SCENE-SAM3D: SEEING MORE OF THE WHOLE VIA VIEW SELECTION AND CONFLICT-AWARE FUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

High-quality 3D scene assets are critical for embodied applications such as robotic manipulation, navigation, and simulation. Despite their strong object priors, recent single-image 3D generation models such as SAM3D remain insufficient for real-world scenes, where severe occlusions, redundant observations, and cross-view inconsistencies make reliable scene generation challenging. To help SAM3D *see more of the whole*, we introduce **Scene-SAM3D**, a training-free framework that extends SAM3D from single-view object generation to calibrated multi-view scene asset generation. For faithful assets from observations of varying quality, Scene-SAM3D selects a compact set of complementary views, reducing redundancy while providing additional evidence for regions occluded in individual views. Based on the selected views, it performs step-efficient latent velocity fusion to integrate multi-view evidence and suppress cross-view conflicts in canonical space. Finally, a lightweight rigid-object Gaussian optimization refines the scene layout within 200 iterations while preserving the generated object geometry. Experiments on Replica and ScanNet++ demonstrate consistent improvements at both instance and scene levels, with our method reducing scene-level CD by 43.8% on Replica and 30.9% on ScanNet++, while cutting flow-model sampling FLOPs and wall-time latency by nearly 20% under the same multi-view setting. Our anonymous project page is available at `https://shorturl.at/zSHiQ`.

1 INTRODUCTION

Faithful and spatially coherent 3D scene assets are essential for realistic digital environments, with applications in immersive AR/VR (Jones et al., 2014; Orts-Escolano et al., 2016; Newcombe et al., 2011; Sra et al., 2018; Xu et al., 2023; Huang et al., 2025b) and embodied simulation (Wang et al., 2024a; Gu et al., 2023; Khanna et al., 2024; Puig et al., 2024; Ma et al., 2023; Majumdar et al., 2024). A practical way to obtain such assets is to reconstruct or generate 3D meshes from 2D visual observations. Recent foundation models (Xiang et al., 2025; Zhao et al., 2025; Li et al., 2025; Chen et al., 2026) bridge this gap in part by turning a *single* natural image into a structured object-centric prediction of shape, appearance, and camera-relative pose, with SAM3D (Chen et al., 2026) as a prominent example. Yet realistic scenes are rarely observed from a single view. Calibrated captures instead produce long multi-view sequences with abundant redundancy, pervasive occlusion, and no obvious "best" view for interpreting each instance ahead of inference.

Generating scene assets under these conditions is therefore a genuinely *multi-view* problem: for each foreground object we seek a complete mesh from complementary multi-view observations, but we simultaneously require that all instances be placed compatibly into a shared world frame consistent with multi-view evidence. To support multi-view conditioning, two common strategies are stochastic view conditioning and multi-view velocity fusion. Stochastic view conditioning (Xiang et al., 2025), which samples different views during generation, provides a simple way to use multiple observations but does not explicitly reconcile them across views. Conversely, multi-view velocity methods such as MultiDiffusion (Bar-Tal et al., 2023), which averages velocity fields induced by multiple views at each step, usually improve reconstruction quality over stochastic conditioning (Chang et al., 2026). However, multi-view generation may introduce *geometry conflicts*. Even with multiple observations, incomplete coverage can leave missing geometry or induce hallucinated structures in unseen regions. Multi-view fusion also implicitly assumes that per-view updates are compatible in a shared canonical

