STRUCTURE-GUIDED CAPACITY ALLOCATION FOR SPARSE-VIEW INDOOR GAUSSIAN SPLATTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Sparse-view indoor reconstruction with 3D Gaussian Splatting remains challenging because limited cross-view supervision leaves weakly textured surfaces and occluded structures poorly constrained. Unrestricted densification can further reinforce unreliable geometry by allocating additional capacity to ambiguous regions. We propose a structure-guided capacity allocation framework that uses explicit Gaussian roles to guide candidate selection and constrain split directions. Starting from scale-consistent multi-view geometry, we initialize ordinary, line, and planar Gaussians, with planar roles further distinguished by texture complexity. For candidate selection, we introduce a type-calibrated score that combines gradient-based fitting demand with rendering contributions normalized within each structural type. The resulting scores prioritize candidates under a shared Top-K growth budget. For growth, persistent structural priors restrict line and high-texture planar split displacements to one-dimensional tangents and two-dimensional tangent subspaces, respectively, while low-texture planar Gaussians are excluded from densification. Descendants inherit these structural attributes, retaining the growth constraints across successive densification stages, while geometric regularization stabilizes subsequent optimization. Experiments demonstrate improved perceptual quality and better preservation of indoor structural details in novel-view renderings compared with representative baselines.

1 INTRODUCTION

3D Gaussian Splatting (3DGS) provides an efficient representation for high-quality novel-view synthesis through explicit anisotropic Gaussian primitives. However, reconstructing indoor scenes from sparse views remains challenging: with limited view overlap, large weakly textured surfaces, occlusions, and thin structures leave geometry and appearance poorly constrained. Gaussians can consequently fit the observed images while producing artifacts at unseen viewpoints. Adaptive density control (ADC) can compound this ambiguity when training-view residuals trigger additional primitives around unreliable geometry (Zhang et al., 2024a; Xu et al., 2025).

A reliable geometric initialization is essential for reducing this ambiguity. Conventional 3DGS initializes Gaussian primitives from sparse Structure-from-Motion (SfM) points, typically reconstructed using COLMAP (Schonberger & Frahm, 2016). In sparse indoor captures, SfM often yields incomplete and uneven point coverage, particularly in weakly textured regions where photometric supervision provides insufficient evidence to recover missing geometry. Recent methods therefore incorporate learned depth priors to improve geometric support (Sun et al., 2026; Yuan et al., 2026; Kim et al., 2026). Beyond the quality of the initial geometry, reconstruction also depends on how Gaussian capacity is distributed and expanded during optimization. Different indoor structures impose distinct representation requirements: weakly textured planes can often be modeled with compact, surface-aligned support, textured surfaces require finer spatial detail, and thin line-like structures benefit from directionally aligned primitives. These differences motivate using structural cues to guide both the initial representation and the allocation of additional capacity during optimization.

Existing indoor reconstruction methods incorporate geometric and structural priors to improve reconstruction quality (Jin et al., 2026; Han et al., 2026; Xiang et al., 2025). In particular, IndoorGS (Ruan et al., 2025) combines geometric-cue-based initialization with type-specific ADC.

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.

Complementary densification studies improve gradient-based candidate discovery (Ye et al., 2024) and develop splitting and growth-control strategies (Deng et al., 2026). These advances provide foundations for deciding when and how to introduce new primitives. Under sparse supervision, however, a large positional gradient indicates fitting demand without establishing whether additional capacity will improve novel-view generalization. Allocating a limited growth budget therefore requires prioritizing candidates across structural roles. Meanwhile, Gaussian covariance orientations remain trainable, so splitting along their current axes may introduce displacements inconsistent with the structural priors used at initialization. These observations motivate two coupled questions: *which Gaussians should receive additional capacity, and along which directions should their descendants be introduced?*

To address these questions, we propose a structure-guided capacity allocation framework that uses persistent Gaussian roles to guide candidate selection and constrain split directions. Starting from jointly inferred, scale-consistent multi-view geometry, we combine line, plane, and texture cues to initialize four Gaussian types—ordinary, line, low-texture planar, and high-texture planar—with sampling densities, orientations, and scales adapted to their structural roles. These roles subsequently determine contribution-calibration groups and eligible growth operations. For candidate selection, we use absolute screen-space gradients to identify Gaussians requiring refinement. We then normalize rendering contributions using within-type percentiles to reduce systematic differences in contribution scale across structural types. The calibrated contributions modulate the gradient signal, yielding a type-calibrated score that combines fitting demand with relative rendering participation. All eligible candidates are ranked by this score and compete under a shared Top-K growth budget at each densification step. For growth, persistent structural priors define admissible split subspaces independently of the evolving covariance orientations. Line Gaussians split along their stored three-dimensional tangents, while high-texture planar Gaussians split within the tangent planes defined by their target normals. Ordinary Gaussians remain eligible for cloning or splitting without structural direction constraints. Low-texture planar Gaussians are excluded from densification to limit growth in homogeneous regions with weak photometric constraints. Descendants inherit their parents' structural attributes, preserving the admissible subspaces across successive split events. These constraints govern child placement during splitting, while Gaussian centers and covariance parameters remain trainable afterward. Depth, normal, and co-planarity regularization further stabilize subsequent geometric optimization.

Our contributions are summarized as follows:

- We propose a structure-guided capacity allocation framework for sparse-view indoor reconstruction that uses persistent Gaussian roles to jointly guide candidate selection and constrain split directions across successive densification stages.
- We introduce a type-calibrated scoring rule that combines gradient-based fitting demand with relative rendering participation to rank densification candidates under a shared Top-K growth budget.
- We develop a structural-subspace growth strategy that constrains split directions using inherited line and plane priors, together with texture-dependent densification policies for planar regions.

2 Related Work

Sparse-View Novel View Synthesis and Reconstruction. Sparse-view novel-view synthesis is severely under-constrained because limited observations provide insufficient supervision. NeRF-based methods address this ambiguity through regularization and geometric priors. FreeNeRF (Yang et al., 2023) regularizes encoding frequencies and near-camera density, while Roessle et al. (2022) use dense depth priors to constrain reconstruction. For indoor scenes, P^2NeRF (Sun et al., 2024) combines correspondence-based initialization with hierarchical depth-consistency regularization. Sparse-view Gaussian methods similarly strengthen geometric support to compensate for incomplete SfM initialization. TWINGS (Kim et al., 2026) aligns monocular-depth backprojections through non-rigid warping, whereas VGA-GS (Sun et al., 2026) combines global depth alignment with view-space Gaussian optimization. StereoGS (Yuan et al., 2026) uses dense multi-view stereo initialization and stereo regularization. Related reconstruction methods incorporate depth priors and geometric or feature consistency to stabilize sparse-view geometry (Huang et al., 2025; Gu et al.,

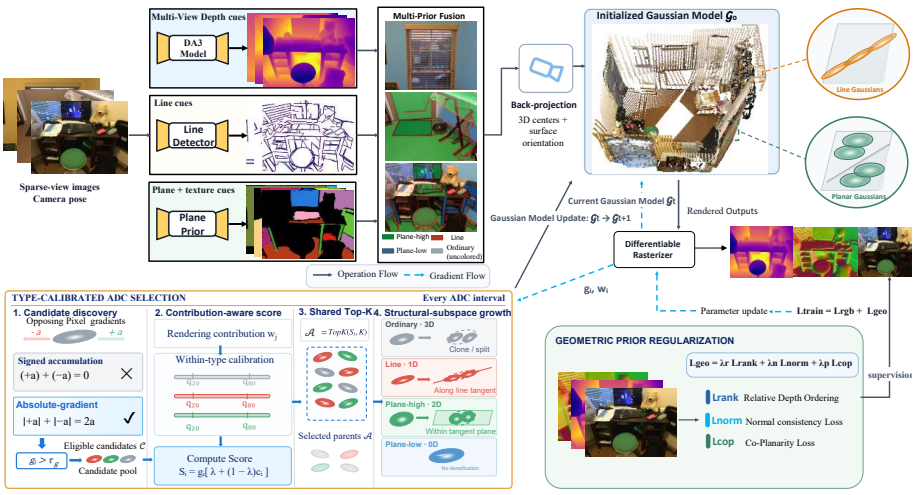


Figure 1: Overview of our structure-guided sparse-view indoor Gaussian splatting framework.

2026). Complementary approaches control representation capacity to improve generalization. DropGaussian (Park et al., 2025) and DropoutGS (Xu et al., 2025) employ Gaussian dropout, with the latter also introducing edge-guided splitting. Fang et al. (2026) extend dropout to spatial neighborhoods and higher-order spherical-harmonic coefficients, while D^{2}GS (Song et al., 2026) adapts it to depth and local Gaussian density. CoR-GS (Zhang et al., 2024a) uses disagreement between paired Gaussian fields for co-pruning and pseudo-view regularization. AbsGS (Ye et al., 2024) reduces gradient cancellation, and Deng et al. (2026) adapt edge-aware candidate scoring, long-axis splitting, and growth control.

Indoor Gaussian Splatting. Indoor scenes contain extended weakly textured surfaces and partially occluded structures whose geometry is difficult to resolve from sparse observations. Existing methods incorporate geometric priors to stabilize reconstruction. DN-Splatter (Turkulainen et al., 2025) uses depth and normal supervision, GaussianRoom (Xiang et al., 2025) couples Gaussian splatting with an SDF representation, and 2DGS-Room (Zhang et al., 2024b) combines seed-guided Gaussian allocation with monocular and multi-view geometric constraints. Other methods explicitly exploit indoor structural regularities. 360-GS (Bai et al., 2025) uses room-layout guidance for panoramic reconstruction, while LighthouseGS (Han et al., 2026) constructs planar scaffolds for panorama-style mobile captures. PlanarSplatting (Tan et al., 2025) represents planar surfaces explicitly, and PlanarGS (Jin et al., 2026) uses vision-language planar priors to guide indoor reconstruction. IndoorGS (Ruan et al., 2025) combines line, plane-like, and SfM cues with type-specific initialization and density control. Its tangent-space densification for line-derived Gaussians uses normals derived from learned Gaussian orientations. Our method couples type-calibrated ranking under a shared growth budget with split subspaces defined by persistent line tangents and target plane normals, independently of covariance rotation. Texture-dependent planar policies enable refinement of textured surfaces while suppressing growth in homogeneous regions.

