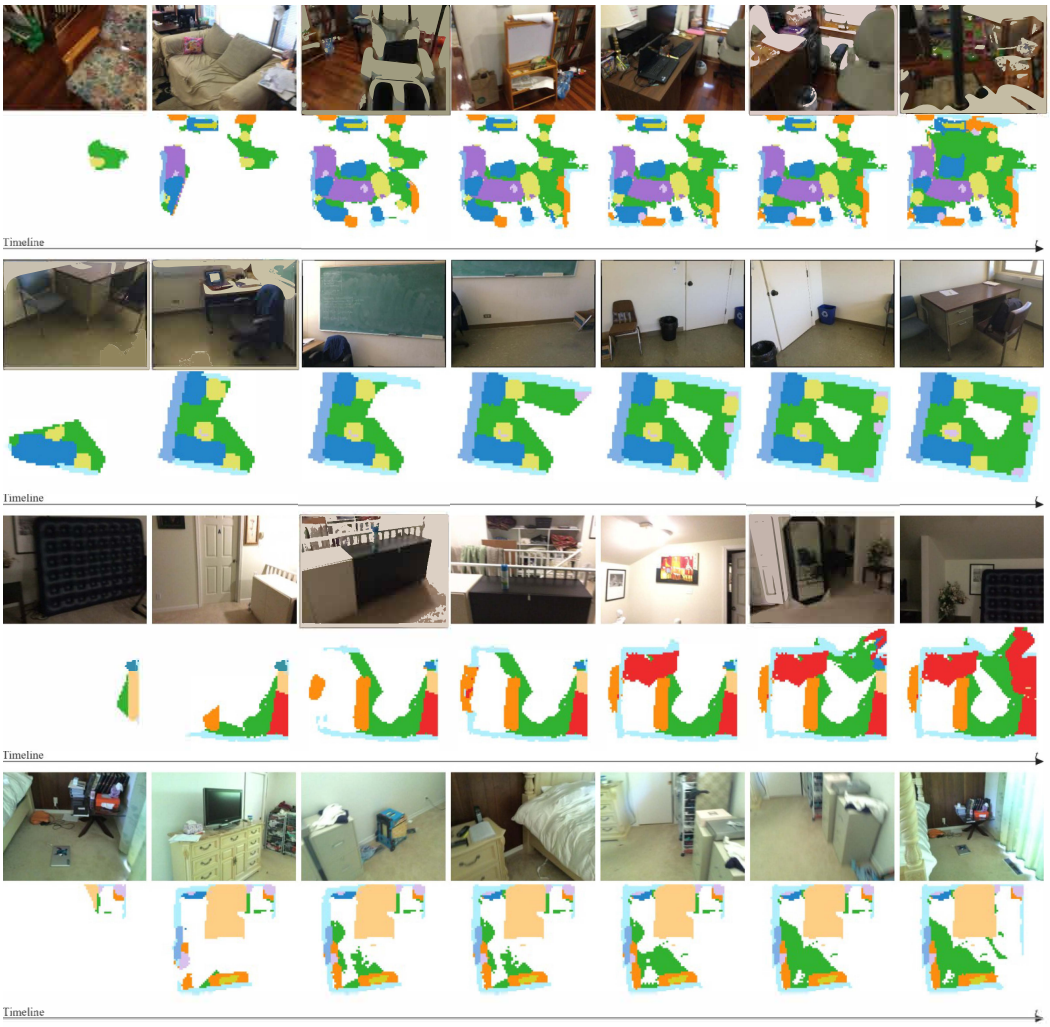

Figure 8: Qualitative results on online frames.



Figure 9: Qualitative results on online frames.