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

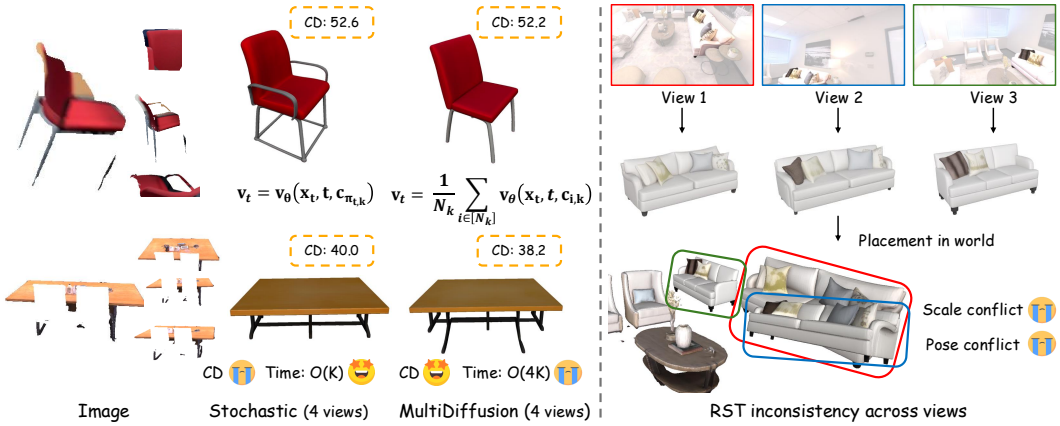


Figure 1: **Our preliminary study on extending SAM3D with multi-view fusion.** Left: stochastic conditioning is efficient but unstable, whereas MultiDiffusion improves geometry at a higher cost. Right: different views produce different predictions of rotation, scale, and translation.

space. When this assumption fails, direct averaging can produce duplicated or layered structures. For example, Fig. 1 (left) shows a hallucinated chair base and duplicated table legs.

Multi-view inference also introduces *pose conflicts*. For the same instance, different views may produce substantially different predictions of rotation, scale, and translation. Once transformed into the shared world frame, these discrepancies lead to incompatible object sizes, orientations, and placements, as shown in Fig. 1 (right). More importantly, the view most favorable for mesh quality is often not the most reliable for downstream layout estimation. This mismatch motivates decoupling object-shape reconstruction from scene-layout refinement.

These observations suggest that multi-view scene generation should be treated as a structured **selection–fusion–alignment** problem rather than as direct multi-image conditioning. First, redundant frames should be filtered out while retaining views that provide complementary 3D evidence. Second, multi-view fusion in a flow-based generator must account for canonical-space disagreement, since uniformly averaging view-conditioned velocity updates can mix incompatible shape hypotheses. Third, after object-level reconstruction, scene-level layout can benefit from the full trajectory of calibrated views, which provides stronger constraints on object placement than any single reconstruction view alone. Together, these considerations motivate a pipeline that selects complementary views, fuses them with conflict awareness, and refines object layout through calibrated multi-view feedback.

We instantiate this design in **Scene-SAM3D**, a training-free pipeline built on frozen SAM3D. Following the selection–fusion–alignment decomposition, our method first performs *occupancy-coverage view selection*, which selects an anchor view and complementary helper views by measuring newly observed object occupancy in the shared world frame. The selected views are then integrated by *step-efficient conflict-aware multiview fusion*, which concentrates multi-view fusion in early sampling steps and suppresses incompatible velocity updates through anchor-guided sign consensus. Finally, *rigid-object gaussian layout refinement* fixes the reconstructed object assets and optimizes their similarity transforms with calibrated multi-view rendering feedback, improving scene-level spatial coherence without modifying the generated meshes.

Without finetuning SAM3D, Scene-SAM3D consistently outperforms single-view and MultiDiffusion fusion baselines. At the instance level, it reduces Chamfer Distance (CD) by 5.36% over single-view SAM3D on Replica and improves Completeness by 6.98% on ScanNet++; compared with SAM3D using MultiDiffusion fusion, it further reduces CD by 3.93%/1.97% on Replica/ScanNet++ while saving nearly 20% flow-model FLOPs and wall-time latency under ten-view inference. At the scene level, it reduces CD by 43.75%/30.91% over SAM3D with ICP alignment on Replica/ScanNet++.

Our contributions are summarized as follows.

- We formulate calibrated multi-view scene asset generation as a coupling between instance-centric flow estimation and globally consistent placement, motivating selection, fusion, and alignment stages that interoperate cleanly with pretrained single-view generators.
- We introduce Scene-SAM3D, a training-free method combining occupancy-based anchor–helper selection for 3D-complementary redundancy reduction, early-stop multi-view latent fusion with

anchor-guided sign consensus to mitigate canonical-space conflicts, and rigid-object Gaussian layout refinement for calibrated multi-view silhouette alignment without editing generated meshes.

- We evaluate on Replica and ScanNet++, demonstrating consistent gains in object-level metrics and stronger scene-scale reconstruction when refining poses with Gaussian objectives, alongside reduced flow-model inference steps relative to exhaustive multi-view velocity fusion.