3 METHOD

We propose a structure-guided capacity allocation framework for sparse-view indoor Gaussian splatting. Starting from jointly inferred, scale-consistent multi-view geometry, we initialize four Gaussian types—ordinary, line, low-texture planar, and high-texture planar—with role-specific sampling densities, orientations, and scales. These persistent structural roles guide two complementary densification decisions: which Gaussians receive additional capacity and along which directions their descendants are introduced. A type-calibrated score combines gradient-based fitting demand with relative rendering participation to rank candidates under a shared Top-K growth budget. Persistent structural priors define admissible split subspaces and are inherited by descendants across successive densification stages. Depth, normal, and co-planarity regularization further stabilize subsequent parameter optimization. Fig. 1 illustrates the framework.

3.1 STRUCTURE-AWARE MULTI-VIEW INITIALIZATION

Our initialization establishes geometric support and assigns structural attributes for subsequent capacity allocation. We combine multi-view depth with line, planar, and texture cues to adapt the distribution and shape of Gaussian primitives to structures. The resulting structural types and directional priors are retained to guide candidate selection and constrain split displacements.

Joint Multi-View Geometry. Given N calibrated images $\{I_v\}_{v=1}^N$ with intrinsics $\mathbf{K}_v$ and camera-to-world poses $\mathbf{T}_v^{c2w}$, we jointly infer depth maps using the pretrained multi-view depth predictor from Depth Anything 3 (DA3) (Lin et al., 2025):

$$\{\widehat{\mathbf{D}}_v\}_{v=1}^N = \mathcal{D}_{\mathrm{DA3}}\left(\{I_v, \mathbf{K}_v, \mathbf{T}_v^{w2c}\}_{v=1}^N\right), \tag{1}$$

where $\mathbf{T}_v^{w2c} = (\mathbf{T}_v^{c2w})^{-1}$ and $\widehat{\mathbf{D}}_v$ denotes the predicted camera-space z-depth map. Joint prediction provides scale-consistent geometric support across the input views (Lin et al., 2025). Let $\mathbf{R}_v^{c2w}$ and $\mathbf{t}_v^{c2w}$ denote the rotation and translation of the source camera. A sampled pixel $\mathbf{p}$ with valid depth and homogeneous coordinate $\widetilde{\mathbf{p}}$ defines a world-space anchor:

$$\mathbf{X}_v(\mathbf{p}) = \mathbf{R}_v^{c2w}\left(\widehat{D}_v(\mathbf{p})\mathbf{K}_v^{-1}\widetilde{\mathbf{p}}\right) + \mathbf{t}_v^{c2w}. \tag{2}$$

Each sampled anchor initializes a Gaussian center, with color taken from $I_v(\mathbf{p})$. The Gaussian retains its source-view association, and its center is optimized through normalized image coordinates and a view-space depth variable. Its structural role determines the sampling policy, initial orientation, and scale, as described below.

3.1.1 STRUCTURAL ROLE ASSIGNMENT.

We assign each Gaussian $\mathcal{G}_i$ a structural type $z_i \in \mathcal{T} = \{\text{ordinary}, \text{line}, \text{plane-low}, \text{plane-high}\}$. Line Gaussians provide directional support along detected structures. Planar Gaussians provide surface-aligned support, with low- and high-texture roles reflecting different requirements for spatial detail. Ordinary Gaussians cover the remaining valid regions. Line assignments take precedence, and their neighborhoods are excluded from planar and ordinary sampling to keep the roles disjoint. These structural types define the groups used to calibrate rendering contributions and determine the eligible growth operations.

Oriented Line Gaussian Initialization. To represent thin structures with directional support, we extract image-space line segments (Ke et al., 2025) and lift their local directions using the predicted depth. For a segment $\boldsymbol{\ell}_j = (\mathbf{p}_j^0, \mathbf{p}_j^1)$ with detector support score s_j, we retain it when

$$s_j \geq \tau_s, \qquad \|\mathbf{p}_j^1 - \mathbf{p}_j^0\|_2 \geq \tau_\ell, \tag{3}$$

where τ_s and τ_ℓ are the score and minimum-length thresholds. Each retained segment is uniformly sampled at a target spacing Δ_ℓ. For a line Gaussian $\mathcal{G}_i$ sampled at pixel $\mathbf{p}$ from segment j in view v, let $\mathbf{t}_j^{2D} = (\mathbf{p}_j^1 - \mathbf{p}_j^0)/\|\mathbf{p}_j^1 - \mathbf{p}_j^0\|_2$ be the unit image-space direction. We estimate its local 3D direction from two neighboring back-projected points.

$$\widetilde{\mathbf{t}}_i = \left[\mathbf{X}_v(\mathbf{p} + \delta\mathbf{t}_j^{2D}) - \mathbf{X}_v(\mathbf{p} - \delta\mathbf{t}_j^{2D})\right] / \left\|\mathbf{X}_v(\mathbf{p} + \delta\mathbf{t}_j^{2D}) - \mathbf{X}_v(\mathbf{p} - \delta\mathbf{t}_j^{2D})\right\|_2, \tag{4}$$

where δ is the sampling offset in pixels along the detected line. The unit tangent $\widetilde{\mathbf{t}}_i$ defines the major axis of the initial Gaussian frame. A local surface normal $\mathbf{m}_i$ derived from the predicted depth provides a reference for constructing the remaining orthogonal axes, with the source-view direction used as a fallback when a valid normal is unavailable. This construction initializes directionally aligned Gaussians without requiring complete 3D line reconstruction.

Scales based on neighboring-point distances are sensitive to sampling density and overlap between views. For line Gaussians, we reduce this dependence by using the source-view pixel footprint as a reference and controlling the directional extents separately:

$$\mathbf{s}_i^{\mathrm{line}} = \rho_v(\mathbf{p})(\gamma_\parallel, \gamma_\perp, \gamma_n), \qquad \rho_v(\mathbf{p}) = \widehat{D}_v(\mathbf{p})/\sqrt{f_{x,v} f_{y,v}}, \tag{5}$$

where $f_{x,v}$ and $f_{y,v}$ are the source-view focal lengths in pixels. The reference scale $\rho_v(\mathbf{p})$ approximates the spatial extent of one image pixel at the predicted depth, using the geometric mean of the horizontal and vertical pixel footprints. The multipliers control the extent along the line, the in-surface cross-line direction, and the normal axis, respectively. We use $\gamma_\parallel > \gamma_\perp > \gamma_n$ to initialize elongated, thin primitives aligned with the detected structure.

Texture-Aware Planar Gaussian Initialization. We obtain planar masks, global plane identities, and target normals from planar preprocessing (Jin et al., 2026), and use texture complexity to determine the initial sampling density of each plane.

For each planar observation, we compute the mean clipped RGB-gradient magnitude and the fraction of strong-gradient pixels to classify its texture level. We then aggregate these per-view labels through pixel-count-weighted voting over observations of the same global plane. The resulting *plane-high* or *plane-low* role is shared across its observations, ensuring consistent texture assignments across views. High-texture planes receive denser sampling to capture fine appearance variations, while low-texture planes use sparser support. For both roles, one axis of the Gaussian frame is aligned with the target plane normal and assigned a reduced scale to produce thin surface-aligned primitives. Valid regions outside the line and planar masks are sampled more coarsely to initialize ordinary Gaussians.

Persistent Structural Attributes. Each Gaussian retains a fixed structural type, together with a line tangent for line primitives or a target normal and global plane identity for planar primitives. Descendants inherit these attributes during densification. Structural types determine the contribution-calibration groups and eligible growth operations, while stored tangents and normals define the admissible split subspaces. Gaussian centers and covariance parameters remain trainable, allowing geometric refinement while preserving the structural references used for capacity allocation.

3.2 Type-Calibrated ADC Selection

Under sparse-view supervision, a large optimization gradient indicates fitting demand but does not establish whether introducing additional primitives will improve generalization. We therefore separate candidate discovery from budget allocation: absolute screen-space gradients identify Gaussians requiring refinement, while type-calibrated rendering contributions modulate their priority under a shared growth budget.

Candidate Discovery. We use component-wise absolute screen-space gradients to reduce cancellation between opposing pixel contributions (Ye et al., 2024). For Gaussian $\mathcal{G}_i$, let $k = 1, \ldots, N_i$ index the training renders between consecutive ADC steps in which it has a positive projected radius. Its accumulated gradient statistic is

$$g_i = \frac{1}{N_i} \sum_{k=1}^{N_i} \sum_{\mathbf{p}} \left(\left| \frac{\partial \ell_k(\mathbf{p})}{\partial x_i^{(k)}} \right| + \left| \frac{\partial \ell_k(\mathbf{p})}{\partial y_i^{(k)}} \right| \right), \tag{6}$$

where $(x_i^{(k)}, y_i^{(k)})$ is the projected Gaussian center and $\ell_k(\mathbf{p})$ denotes the pixel-wise photometric reconstruction loss. The resulting absolute screen-space positional gradients are accumulated over the N_i eligible renders and averaged to obtain g_i. The gradient accumulators and render counts are reset after each densification step.