2 Preliminaries and Problem Setup

SAM3D. Before introducing our multi-view extension, we briefly review the base model SAM3D. SAM3D adopts a two-stage flow-matching architecture, where each stage consists of a conditional flow model followed by a task-specific decoder. For each object k, given an input RGB image $\mathbf{I}$ and an instance mask $\mathbf{M}$, the Stage 1 Geometry Model defines a conditional generative process $p(\mathbf{O}_k, \mathbf{T}_k^{w\leftarrow o} \mid \mathbf{I}, \mathbf{M})$ for coarse shape and object placement. At inference time, it generates a canonical occupancy grid $\mathbf{O}_k \in \mathbb{R}^{64\times64\times64}$ and predicts an object-to-world transform $\mathbf{T}_k^{w\leftarrow o} = (\mathbf{R}_k, \mathbf{s}_k, \mathbf{t}_k)$, where $\mathbf{R}_k$, $\mathbf{s}_k$, and $\mathbf{t}_k$ denote the rotation, scale, and translation of object k, respectively. Given a calibrated camera pose C_i, the predicted camera-relative transform can be lifted to the world frame. Conditioned on $\mathbf{O}_k$, the Stage 2 Texture and Refinement Model defines $p(\mathbf{Z}_k \mid \mathbf{I}, \mathbf{M}, \mathbf{O}_k)$ and generates a refined asset representation $\mathbf{Z}_k$ that encodes refined geometry and texture. Task-specific decoders then convert $\mathbf{Z}_k$ into the final textured asset $\mathcal{G}_k$, such as a Gaussian or mesh representation.

Problem Definition. We consider *scene-level* asset generation from a calibrated video sequence containing K foreground objects and N views. For each object k, we define its object-centric observation set as $\mathcal{D}_k = \{(\mathbf{I}_i, \mathbf{M}_i, \mathbf{C}_i, \mathbf{K})\}_{i=1}^{N_k}$, where N_k denotes the number of valid observations, and $\mathbf{I}_i$, $\mathbf{M}_i$, $\mathbf{C}_i$, and $\mathbf{K}$ denote the RGB view, instance mask, camera-to-world pose, and shared camera intrinsics, respectively. Given these observations, our **goal** is to generate a complete scene asset $\mathcal{S} = \{(\mathcal{G}_k, \mathbf{T}_k^{w\leftarrow o})\}_{k=1}^{K}$ and estimate its placement in the world coordinate system. We denote the object-to-world transform as $\mathbf{T}_k^{w\leftarrow o} = (\mathbf{R}_k, \mathbf{s}_k, \mathbf{t}_k)$, where $\mathbf{R}_k \in \mathrm{SO}(3)$, $\mathbf{s}_k \in \mathbb{R}_+^3$, and $\mathbf{t}_k \in \mathbb{R}^3$. Compared with isolated instance generation, scene-level generation is substantially more demanding: objects are observed through *partial* and often *occluded* views, requiring multi-view evidence for completion, while all recovered assets must be *placed consistently* in a shared world frame.

Multiview Instance Mesh Generation. Modern image-to-3D models (Xiang et al., 2025; Chen et al., 2026) reconstruct plausible geometry and texture from a single image, but effectively leveraging multiple observations remains underexplored. A simple extension is ❶ *Stochastic view conditioning* (Xiang et al., 2025). For a flow-based generator with conditional velocity $\mathbf{v}_\theta$ and intermediate 3D latent state $\mathbf{x}_t$, one may sample a valid view $\pi_{t,k} \sim q(\cdot \mid \mathcal{D}_k)$ at each generation step and update the latent state by $\mathbf{x}_{t+\Delta t} = \mathbf{x}_t + \Delta t \cdot \mathbf{v}_\theta(\mathbf{x}_t, t, \mathbf{c}_{\pi_{t,k}})$, where $\mathbf{c}_{\pi_{t,k}}$ denotes the corresponding condition. This strategy is computationally cheap because each step uses only *one* view condition. Yet it does not explicitly aggregate cross-view evidence, so the generation trajectory may be driven by incompatible single-view interpretations across timesteps. A more direct alternative is ❷ *MultiDiffusion velocity fusion* (Bar-Tal et al., 2023), where all selected views contribute to the update through $\mathbf{x}_{t+\Delta t} = \mathbf{x}_t + \Delta t \cdot \frac{1}{N_k} \sum_{i\in[N_k]} \mathbf{v}_\theta(\mathbf{x}_t, t, \mathbf{c}_{i,k})$. This formulation improves cross-view geometric consistency and object completeness, but introduces two limitations in scene settings. First, the generation stage is dominated by repeated evaluations of the flow model. Therefore, applying MultiDiffusion fusion makes the inference cost scale with both the number of objects K and the number of selected views N_k, approximately as $T \sum_{k=1}^{K} N_k \,\mathrm{Cost}(\mathbf{v}_\theta)$, where T denotes the total number of flow inference steps across all generation stages. Second, uniform fusion assumes that all view conditions are compatible. When different views induce different canonical object states, averaging their velocities can merge inconsistent modes and produce duplicated or layered geometry.