Rendering Contribution. To complement fitting demand, we measure how strongly each Gaussian participates in the current renderings. Over the same accumulation window, we compute

$$w_i^{(k)} = \sum_{\mathbf{p}} \alpha_i^{(k)}(\mathbf{p}) T_i^{(k)}(\mathbf{p}), \qquad w_i = \frac{1}{N_i} \sum_{k=1}^{N_i} w_i^{(k)}, \tag{7}$$

where $\alpha_i^{(k)}(\mathbf{p})$ is the per-pixel alpha and $T_i^{(k)}(\mathbf{p})$ is the transmittance accumulated before $\mathcal{G}_i$ along the corresponding ray. Their product is its effective blending weight at pixel $\mathbf{p}$. Thus, $w_i^{(k)}$ is the total blending weight in one render, and w_i is its average over the current ADC window.

Raw contributions depend on projected footprint, opacity, and visibility, so their distributions can differ across structural types. Broad surface primitives, for example, may accumulate larger contributions than thin line primitives because they participate in more pixels. Using raw contributions directly can therefore favor types with larger contribution scales. We instead calibrate each candidate's contribution relative to candidates of the same structural type.

Type Calibration and Budget Allocation. Let $\mathcal{C}$ contain Gaussian indices satisfying the gradient threshold and positive rendering contribution. Let $z_i \in \mathcal{T}$ denote the structural type of $\mathcal{G}_i$, and define $\mathcal{C}_z = \{i \in \mathcal{C} \mid z_i = z\}$, we calibrate rendering contributions using within-type percentiles:

$$q_r^{(z)} = Q_r(\{w_j \mid j \in \mathcal{C}_z\}), \qquad c_i = \mathrm{clip}\left(\frac{w_i - q_{20}^{(z_i)}}{q_{80}^{(z_i)} - q_{20}^{(z_i)} + \epsilon}, 0, 1\right). \tag{8}$$

Here, $Q_r(\cdot)$ denotes the r-th percentile operator, $q_r^{(z)}$ is the corresponding percentile among candidates of type z, and $\epsilon > 0$ stabilizes the normalization. The calibrated contribution $c_i \in [0, 1]$ expresses rendering participation relative to candidates of the same type. Percentile clipping limits the influence of extreme values and reduces systematic contribution-scale differences across types.

We then use the calibrated contribution to modulate the gradient score and select candidates under a shared growth budget:

$$S_i = g_i\left[\lambda + (1 - \lambda)c_i\right], \qquad \mathcal{A} = \mathrm{TopK}_{i \in \mathcal{C}}(S_i, K). \tag{9}$$

We set $\lambda = 0.5$, bounding the contribution factor within $[0.5, 1]$ and retaining at least half of each candidate's gradient score. Candidates compete under a shared Top-K budget without type-specific quotas, yielding a selected set $\mathcal{A}$ of at most K parents. This budget bounds the number of selected parents at each ADC step, while the type-specific growth rules below determine their permitted densification operations.

3.3 Structural-Subspace Growth

Candidate selection determines which Gaussians receive additional capacity, but their split directions also affect the resulting geometry. Under sparse supervision, displacements along weakly constrained directions can introduce primitives inconsistent with the structural priors. Moreover, Gaussian covariance orientations remain trainable, so their current axes may no longer align with the structures identified at initialization. We therefore use persistent line and plane priors to define admissible split subspaces independently of the evolving covariance orientations.

Growth Eligibility. For selected candidates, small ordinary Gaussians are cloned, whereas large ordinary Gaussians, line Gaussians, and high-texture planar Gaussians are split. Low-texture planar Gaussians are excluded from densification to limit capacity growth in homogeneous regions with weak photometric constraints.

Admissible Split Subspaces. For Gaussian $\mathcal{G}_i$ with world-space covariance $\mathbf{\Sigma}_i$, we define the orthogonal projector onto its admissible displacement subspace as

$$\mathbf{P}_i = \begin{cases} \mathbf{I}, & z_i = \text{ordinary}, \\ \widetilde{\mathbf{t}}_i \widetilde{\mathbf{t}}_i^\top, & z_i = \text{line}, \\ \mathbf{I} - \mathbf{n}_i \mathbf{n}_i^\top, & z_i = \text{plane-high}, \\ \mathbf{0}, & z_i = \text{plane-low}, \end{cases} \tag{10}$$

where $\widetilde{\mathbf{t}}_i$ and $\mathbf{n}_i$ are the stored unit line tangent and target plane normal, respectively. The projector encodes the directions permitted by the structural role.

For each selected Gaussian eligible for splitting, we have $\mathrm{rank}(\mathbf{P}_i) > 0$ and choose $\mathbf{d}_i = \mathbf{v}_{\max}(\mathbf{P}_i \mathbf{\Sigma}_i \mathbf{P}_i)$, where $\mathbf{v}_{\max}(\cdot)$ denotes a unit eigenvector associated with the largest eigenvalue. This direction maximizes spatial variance within the admissible subspace. The structural prior fixes the subspace, while the current covariance determines the principal direction within it. For ordinary Gaussians, $\mathbf{d}_i$ is the principal covariance direction. Line Gaussians satisfy $\mathbf{d}_i = \pm\widetilde{\mathbf{t}}_i$, while high-texture planar Gaussians use the principal direction within the target tangent plane, satisfying $\mathbf{d}_i^\top \mathbf{n}_i = 0$. The admissible subspaces have dimensions 3, 1, 2, and 0 for ordinary, line, high-texture planar, and low-texture planar Gaussians, respectively. These dimensions describe the permitted displacement directions; each split places two children along a single selected direction.

Child Construction and Attribute Inheritance. Each parent selected for splitting produces two children at $\boldsymbol{\mu}_i^{\pm} = \boldsymbol{\mu}_i \pm \eta s_i \mathbf{d}_i$, where $\boldsymbol{\mu}_i$ is its center and s_i its scale along $\mathbf{d}_i$ in the structure-aligned frame. We set $\eta = 1.35$, combining the relative split offset and scale multiplier used in LAS (Deng

Table 1: Quantitative comparison on ScanNet, Replica, and NRGBD. "+ DA3" replaces each baseline's original depth prior with shared DA3 predictions while retaining the remaining pipeline and settings. The best and second-best results are boldfaced and underlined, respectively.

Method	ScanNet			Replica			NRGBD		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
P^2NeRF (Sun et al., 2024)	20.53	0.737	0.427	30.50	0.906	0.166	–	–	–
DropGaussian (Park et al., 2025)	19.71	0.731	0.358	31.85	0.948	0.093	31.87	0.950	0.088
StereoGS (Yuan et al., 2026)	20.89	0.763	0.340	30.97	0.933	0.150	31.80	0.945	0.113
D^2GS (Song et al., 2026)	19.25	0.720	0.364	31.72	0.947	0.091	31.71	0.946	0.088
D^2GS + DA3	18.95	0.712	0.370	31.68	0.948	0.090	31.67	0.946	0.089
TWINGS (Kim et al., 2026)	20.90	0.767	0.319	32.72	0.953	0.086	**32.88**	**0.955**	<u>0.079</u>
TWINGS + DA3	20.91	0.768	0.319	32.98	<u>0.954</u>	<u>0.085</u>	<u>32.75</u>	<u>0.954</u>	0.080
VGA-GS (Sun et al., 2026)	21.38	0.763	0.307	31.46	0.932	0.123	31.72	0.949	0.085
VGA-GS + DA3	<u>21.68</u>	<u>0.777</u>	<u>0.296</u>	<u>33.18</u>	0.953	0.094	31.62	0.951	0.081
Ours	**21.92**	**0.786**	**0.287**	**33.50**	**0.958**	**0.078**	32.18	**0.955**	**0.074**

et al., 2026), and retain its child-scale contraction and opacity initialization. High-texture planar children are initialized with the orientation frame $[\mathbf{d}_i, \mathbf{n}_i \times \mathbf{d}_i, \mathbf{n}_i]$. The child centers are projected into the parent's source view and re-encoded using the same normalized image coordinates and view-space depth parameterization. Each child inherits the parent's structural type and associated priors: the stored tangent for line Gaussians, or the target normal and global plane identity for planar Gaussians. These attributes remain fixed and define the admissible subspaces for subsequent splits. The constraints govern child placement at split events; centers and covariance parameters remain trainable afterward, with geometric regularization stabilizing subsequent optimization.

3.4 Geometric Prior Regularization

Structure-guided densification controls capacity allocation and split displacements, while subsequent parameter updates remain subject to sparse photometric supervision. We therefore reuse the geometric priors underlying initialization to stabilize relative depth ordering, local surface orientation, and regional planarity. For the current training view, let D_{prior} denote the fixed depth prior used for initialization and D_{r} the rendered depth.

Relative Depth Ordering. To reduce sensitivity to errors in absolute depth values, we supervise their relative ordering using a group-wise ranking loss (Sun et al., 2026). Valid pixels are sorted by D_{prior} and divided into B ordered groups from near to far. For rendered depth samples $\hat{d}_a$ and $\hat{d}_b$ paired from groups a and b, the ranking penalty is

$$\mathcal{L}_{\text{rank}} = \sum_{a=1}^{B} \sum_{b=1}^{B} \max\Big(\text{sgn}(b-a)\left(\hat{d}_a - \hat{d}_b\right), 0\Big). \tag{11}$$

The penalty is evaluated over cross-group sample pairings, with pixels shuffled within each group during training. It penalizes violations of the prior-induced ordering without requiring rendered depths to match the predicted values exactly.

Normal Consistency. To constrain local surface orientation, we derive reference normals $\mathbf{N}_{\text{prior}}$ from the fixed depth prior using finite differences. Let $\widehat{\mathbf{N}}_{\text{r}}$ denote the alpha-composited Gaussian normals and $\widehat{\mathbf{N}}_{\text{d}}$ the normals computed from rendered depth. All normals are expressed in a common coordinate frame and normalized to unit length. Defining $\hbar_n(\mathbf{a}, \mathbf{b}) = 1 - \mathbf{a}^\top \mathbf{b}$, we minimize

$$\mathcal{L}_{\text{norm}} = \mathbb{E}_{\mathbf{p}\in\Omega_n}\left[\hbar_n\Big(\widehat{\mathbf{N}}_r(\mathbf{p}), \mathbf{N}_{\text{prior}}(\mathbf{p})\Big) + \hbar_n\Big(\widehat{\mathbf{N}}_d(\mathbf{p}), \mathbf{N}_{\text{prior}}(\mathbf{p})\Big)\right], \tag{12}$$

where Ω_n contains pixels with valid normals. The two terms align the rendered Gaussian orientations and the depth-induced surface orientations with the reference geometry.

Co-Planarity. To stabilize extended planar regions, we impose consistency with a plane fitted to the current rendered geometry (Jin et al., 2026). For each planar region, we back-project D_{r}, fit a plane to the resulting points, and compute its induced depth D_π. The co-planarity loss is

$$\mathcal{L}_{\text{cop}} = \mathbb{E}_{\mathbf{p}\in\mathcal{P}}\left|D_{\text{r}}(\mathbf{p}) - D_\pi(\mathbf{p})\right|, \tag{13}$$

where $\mathcal{P}$ denotes valid planar pixels and $D_\pi(\mathbf{p})$ is evaluated using the plane fitted to the corresponding region.



(a) P^{2}NeRF (b) DropGaussian (c) D^{2}GS (d) TWINGS (e) StereoGS (f) VGA-GS (g) Ours (h) GT

Figure 2: Qualitative comparison of novel-view renderings on ScanNet.

The complete training objective is

$$\mathcal{L}_{\text{train}} = \mathcal{L}_{\text{RGB}} + \lambda_r \mathcal{L}_{\text{rank}} + \lambda_n \mathcal{L}_{\text{norm}} + \lambda_p \mathcal{L}_{\text{cop}}, \tag{14}$$

where $\mathcal{L}_{\text{RGB}} = (1 - \lambda_s)\mathcal{L}_1 + \lambda_s \mathcal{L}_{\text{SSIM}}$ with $\lambda_s = 0.2$. Together, these terms support appearance fitting and stabilize geometric optimization between successive densification steps.

4 Experiments

Our implementation uses the view-space Gaussian representation of VGA-GS (Sun et al., 2026) and the RaDe-GS rasterizer (Zhang et al., 2026). We optimize each scene for 10K iterations, with a default shared per-step growth budget of $K = 2000$ for ScanNet and $K = 3000$ for Replica and NRGBD. We evaluate on ScanNet (Dai et al., 2017), Replica (Straub et al., 2019), and Neural RGB-D (NRGBD) (Azinović et al., 2022), following the scenes and train–test splits of VGA-GS. ScanNet includes three scenes with 18–20 training and eight test views per scene (Roessle et al., 2022). Replica includes eight scenes with 20 training and 20 test views per scene (Sun et al., 2024); NRGBD uses the same view counts. We compare with P^{2}NeRF (Sun et al., 2024), DropGaussian (Park et al., 2025), D^{2}GS (Song et al., 2026), TWINGS (Kim et al., 2026), StereoGS (Yuan et al., 2026), and VGA-GS (Sun et al., 2026), using their official implementations and recommended configurations. All methods use identical train–test splits, input views, image resolutions, and evaluation protocols. We additionally evaluate D^{2}GS, TWINGS, and VGA-GS with the same cached DA3 depth maps used by our method, retaining their remaining pipelines and settings.

4.1 Comparison with State-of-the-Art Methods

Table 1 shows that our method achieves the best or tied-best results in eight of nine dataset–metric combinations. It leads all three metrics on ScanNet and Replica. On NRGBD, it matches TWINGS in SSIM (0.955) and achieves the lowest LPIPS (0.074), although TWINGS retains the highest PSNR. These results indicate strong structural similarity and perceptual quality across the three benchmarks, with a less uniform advantage in pixel-wise reconstruction accuracy. Under the same DA3 prior, our method still exceeds VGA-GS+DA3 by 0.24 and 0.32 dB in PSNR on ScanNet and Replica, respectively, while reducing LPIPS by 0.009 and 0.016. It also improves all three metrics on NRGBD. Thus, replacing the depth source alone does not reproduce the performance of the complete framework. The component ablations below further examine the contributions of initialization, regularization, and capacity allocation. Fig. 2 presents representative ScanNet results. Compared with several baselines, our method produces more coherent chair frames and door boundaries, with fewer local distortions and discontinuities. It also preserves details on textured surfaces while maintaining cleaner weakly textured regions. This balance is consistent with structural guidance that

Table 2: Training time and Gaussian counts on ScanNet and Replica.

Method	ScanNet		Replica	
	Time (min)↓	#Gaussians↓	Time (min)↓	#Gaussians↓
P^2NeRF	461.80	–	381.63	–
VGA-GS	5.95	128.2K	4.35	68.2K
DropGaussian	**2.08**	201.1K	**1.96**	416.1K
D^2GS	3.38	217.7K	2.97	422.4K
StereoGS	164.24	295.6K	135.09	150.1K
TWINGS	20.88	213.6K	18.96	403.6K
Ours	4.97	172.5K	4.08	208.1K

Table 3: Ablation study on ScanNet.

SAI.	GR	TACS	SSG	PSNR↑	SSIM↑	LPIPS↓	Depth RMSE↓	#Gaussians↓
				19.06	0.703	0.378	0.581	332.0K
✓				21.23	0.753	0.316	0.197	426.5K
✓	✓			21.68	0.771	0.310	0.153	620.2K
✓	✓	✓		21.74	0.777	0.292	0.154	227.3K
✓	✓		✓	21.77	0.781	0.302	0.152	260.8K
✓	✓	✓	✓	**21.92**	**0.786**	**0.287**	**0.151**	**172.5K**

constrains split displacements and differentiates growth according to planar texture. The visual results reflect the combined effects of geometric initialization, capacity allocation, and regularization.

Training Time and Gaussian Counts. Table 2 compares training time and representation size. Our method takes 4.97 and 4.08 minutes on ScanNet and Replica, using 172.5K and 208.1K Gaussians, respectively. It trains faster than VGA-GS and uses fewer Gaussians than DropGaussian, D^2GS, and TWINGS on both datasets. DropGaussian has the shortest reported training time, while VGA-GS uses the fewest Gaussians; StereoGS also uses fewer Gaussians than our method on Replica. Together with Table 1, these results indicate a favorable balance among rendering quality, representation size, and training time. The measurements compare complete training configurations rather than isolating individual components.

4.2 Ablation Studies and Analysis

Table 3 evaluates structure-aware initialization (SAI), geometric regularization (GR), and the two capacity-allocation components on ScanNet. Here, TACS denotes type-calibrated candidate selection, and SSG denotes structural-subspace growth. We first add SAI and GR to the baseline, then enable TACS and SSG separately and jointly from the SAI+GR configuration. SAI increases PSNR from 19.06 to 21.23 dB and reduces depth RMSE from 0.581 to 0.197. Adding GR further improves PSNR to 21.68 dB and depth RMSE to 0.153. These improvements highlight the importance of reliable initialization and geometric regularization under sparse photometric supervision. Starting from SAI+GR, TACS reduces LPIPS from 0.310 to 0.292, while SSG increases SSIM from 0.771 to 0.781. Their depth RMSE values remain close to that of SAI+GR, at 0.154 and 0.152, respectively. Combining both components achieves the best values across all four quality metrics in this ablation: 21.92 dB PSNR, 0.786 SSIM, 0.287 LPIPS, and 0.151 depth RMSE. It also reduces the Gaussian count from 620.2K to 172.5K, a reduction of 72.2%. These results support jointly controlling candidate selection and structural growth to improve rendering quality and representation compactness on top of the geometry established by SAI and GR.

5 Conclusion

In this paper, we presented a structure-guided capacity allocation framework for sparse-view indoor Gaussian splatting. Rather than treating densification as an unconstrained increase in model capacity, our approach assigns persistent structural roles to Gaussian primitives and uses them to determine both which primitives should grow and along which directions their descendants should be introduced. Type-calibrated contribution scoring prioritizes candidates under a shared Top-K budget, while persistent structural priors constrain split directions. Combined with multi-view initialization and geometric regularization, the framework improves novel-view rendering quality and preserves indoor structural details.

AI USE STATEMENT

We used an OpenAI language model to assist with drafting and polishing the manuscript, including improving clarity, grammar, and organization. The authors reviewed and revised the AI-assisted text and take full responsibility for the final content of the paper.

REFERENCES

Dejan Azinović, Ricardo Martin-Brualla, Dan B Goldman, Matthias Nießner, and Justus Thies. Neural RGB-D surface reconstruction. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6280–6291, 2022.

Jiayang Bai, Letian Huang, Jie Guo, Wen Gong, Yuanqi Li, and Yanwen Guo. 360-GS: Layout-guided panoramic gaussian splatting for indoor roaming. In *2025 International Conference on 3D Vision (3DV)*, pp. 1042–1053. IEEE, 2025.

Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 2432–2443. IEEE, 2017.

Xiaobin Deng, Changyu Diao, Min Li, Ruohan Yu, and Duanqing Xu. Improving densification in 3d gaussian splatting for high-fidelity rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 223–232, 2026.

Shuangkang Fang, I-Chao Shen, Xuanyang Zhang, Zesheng Wang, Yufeng Wang, Wenrui Ding, Gang YU, and Takeo Igarashi. Dropping anchor and spherical harmonics for sparse-view gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 33312–33322, June 2026.

Meiying Gu, Jiawei Zhang, Jiahe Li, Xiaohan Yu, Haonan Luo, Jin Zheng, and Xiao Bai. Sparsesurf: Sparse-view 3d gaussian splatting for surface reconstruction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 4311–4319, 2026.

Seungoh Han, Jaehoon Jang, Hyunsu Kim, Jaeheung Surh, Junhyung Kwak, Hyowon Ha, and Kyungdon Joo. Lighthousegs: Indoor structure-aware 3d gaussian splatting for panorama-style mobile captures. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3441–3450, 2026.