3 Proposed Approach

As illustrated in Fig. 2 and Fig. 3, Scene-SAM3D follows a selection–fusion–alignment pipeline for multi-view scene asset generation. Given calibrated observations, it first selects an anchor-helper view set (§3.1), where helper views complement anchor-unobserved regions through world-space occupancy coverage. It then performs step-efficient conflict-aware latent fusion (§3.2) to integrate early multi-view cues while suppressing inconsistent updates. Finally, rigid-object Gaussian layout

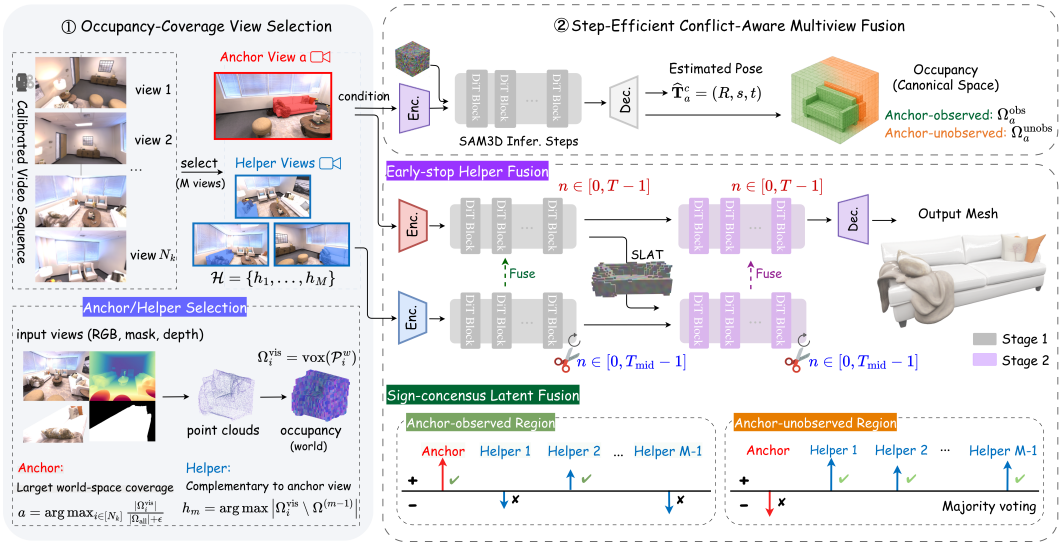


Figure 2: **Overview of Scene-SAM3D view selection and multi-view fusion.** Given calibrated multi-view observations, Scene-SAM3D first selects a compact anchor-helper view set using world-space occupancy coverage. Based on the selected views, its performs step-efficient conflict-aware multiview fusion: the anchor view provides the main generation trajectory, while helper views are early-stopped and fused to inject complementary multi-view cues. Sign consensus further filters inconsistent latent updates in anchor-observed regions and aggregates reliable evidence in anchor-unobserved regions.

optimization (§3.3) improves scene coherence without altering object geometry or appearance. We describe the first two components for a fixed object k and omit the index when clear.

3.1 Occupancy-Coverage View Selection

Using all N_k observations is both redundant and expensive. We therefore select a compact view set consisting of one anchor view a and M helper views $\mathcal{H} = \{h_1, \ldots, h_M\}$, denoted as $\mathcal{Q} = \{a\} \cup \mathcal{H}$. The anchor provides the primary appearance and canonical reference, while helper views are selected only to complement object regions that are weakly observed or invisible from the anchor.

Anchor View Selection. To measure view coverage in 3D rather than in the image plane, we project each masked observation into the world coordinate system. Given an estimated depth map $\mathbf{D}_i$, the world-space point set of view i is

$$\mathcal{P}_i^w = \left\{ \mathbf{C}_i \begin{bmatrix} \mathbf{D}_i(\mathbf{u})\mathbf{K}^{-1}\bar{\mathbf{u}} \\ 1 \end{bmatrix} \,\middle|\, \mathbf{M}_i(\mathbf{u}) = 1 \right\}, \tag{1}$$

where $\bar{\mathbf{u}}$ is the homogeneous image coordinate and $\mathbf{C}_i$ is the camera-to-world pose. We voxelize $\mathcal{P}_i^w$ in the shared world frame and obtain the visible occupancy of view i as $\Omega_i^{\text{vis}} = \text{vox}(\mathcal{P}_i^w)$. The per-view 3D coverage score is defined as the *volume* of the merged occupancy and thus anchor selection reduces to choosing the view with the largest world-space occupied volume:

$$c_i = \frac{|\Omega_i^{\text{vis}}|}{|\Omega_{\text{all}}| + \epsilon}, \quad \Omega_{\text{all}} = \bigcup_{i=1}^{N_k} \Omega_i^{\text{vis}}, \quad a = \arg\max_i c_i, \tag{2}$$

where $|\Omega_{\text{all}}|$ is shared by all views. Notably, this criterion *differs* from selecting the largest 2D mask $|\mathbf{M}_i|$: a close-up view may contain many object pixels, but after voxelization, repeated samples from the same local surface occupy only a small number of voxels. Thus, occupancy coverage favors views that explain a broader 3D extent of the object rather than views with merely larger image footprints.

Helper View Selection. After selecting the anchor, we greedily add helper views that contribute new 3D evidence beyond the regions already covered. Starting from $\Omega^{(0)} = \Omega_a^{\text{vis}}$, the m-th helper view is selected by maximizing its complementary occupancy:

$$h_m = \arg\max_{i \notin \{a, h_1, \ldots, h_{m-1}\}} \left| \Omega_i^{\text{vis}} \setminus \Omega^{(m-1)} \right|, \quad \Omega^{(m)} = \Omega^{(m-1)} \cup \Omega_{h_m}^{\text{vis}}. \tag{3}$$

In this way, the selected view set separates view reliability from view quantity: the anchor stabilizes canonical generation, while helper views are introduced only when they provide genuinely complementary 3D support. Examples of the selected anchor and helper views are provided in Fig. 7.

3.2 STEP-EFFICIENT CONFLICT-AWARE MULTIVIEW FUSION

The selected view set $\mathcal{Q} = \{a\} \cup \mathcal{H}$ reduces view redundancy, but does not resolve how these views should be fused. While *multiview velocity fusion* (Sec. 2) can aggregate complementary cues, applying it throughout the full trajectory is both *costly* and *unstable*. It costs $\mathcal{O}((M+1)T)$ conditional evaluations per object, and different view conditions may induce misaligned canonical geometries. Consequently, direct averaging can mix incompatible velocity updates and produce duplicated or layered structures. To address these issues, we introduce two strategies as follows:

Early-stop Helper Fusion. Let $\mathbf{x}_n$ denote the intermediate latent state at n-th inference step, where $n = 0, 1, ..., T-1$. Let $t_n \in [0, 1]$ be the corresponding flow time used by the sampler. Given a view condition $\mathbf{c}_i$, the conditional velocity is predicted as $\mathbf{v}_\theta(\mathbf{x}_n, t_n, \mathbf{c}_i)$ A direct fusion strategy updates the latent using the average velocity from all selected views:

$$\bar{\mathbf{v}}_n^{\text{dir}} = \frac{1}{|\mathcal{Q}|} \sum_{i \in \mathcal{Q}} \mathbf{v}_\theta(\mathbf{x}_n, t_n, \mathbf{c}_i), \qquad \mathbf{x}_{n+1} = \mathbf{x}_n + \Delta t_n \cdot \bar{\mathbf{v}}_n^{\text{dir}}. \tag{4}$$

Empirically, we observe that multi-view cues are most beneficial during early sampling (refer to Fig. 6), when the trajectory determines coarse object structure and recovers missing geometry. In contrast, fusion applied from the middle to the end of the trajectory can amplify canonical conflicts, as the generator increasingly refines view-specific details. This motivates a ***step-efficient*** strategy: helper views assist the structure-forming phase, but the final trajectory should be stabilized by the anchor.