Han Huang, Yulun Wu, Chao Deng, Ge Gao, Ming Gu, and Yu-Shen Liu. FatesGS: Fast and accurate sparse-view surface reconstruction using gaussian splatting with depth-feature consistency. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 3644–3652, 2025.

Sergio Izquierdo, Mohamed Sayed, Michael Firman, Guillermo Garcia-Hernando, Daniyar Turmukhambetov, Javier Civera, Oisin Mac Aodha, Gabriel Brostow, and Jamie Watson. Mvsanywhere: Zero-shot multi-view stereo. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11493–11504. IEEE, 2025.

Xirui Jin, Renbiao Jin, Boying Li, Danping Zou, and Wenxian Yu. Planargs: High-fidelity indoor 3d gaussian splatting guided by vision-language planar priors. *Advances in Neural Information Processing Systems*, 38:36342–36370, 2026.

Zeran Ke, Bin Tan, Xianwei Zheng, Yujun Shen, Tianfu Wu, and Nan Xue. Scalelsd: Scalable deep line segment detection streamlined. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6327–6336. IEEE, 2025.

Hyeseong Kim, Geonhui Son, Deukhee Lee, and Dosik Hwang. TWINGS: Thin plate splines warp-aligned initialization for sparse-view gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Haotong Lin, Sili Chen, Junhao Liew, Donny Y Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

Hyunwoo Park, Gun Ryu, and Wonjun Kim. Dropgaussian: Structural regularization for sparse-view gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21600–21609, 2025.

Barbara Roessle, Jonathan T Barron, Ben Mildenhall, Pratul P Srinivasan, and Matthias Nießner. Dense depth priors for neural radiance fields from sparse input views. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12882–12891. IEEE, 2022.

Cong Ruan, Yuesong Wang, Tao Guan, Bin Zhang, and Lili Ju. Indoorgs: Geometric cues guided gaussian splatting for indoor scene reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 844–853, 2025.

Johannes L. Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Meixi Song, Xin Lin, Dizhe Zhang, Haodong Li, Xiangtai Li, Bo Du, and Lu Qi. D^2gs: Depth-and-density guided gaussian splatting for stable and accurate sparse-view reconstruction. In *International Conference on Learning Representations*, 2026.

Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, et al. The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019.

Xiaotian Sun, Qingshan Xu, Xinjie Yang, Yu Zang, and Cheng Wang. Global and hierarchical geometry consistency priors for few-shot nerfs in indoor scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20530–20539, 2024.

Xiaotian Sun, Qingshan Xu, Mengyin Liu, Siqi Shen, and Cheng Wang. Reconstructing sparse-view indoor scenes in view space with global monocular prior alignment. *IEEE Transactions on Circuits and Systems for Video Technology*, 2026.

Bin Tan, Rui Yu, Yujun Shen, and Nan Xue. Planarsplatting: Accurate planar surface reconstruction in 3 minutes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1190–1199, 2025.

Matias Turkulainen, Xuqian Ren, Iaroslav Melekhov, Otto Seiskari, Esa Rahtu, and Juho Kannala. Dn-splatter: Depth and normal priors for gaussian splatting and meshing. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2421–2431, 2025.

Haodong Xiang, Xinghui Li, Kai Cheng, Xiansong Lai, Wanting Zhang, Zhichao Liao, Long Zeng, and Xueping Liu. Gaussianroom: Improving 3d gaussian splatting with sdf guidance and monocular cues for indoor scene reconstruction. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2686–2693. IEEE, 2025.

Yexing Xu, Longguang Wang, Minglin Chen, Sheng Ao, Li Li, and Yulan Guo. Dropoutgs: Dropping out gaussians for better sparse-view rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 701–710, 2025.

Jiawei Yang, Marco Pavone, and Yue Wang. Freenerf: Improving few-shot neural rendering with free frequency regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8254–8263, 2023.

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.

Wenhao Yuan, Yiyuan Ge, and Deli Cai. StereoGS: Sparse-view 3d gaussian splatting via stereo priors. In *European Conference on Computer Vision*, 2026.

Baowen Zhang, Chuan Fang, Rakesh Shrestha, Yixun Liang, Xiao-Xiao Long, and Ping Tan. RaDe-GS: Rasterizing depth in gaussian splatting. *ACM Transactions on Graphics*, 45(2):1–14, 2026.

Jiawei Zhang, Jiahe Li, Xiaohan Yu, Lei Huang, Lin Gu, Jin Zheng, and Xiao Bai. Cor-gs: Sparse-view 3d gaussian splatting via co-regularization. In *European Conference on Computer Vision*, pp. 335–352. Springer, 2024a.

Wanting Zhang, Haodong Xiang, Zhichao Liao, Xiansong Lai, Xinghui Li, and Long Zeng. 2dgs-room: Seed-guided 2d gaussian splatting with geometric constrains for high-fidelity indoor scene reconstruction. *arXiv preprint arXiv:2412.03428*, 2024b.

SUPPLEMENTARY MATERIAL

This supplement provides implementation details, additional ablations, and further quantitative and qualitative evaluations of the proposed framework.

A IMPLEMENTATION AND EVALUATION DETAILS

A.1 TRAINING CONFIGURATION

We optimize each scene for 10K iterations using Adam, progressively increasing the spherical-harmonic degree from 0 to 3. All experiments are conducted on a single NVIDIA GeForce RTX 3090 GPU with 24 GB of memory. Type-calibrated ADC is performed every 100 iterations from iteration 600 to 9,900, using an absolute-gradient threshold of 6×10^{-4}. Unless otherwise specified, the shared per-step growth budget is $K = 2000$ for ScanNet and $K = 3000$ for Replica and NRGBD. All eligible structural types compete under this shared budget without type-specific quotas.

For the geometric terms in Eq. (14), we set $\lambda_r = 1.0$, $\lambda_n = 0.05$, and $\lambda_p = 1.0$. The co-planarity loss is activated after $5{,}000$ iterations.

A.2 EVALUATION PROTOCOL

We compute PSNR, SSIM, and LPIPS following the 3DGS evaluation protocol adopted by Drop-Gaussian (Park et al., 2025) and StereoGS (Yuan et al., 2026), using the VGG backbone for LPIPS. Image metrics are evaluated on each held-out test view and averaged over the test views of each scene. Depth RMSE follows VGA-GS (Sun et al., 2026) and is computed over pixels with valid ground-truth depth. We follow the scene selections and standard train–test splits of VGA-GS, as summarized in Sec. 4.

A.3 MULTI-VIEW DEPTH INFERENCE

We use Depth Anything 3 (DA3) (Lin et al., 2025) with the `da3nested-giant-large` model for multi-view depth inference. All sparse training images are processed jointly with their calibrated camera intrinsics and poses. The supplied camera-to-world poses are inverted to obtain the world-to-camera extrinsics expected by DA3. We use a processing resolution of 504 pixels with the `upper_bound_resize` strategy and resize the predicted depth maps to the training-image resolution using bilinear interpolation. Our pipeline applies no additional per-view scale or shift alignment to these predictions. The same inference configuration is used across datasets.

A.4 GLOBAL PLANE ASSOCIATION AND TEXTURE CLASSIFICATION

Cross-view plane association. We first obtain per-view planar masks and semantic labels using LP3 (Jin et al., 2026). Planar pixels with valid predicted depth are back-projected into the world coordinate system, and each planar observation is fitted using principal component analysis (PCA). The covariance eigenvector with the smallest eigenvalue defines the plane normal.

We then associate observations across views according to semantic compatibility and 3D geometric consistency. Associated observations must have a normal difference below 10° and a plane-offset difference below $\max(0.08\,\mathrm{m}, 0.006 D_{\mathrm{scene}})$, together with consistent spatial overlap or centroid proximity. Here, D_{scene} is the scene-scale quantity used by the association procedure. Compatible observations share a global plane identity. To recover missed floor regions, unassigned pixels with valid depth are associated with the nearest eligible global floor plane when their point-to-plane distance is below $0.05\,\mathrm{m}$.

Per-observation texture classification. Let $g(\mathbf{p})$ denote the RGB-gradient magnitude at a planar pixel $\mathbf{p}$. We compute a robust threshold over planar pixels:

$$\tau_c = \max(0.03,\, \mathrm{median}(g) + 2.0 \times 1.4826 \times \mathrm{MAD}(g))\,, \tag{15}$$

Table 4: Depth-prior sensitivity on ScanNet and Replica. Only the predictor used for initialization and geometric regularization is changed. Best results are boldfaced.

	ScanNet			Replica		
Depth Prior	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
MVSAnywhere	21.86	**0.791**	**0.281**	33.10	0.954	0.0794
DA3	**21.92**	0.786	0.287	**33.50**	**0.958**	**0.0780**

where MAD is the median absolute deviation. For each planar observation, gradient magnitudes are clipped at its 95th percentile to limit the influence of isolated strong edges. We compute the mean clipped magnitude as the texture score and the fraction of pixels with gradient magnitude above τ_c as the texture density. An observation is classified as high-texture when

$$\text{score} > 0.02 \quad \text{and} \quad \text{density} > 0.05. \tag{16}$$

Observations containing fewer than 300 valid planar pixels are classified as low-texture to avoid decisions based on unstable statistics.

Consistent texture roles across views. We aggregate the observation-level decisions within each global plane using pixel-count-weighted voting. A plane is assigned the high-texture role when the weighted proportion of high-texture votes is at least 0.5. All observations of that plane share the resulting label. This association links initial sampling density to a persistent growth policy: high-texture planar Gaussians remain eligible for tangent-plane splitting, whereas low-texture planar Gaussians are excluded from densification. The latter remain subject to continuous parameter optimization and geometric regularization.

A.5 Line Priors and Structural Sampling

Line extraction and lifting. ScaleLSD (Ke et al., 2025) segment coordinates are rescaled to the training-image resolution. We retain segments with detector support score $s_j \geq \tau_s = 5.0$ and image-space length $L_j \geq \tau_\ell = 15$ pixels. Each retained segment is sampled at a target spacing of $\Delta_\ell = 3$ pixels, yielding $M_j = \max(\lceil L_j/\Delta_\ell \rceil + 1,\, 2)$ samples. To lift the image-space direction, we use symmetric offsets with a half-baseline of $\delta = 2$ pixels and bilinearly sample the DA3 depth at both locations. Offset locations are clamped to the image boundary; an invalid offset depth is replaced by the depth at the segment sample.

Line-aligned initialization. When a valid DA3-derived surface normal is unavailable, the source-view direction provides the reference normal for constructing the line-aligned frame. The initial scale multipliers are $(\gamma_\parallel, \gamma_\perp, \gamma_n) = (2.0, 0.75, 0.15)$ along the line, cross-line, and normal axes of this frame, respectively. The stored line tangent subsequently defines the admissible split direction independently of the learned covariance orientation.

Disjoint structural sampling. The rasterized line mask is dilated by one pixel. An additional two-pixel dilation defines the line neighborhoods excluded from planar and ordinary sampling. We use grid strides of 6 pixels for high-texture planar regions, 20 pixels for low-texture planar regions, and 30 pixels for ordinary regions. The line thresholds, scale multipliers, mask dilations, and grid strides are shared across all datasets.

B Depth-Prior Analysis

B.1 Common-Depth-Prior Baseline Comparisons

For the D^{2}GS, TWINGS, and VGA-GS variants in Table 1, we replace the original depth predictions with the same cached DA3 maps used by our method. Each baseline otherwise retains its original pipeline and settings, including its procedures for incorporating depth. In particular, TWINGS retains its initialization and alignment procedure, while VGA-GS retains its initialization and depth-alignment pipeline. Optimization, regularization, and densification settings remain unchanged within each original/DA3 pair.

Table 5: Sensitivity to the shared per-step Top-K growth budget on ScanNet. We report reconstruction quality and the average final Gaussian count. The default is $K = 2000$.

K	Depth RMSE↓	PSNR↑	SSIM↑	LPIPS↓	#Gaussians
500	0.151	21.87	0.7850	0.3039	**90.9K**
1,000	0.152	**21.92**	0.7855	0.2950	115.9K
2,000	0.151	**21.92**	**0.7858**	0.2873	172.5K
3,000	**0.150**	21.88	0.7839	0.2842	234.9K
4,000	0.153	21.85	0.7826	**0.2836**	299.0K

Table 6: Effect of the contribution-modulation parameter λ on ScanNet and Replica. Setting $\lambda = 1$ recovers gradient-only ranking, $S_i = g_i$. The default is $\lambda = 0.5$, and all other settings are fixed. Best results are shown in bold.

	ScanNet				Replica			
λ	PSNR↑	SSIM↑	LPIPS↓	Depth RMSE↓	PSNR↑	SSIM↑	LPIPS↓	Depth RMSE↓
0.3	**21.94**	**0.786**	0.288	0.155	**33.530**	**0.958**	0.079	0.069
0.5	21.92	**0.786**	**0.287**	**0.151**	33.504	**0.958**	**0.078**	**0.067**
0.7	21.83	0.784	0.289	0.155	33.482	**0.958**	**0.078**	0.069
1.0	21.79	0.778	0.289	0.153	33.443	0.957	0.079	0.071

Table 7: Ablation study on Replica. SAI: structure-aware initialization; GR: geometric regularization; TACS: type-calibrated candidate selection; SSG: structural-subspace growth. Best results in each column are boldfaced.

SAI	GR	TACS	SSG	PSNR↑	SSIM↑	LPIPS↓	Depth RMSE↓	#Gaussians↓
				31.57	0.941	0.085	0.258	473.9K
✓				33.08	0.952	0.091	0.086	203.5K
✓	✓			32.88	0.952	0.095	0.082	231.1K
✓	✓	✓		33.22	0.956	**0.071**	0.093	289.9K
✓	✓		✓	33.12	0.953	0.099	**0.064**	**118.4K**
✓	✓	✓	✓	**33.50**	**0.958**	0.078	0.067	208.1K

These paired comparisons assess the effect of changing the depth source within each baseline. Comparisons with our method use a common depth source while retaining differences in initialization, structural priors, and growth policies. The component and growth-budget ablations provide complementary evidence about these design choices.

B.2 Sensitivity to Different Depth Priors

Table 4 replaces DA3 with MVSAnywhere (Izquierdo et al., 2025) for initialization and geometric regularization while keeping the remaining settings fixed. On ScanNet, DA3 yields slightly higher PSNR (21.92 vs. 21.86 dB), whereas MVSAnywhere achieves better SSIM (0.791 vs. 0.786) and LPIPS (0.281 vs. 0.287). On Replica, DA3 performs better across all three metrics, achieving 33.50 dB PSNR and 0.0780 LPIPS, compared with 33.10 dB and 0.0794 for MVSAnywhere. These results demonstrate compatibility with both tested predictors, while showing that reconstruction quality remains sensitive to the geometric prior.

C Additional Ablation Studies

C.1 Sensitivity to the Shared Growth Budget

Table 5 examines the shared per-step Top-K growth budget on ScanNet. Increasing K from 500 to 4,000 raises the average final Gaussian count from 90.9K to 299.0K and consistently improves LPIPS from 0.3039 to 0.2836. PSNR varies only within 21.85–21.92 dB, while SSIM peaks at

$K = 2000$ and depth RMSE changes non-monotonically. Larger budgets therefore primarily benefit perceptual quality in this experiment, with diminishing returns as representation size increases.

The default $K = 2000$ achieves joint-best PSNR (21.92 dB) and the highest SSIM (0.7858) with 172.5K Gaussians. Doubling the budget to 4,000 further reduces LPIPS by 0.0037 but increases the Gaussian count to 299.0K, with lower PSNR and SSIM. These results support $K = 2000$ as a balance between rendering quality and representation size on ScanNet.

C.2 Effect of Contribution Modulation

Table 6 evaluates the contribution-modulation parameter in $S_i = g_i[\lambda + (1-\lambda)c_i]$, where c_i denotes the rendering contribution calibrated within each structural type. Setting $\lambda = 1$ recovers gradient-only ranking, while smaller values increase the influence of relative rendering participation. All other settings, including the shared Top-K growth budget, remain fixed.

Compared with gradient-only ranking, the default $\lambda = 0.5$ improves all four metrics on both datasets. On ScanNet, PSNR increases from 21.79 to 21.92 dB, while LPIPS and depth RMSE decrease from 0.289 to 0.287 and from 0.153 to 0.151, respectively. On Replica, PSNR increases from 33.443 to 33.504 dB and depth RMSE decreases from 0.071 to 0.067, accompanied by improved SSIM and LPIPS. These results support incorporating relative rendering participation into gradient-based candidate selection. This comparison evaluates the combined effect of contribution modulation and within-type calibration.

The parameter sweep reveals similar trade-offs across both datasets. The setting $\lambda = 0.3$ achieves the highest PSNR, whereas $\lambda = 0.5$ yields the lowest depth RMSE and the best or tied-best LPIPS. SSIM is tied between these two settings at the reported precision. Increasing λ from 0.5 to 0.7 lowers PSNR and increases depth RMSE on both datasets, without improving SSIM or LPIPS. These results support $\lambda = 0.5$ as a balanced default for appearance quality and geometric accuracy.

C.3 Ablation Study on Replica

Table 7 extends the main-paper component ablation to Replica. We evaluate structure-aware initialization (SAI), geometric regularization (GR), type-calibrated candidate selection (TACS), and structural-subspace growth (SSG), enabling TACS and SSG separately and jointly from the SAI+GR configuration. SAI increases PSNR from 31.57 to 33.08 dB and reduces depth RMSE from 0.258 to 0.086 with fewer Gaussians, although LPIPS increases from 0.085 to 0.091. Adding GR further reduces depth RMSE to 0.082, with small decreases in PSNR and perceptual quality. These results show the geometric benefits of initialization and regularization, together with their trade-offs in appearance fitting. Starting from SAI+GR, TACS improves PSNR to 33.22 dB and achieves the lowest LPIPS of 0.071, while increasing depth RMSE and the Gaussian count. SSG instead yields the lowest depth RMSE (0.064) and the fewest Gaussians (118.4K), but increases LPIPS to 0.099. Candidate selection and structural growth therefore affect appearance, geometry, and representation size differently.

Combining both components achieves the highest PSNR and SSIM, at 33.50 dB and 0.958. Compared with SAI+GR, the full model improves all four quality metrics while reducing the Gaussian count from 231.1K to 208.1K. It also improves all four metrics over SAI alone with similar final counts (208.1K vs. 203.5K). These results support their combined use for balanced reconstruction quality, although individual configurations attain lower LPIPS, depth RMSE, or Gaussian counts.

C.4 Qualitative Ablation Study

Figure 3 presents a cumulative ablation with alternating RGB and depth rows. SAI reduces the severe distortions of the base configuration and establishes more coherent geometric support. GR further stabilizes the recovered depth, after which SSG and TACS are introduced successively to constrain structural growth and prioritize candidate selection. The full configuration combines these mechanisms to improve local appearance while retaining coherent depth structure. The figure illustrates their cumulative effect; the main-paper ablation and Table 7 also evaluate TACS and SSG separately from the same SAI+GR starting configuration.