We therefore fuse anchor and helper views only before a midpoint T_{mid}, and then continue with anchor-only inference:

$$\bar{\mathbf{v}}_n = \begin{cases} \text{Fuse}\left(\{\mathbf{v}_\theta(\mathbf{x}_n, t_n, \mathbf{c}_i)\}_{i \in \mathcal{Q}}\right), & n \le T_{\text{mid}} - 1, \\ \mathbf{v}_\theta(\mathbf{x}_n, t_n, \mathbf{c}_a), & n > T_{\text{mid}} - 1, \end{cases} \qquad \mathbf{x}_{n+1} = \mathbf{x}_n + \Delta t_n \cdot \bar{\mathbf{v}}_n. \tag{5}$$

In practice, $\text{Fuse}(\cdot)$ is implemented as a lightweight consensus operation described below. The same early-stop fusion rule is applied independently to both Stage 1 and Stage 2 of Scene-SAM3D, with the latent state and velocity instantiated in the corresponding stage-specific latent space. Compared with running all $M+1$ selected views for all T steps, early-stop fusion reduces the per-object conditional evaluations from $\mathcal{O}((M+1)T)$ to $\mathcal{O}(T + MT_{\text{mid}})$, while preserving helper-view information during the stage where it has the strongest influence on geometry.

Sign-Consensus Latent Fusion. Early stopping limits the duration of multi-view interaction, but conflicting helper signals can still enter the fused update before T_{mid}. We therefore define $\text{Fuse}(\cdot)$ in Eq. (5) as a sign-consensus operator over the structured velocity latent. At step n, each selected view $i \in \mathcal{Q}$ predicts a conditional velocity in the structured latent grid: $\mathbf{Z}_{i,n} = \mathbf{v}_\theta(\mathbf{x}_n, t_n, \mathbf{c}_i) \in \mathbb{R}^{16\times16\times16\times d}$. For readability, we omit the step index n below and write $\mathbf{Z}_i$ for $\mathbf{Z}_{i,n}$. We first identify which latent cells are directly observed by the anchor. SAM3D Stage-1 predicts an anchor camera-relative transform $\widehat{\mathbf{T}}_a^c = (\widehat{\mathbf{R}}_a^c, \widehat{\mathbf{s}}_a^c, \widehat{\mathbf{t}}_a^c)$, which maps a canonical object point $\mathbf{p}^o$ to the anchor camera frame:

$$\mathbf{p}_a^c = \widehat{\mathbf{R}}_a^c \left(\widehat{\mathbf{s}}_a^c \odot \mathbf{p}^o\right) + \widehat{\mathbf{t}}_a^c. \tag{6}$$

Thus, each visible anchor pixel $\mathbf{u} \in \{\mathbf{u} \mid \mathbf{M}_a(\mathbf{u}) = 1\}$ can be mapped to canonical object frame by:

$$\mathbf{p}^o = (\widehat{\mathbf{s}}_a^c)^{-1} \odot \left(\widehat{\mathbf{R}}_a^c\right)^\top \left(\mathbf{p}_a^c - \widehat{\mathbf{t}}_a^c\right), \qquad \mathbf{p}_a^c = \mathbf{D}_a(\mathbf{u})\mathbf{K}^{-1}\bar{\mathbf{u}}. \tag{7}$$

Voxelizing these canonical points on the structured latent grid gives the anchor-observed region $\Omega_a^{\text{obs}} \subseteq \Omega$. The remaining region is $\Omega_a^{\text{unobs}} = \Omega \setminus \Omega_a^{\text{obs}}$. All sign operations below are applied element-wise over spatial cells and latent channels.

- **Case A Anchor-observed region:** For $u \in \Omega_a^{\text{obs}}$, the anchor is treated as the reference because it provides the most reliable local evidence. Only proposals whose signs *agree* with the *anchor* are allowed to contribute:

$$\mathcal{I}_{\text{obs}}(u) = \{i \in \mathcal{Q} \mid \text{sgn}(\mathbf{Z}_i(u)) = \text{sgn}(\mathbf{Z}_a(u))\}. \tag{8}$$

The fused structured velocity is then

$$\widetilde{\mathbf{Z}}(u) = \frac{1}{|\mathcal{I}_{\text{obs}}(u)|} \sum_{i \in \mathcal{I}_{\text{obs}}(u)} \mathbf{Z}_i(u), \qquad u \in \Omega_a^{\text{obs}}. \tag{9}$$