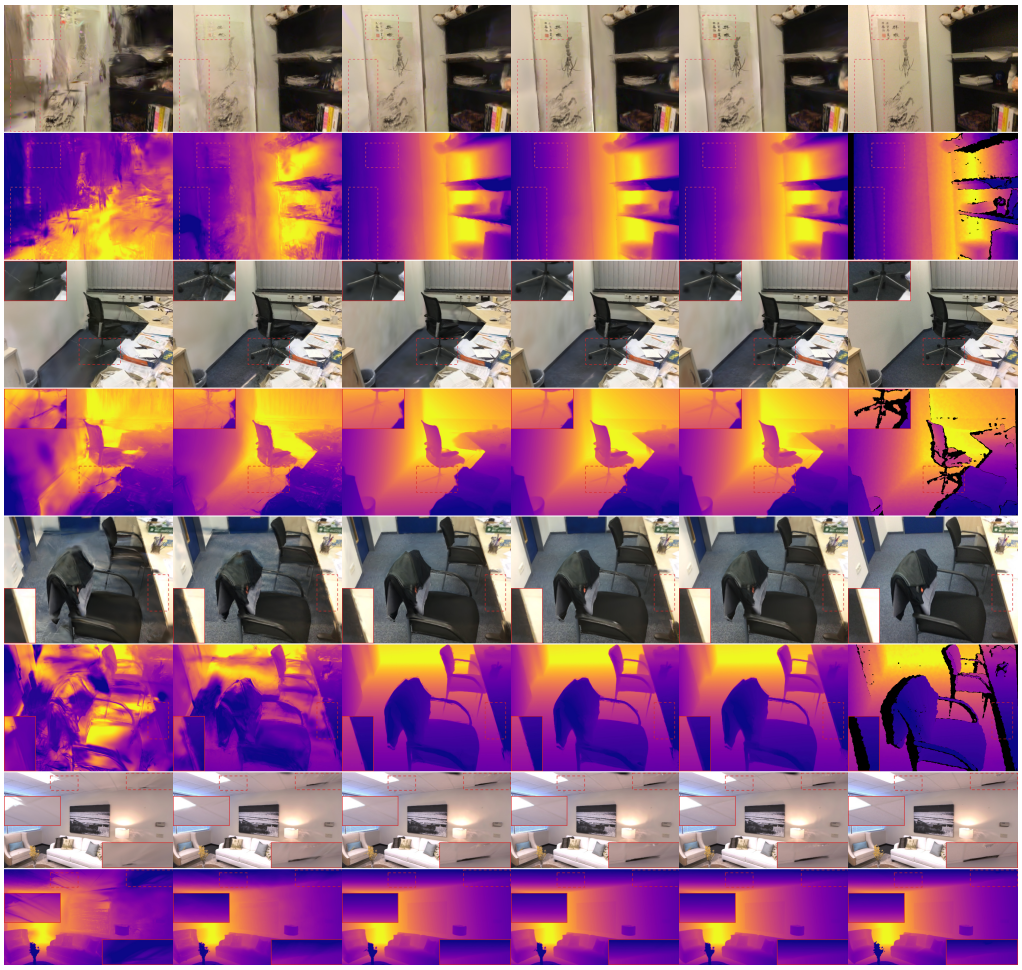


Figure 3: Representative qualitative ablation on ScanNet and Replica. RGB renderings and depth maps are shown in alternating rows. Components are added cumulatively from (a) to (e), with (e) corresponding to the full model.

D Additional Reconstruction Results

D.1 Reconstruction with Varying Numbers of Input Views

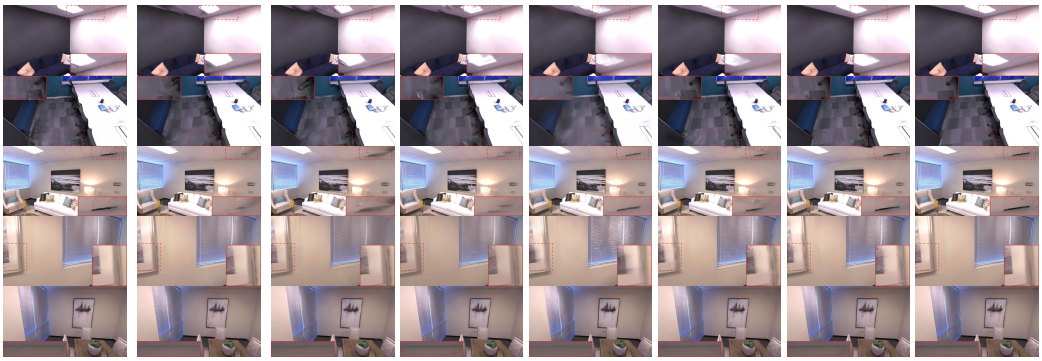
Table 8 evaluates Replica with 10, 15, 25, and 30 training views while keeping the test protocol fixed. Our method achieves the best or tied-best PSNR, SSIM, and LPIPS at every evaluated view count. Its advantage also holds over VGA-GS+DA3, which uses the same depth source. With 10 views, our method improves PSNR from 26.99 to 27.59 dB and SSIM from 0.904 to 0.916, while reducing LPIPS from 0.143 to 0.116. With 15 views, it improves PSNR by 0.70 dB and reduces LPIPS by 0.021 relative to the same baseline.

All evaluated methods lose quality as training views become sparser. Our method retains a relative advantage, particularly in perceptual quality at 10 and 15 views, supporting the effectiveness of the complete framework under limited observations. Figures 6, 7, 8, and 9 provide corresponding visual comparisons. Severe observation gaps nevertheless remain challenging, as structural guidance cannot fully recover unobserved geometry and appearance.



(a) DropGaussian (b) D^{2}GS (c) TWINGS (d) StereoGS (e) VGA-GS (f) Ours (g) GT

Figure 4: Qualitative comparison of novel-view renderings on Neural RGB-D.



(a) P^{2}NeRF (b) DropGaussian (c) D^{2}GS (d) TWINGS (e) StereoGS (f) VGA-GS (g) Ours (h) GT

Figure 5: Qualitative comparison of novel-view renderings on Replica.

Table 8: Quantitative comparison on Replica with varying numbers of training views. The test protocol is kept fixed across all settings. Best results are boldfaced.

Method	30 Views			25 Views			15 Views			10 Views		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
D^{2}GS	34.71	0.964	**0.066**	33.34	0.958	0.075	28.83	0.918	0.123	25.03	0.870	0.177
TWINGS	35.12	**0.967**	**0.066**	34.16	0.963	0.073	29.94	0.929	0.113	25.93	0.876	0.178
StereoGS	32.04	0.941	0.144	31.59	0.938	0.147	29.14	0.917	0.163	26.51	0.889	0.184
VGA-GS	33.72	0.951	0.104	32.61	0.944	0.111	28.67	0.905	0.148	25.56	0.869	0.178
VGA-GS + DA3	35.19	0.964	0.076	34.27	0.960	0.083	30.44	0.934	0.114	26.99	0.904	0.143
Ours	**35.27**	**0.967**	**0.066**	**34.58**	**0.964**	**0.072**	**31.14**	**0.943**	**0.093**	**27.59**	**0.916**	**0.116**

D.2 Additional Novel-View Rendering Comparisons

Figure 4 and Figure 5 complement the main-paper ScanNet examples with results on Replica and Neural RGB-D. The examples include extended planar surfaces, object boundaries, and thin furniture structures. Our renderings show clearer boundaries and more coherent local appearance in the illustrated regions, consistent with the quantitative gains in structural similarity and perceptual quality. These comparisons illustrate the combined effects of structural initialization, capacity allocation, and geometric regularization.

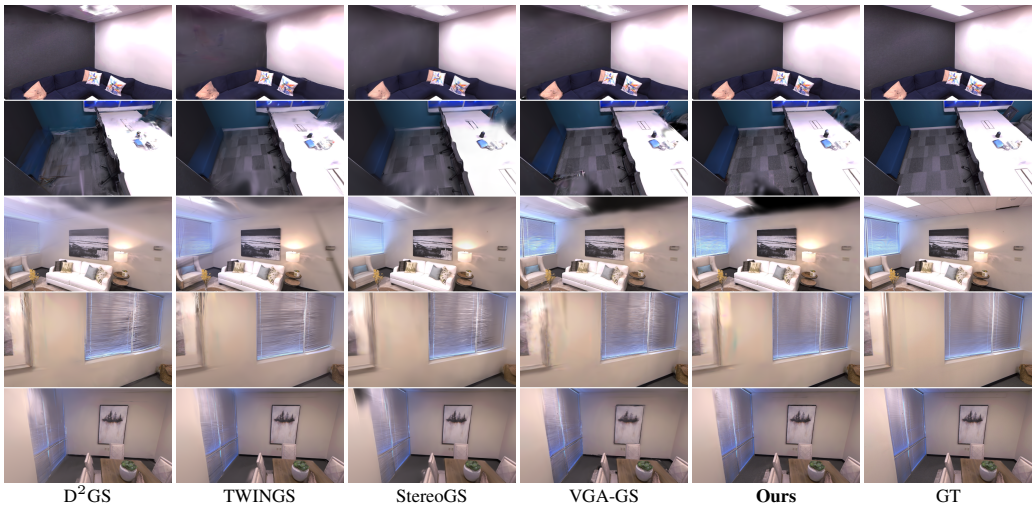


Figure 6: Qualitative comparison on Replica using 10 input views.