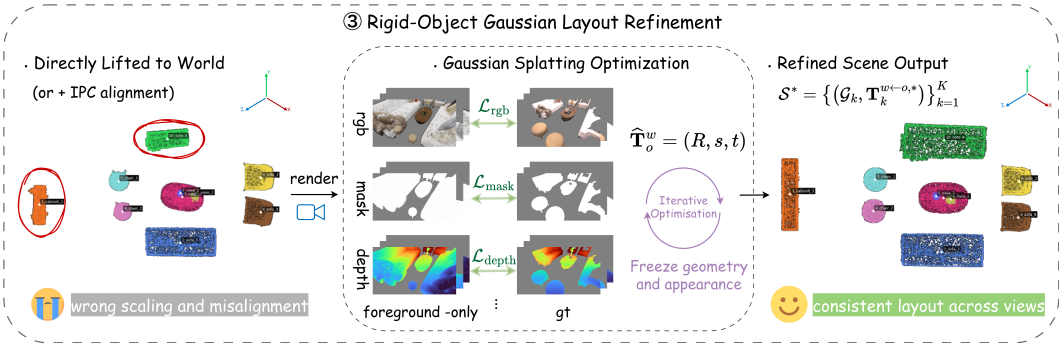


Figure 3: **Rigid-object Gaussian layout refinement.** Scene-SAM3D refines the scene by optimizing object-level similarity transforms in a foreground-only Gaussian representation.

- **Case B Anchor-unobserved region:** For $u \in \Omega_a^{\text{unobs}}$, the anchor no longer has privileged evidence. We instead let all selected views vote for the dominant direction:

$$\mathcal{I}_{\text{unobs}}(u) = \{i \in \mathcal{Q} \mid \text{sgn}(\mathbf{Z}_i(u)) = s(u)\}, s(u) = \text{sgn}\left(\sum_{i\in\mathcal{Q}} \text{sgn}(\mathbf{Z}_i(u))\right), \tag{10}$$

and average only the proposals that agree with this majority sign:

$$\widetilde{\mathbf{Z}}(u) = \frac{1}{|\mathcal{I}_{\text{unobs}}(u)|} \sum_{i\in\mathcal{I}_{\text{unobs}}(u)} \mathbf{Z}_i(u). \tag{11}$$

Finally, the sign-consensus operator is defined as $\text{Fuse}\left(\{\mathbf{v}_\theta(\mathbf{x}_t, t, \mathbf{c}_i)\}_{i\in\mathcal{Q}}\right) = \widetilde{\mathbf{Z}}_t$. Thus, anchor-visible regions are protected by the most reliable canonical reference, while anchor-unobserved regions are completed through multi-view consensus. This suppresses local canonical conflicts without discarding complementary geometry from helper views.

3.3 Rigid-Object Gaussian Layout Refinement

As shown in Fig. 3, generated objects remain in object-centric coordinates, while the pose and scale lifted from a single view may be inaccurate in the metric world frame. A simple solution is: (1) direct mask-depth backprojection (BP), or (2) refining the BP with Iterative Closest Point (ICP) (Besl & McKay, 1992) against masked depth points. However, BP is sensitive to mask and depth noise, and ICP is often under-constrained for truncated or heavily occluded objects. We therefore introduce a lightweight rigid-object Gaussian optimization that exploits calibrated multiview evidence. For each eligible generated asset $\mathcal{G}_k$, we construct a downsampled object-centric Gaussian proxy $\mathcal{P}_k = \{\boldsymbol{\mu}_{k,j}^o\}_{j=1}^{J_k}$, where j indexes Gaussian primitives within instance k. The Gaussian geometry, opacity, and appearance of all proxies remain fixed, and the K eligible instances are optimized jointly by updating only their object-level similarity transforms $\mathbf{T}_k^{w\leftarrow o} = (\mathbf{R}_k, s_k, \mathbf{t}_k)$. For each instance, the same transform is shared by all its Gaussian primitives; thus, for $j = 1, \ldots, J_k$, the transformed world-space mean is

$$\boldsymbol{\mu}_{k,j}^w = s_k \mathbf{R}_k \left(\boldsymbol{\mu}_{k,j}^o - \mathbf{c}_k^o\right) + \mathbf{t}_k, \tag{12}$$