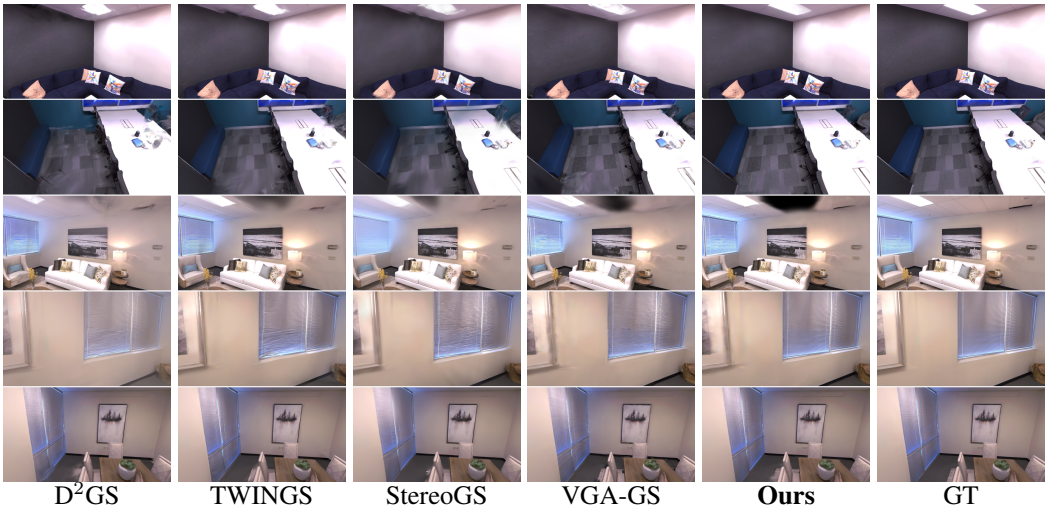


Figure 7: Comparison on Replica with 15 input views.

D.3 Qualitative Evaluation of Geometric Reconstruction

To further evaluate geometric reconstruction quality, we provide additional qualitative mesh comparisons in Fig. 10. For a fair comparison, we first render depth maps for PlanarGS, VGA-GS, and our method, and then fuse them into meshes using the same TSDF fusion implementation adopted from VGA-GS, with identical fusion parameters across all methods. For Replica, the reference meshes are reconstructed from the ground-truth depth maps using the same TSDF fusion pipeline, while for ScanNet we directly use the ground-truth meshes provided by the dataset.

As shown in Fig. 10, our method better preserves high-frequency geometric details, including thin furniture structures and object boundaries, while recovering smoother and more coherent planar surfaces. Compared with PlanarGS and VGA-GS, the detailed views show clearer structural contours, fewer spurious surface irregularities, and fewer floating fragments. These improvements yield reconstructions that are visually closer to the GT references in both local detail and planar structure. The results complement the novel-view rendering comparisons and are consistent with the combined use of structure-aware initialization, structure-guided capacity allocation, and geometric regularization.

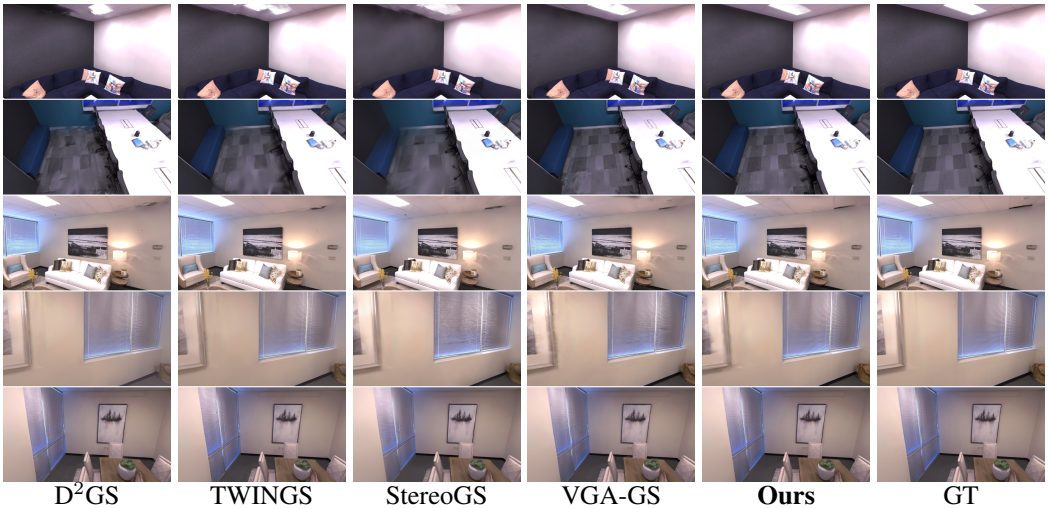


Figure 8: Comparison on Replica with 25 input views.

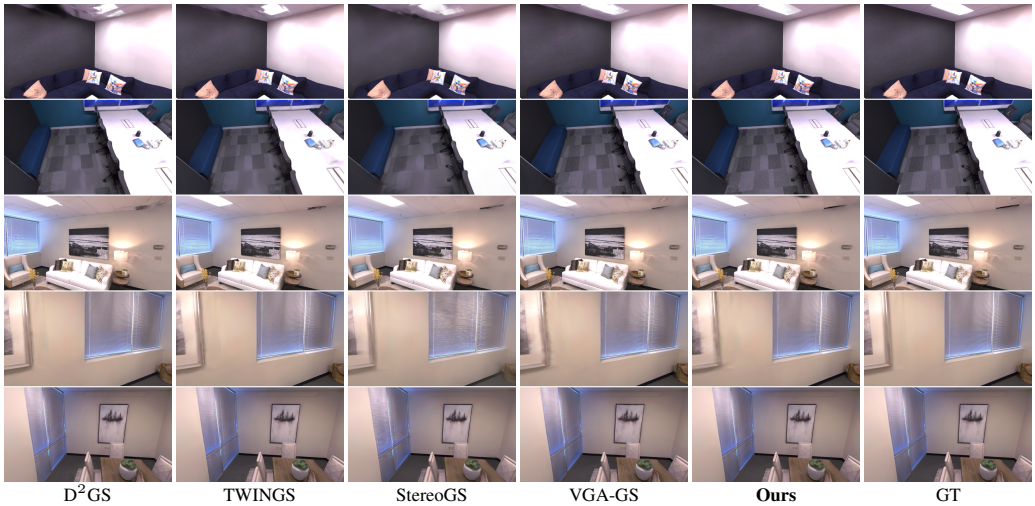


Figure 9: Qualitative comparison on Replica using 30 input views.

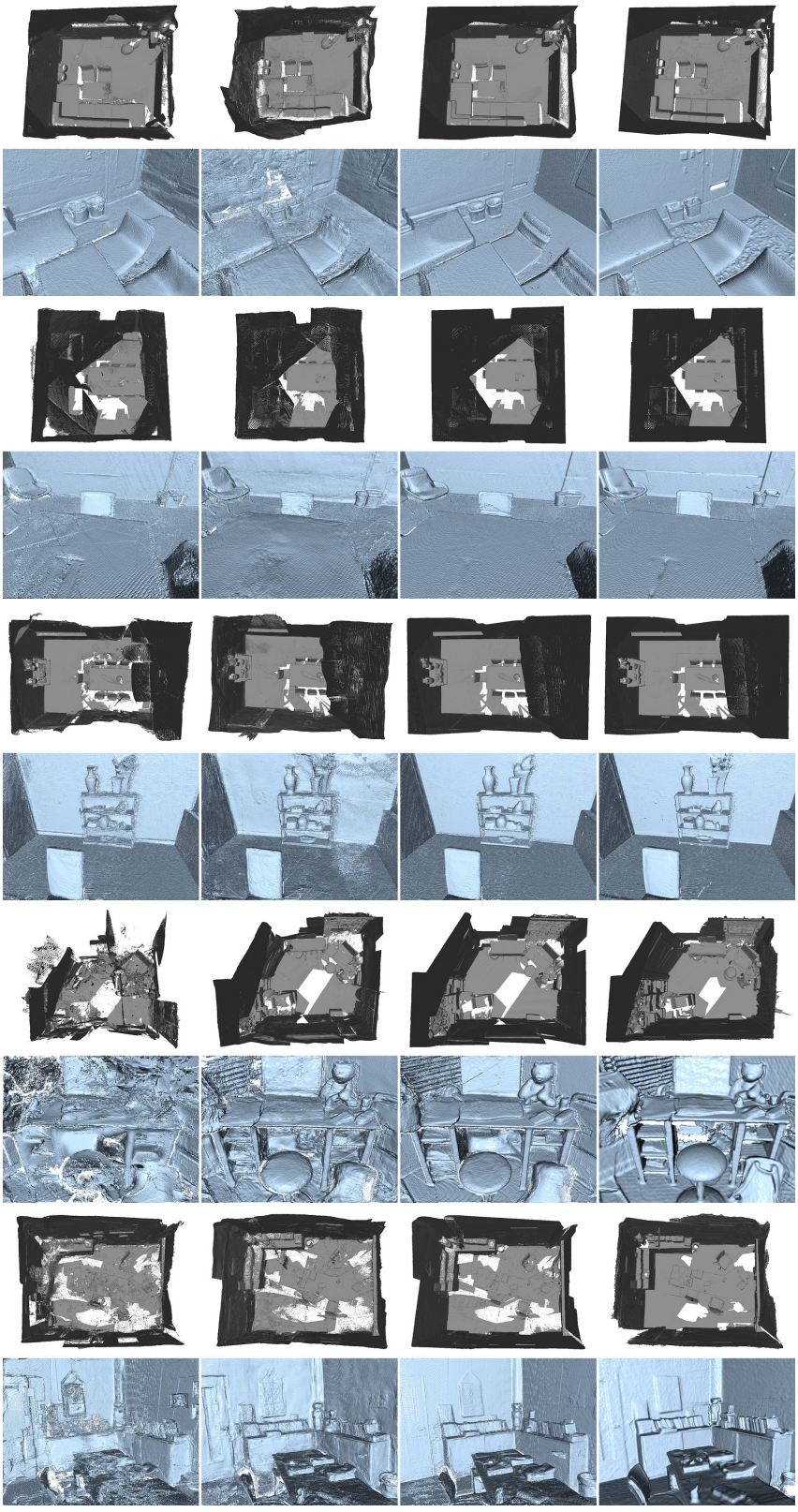


Figure 10: Qualitative comparison of mesh reconstruction on Replica and ScanNet. From top to bottom: Replica *Office 0*, Replica *Office 4*, Replica *Room 2*, ScanNet *Scene 0710-00*, and ScanNet *Scene 0758-00*. For each scene, the upper row shows an overview, and the lower row provides a corresponding close-up view.