where $\mathbf{c}_k^o$ is the canonical object center. The Gaussian orientations and scales are transformed accordingly by the same $\mathbf{R}_k$ and s_k, while all canonical Gaussian parameters remain unchanged. For each valid instance-view pair (k, i), given the camera-to-world pose $\mathbf{C}_{k,i}$ and intrinsics $\mathbf{K}$, we render the transformed proxy into view i with a differentiable Gaussian rasterizer, producing an opacity map $\widehat{\mathbf{A}}_{k,i}(\mathbf{T}_k^{w\leftarrow o})$. We also render all optimized instances jointly to obtain scene-level color, depth, and opacity for additional multiview constraints. The object transforms are optimized with

$$\mathcal{L}_{\text{layout}} = \lambda_m \mathcal{L}_{\text{mask}} + \lambda_{\text{rgb}} \mathcal{L}_{\text{fg}} + \lambda_d \mathcal{L}_{\text{depth}} + \lambda_b \mathcal{L}_{\text{bnd}} + \lambda_c \mathcal{L}_{\text{comp}} + \lambda_p \mathcal{L}_{\text{pen}} + \mathcal{L}_{\text{reg}}. \tag{13}$$

Here, $\mathcal{L}_{\text{mask}}$, $\mathcal{L}_{\text{fg}}$, $\mathcal{L}_{\text{depth}}$, and $\mathcal{L}_{\text{bnd}}$ supervise instance masks, foreground appearance, depth, and silhouette boundaries, respectively. Meanwhile, $\mathcal{L}_{\text{comp}}$ enforces scene-level rendering consistency, $\mathcal{L}_{\text{pen}}$ penalizes inter-object penetration, and $\mathcal{L}_{\text{reg}}$ regularizes the solution toward the initial layout. Since the optimization only updates object-level RST parameters rather than Gaussian geometry or appearance, it is lightweight and typically converges within a few hundred iterations. The final output is the refined scene asset

$$\mathcal{S}^* = \left\{\left(\mathcal{G}_k, \mathbf{T}_k^{w\leftarrow o,*}\right)\right\}_{k=1}^K. \tag{14}$$

Table 1: **Instance-level quantitative comparison with single-view methods and reproduced multi-view baselines.** "random" denotes random selection, while "coverage" denotes our occupancy-coverage strategy. MD + SAM3D applies MultiDiffusion fusion to SAM3D. Although InstantMesh takes one image, it synthesizes multiple intermediate views and is considered a multi-view method.

		Replica							ScanNet++			
		Geometry				*Reconstruction*			*Geometry*			
Method	**View**	CD↓	Comp↓	NC↑	F@0.05↑	PSNR↑	SSIM↑	LPIPS↓	CD↓	Comp↓	NC↑	F@0.05↑
Single-view methods												
Trellis$_{(CVPR'25)}$	1(random)	114.48	93.89	58.44	0.33	9.80	0.60	0.54	143.48	128.81	54.66	0.22
	1(coverage)	90.41	81.83	62.99	0.42	9.97	0.63	0.57	115.41	112.13	62.00	0.32
Amodal3R$_{(ICCV'25)}$	1(random)	73.38	70.91	66.13	0.50	10.92	0.65	0.51	102.77	85.34	61.77	0.34
	1(coverage)	58.87	52.46	68.70	0.58	11.11	0.64	0.48	75.78	59.19	68.58	0.48
SAM3D$_{(CVPR'26)}$	1(random)	61.96	55.63	68.30	0.59	12.55	0.71	0.46	79.25	60.49	67.83	0.46
	1(coverage)	41.60	33.20	**75.26**	0.71	13.08	0.70	0.41	57.71	39.71	73.35	0.60
Multi-view methods												
InstantMesh$_{(arXiv'24)}$	1(random)	93.29	105.31	62.22	0.40	11.42	0.68	0.49	111.76	112.79	58.49	0.30
	1(coverage)	70.92	77.31	67.27	0.50	11.78	0.69	0.53	83.44	75.77	64.77	0.41
ReconViaGen$_{(ICLR'26)}$	4(random)	81.24	89.79	64.25	0.45	9.39	0.60	0.53	109.26	114.83	57.04	0.33
	4(coverage)	70.76	75.15	65.62	0.53	9.42	0.60	0.51	92.80	91.56	59.59	0.41
MV-SAM3D$_{(arXiv'26)}$	4(random)	44.70	35.59	74.39	0.69	13.13	0.71	0.42	59.96	38.86	72.82	0.57
	4(coverage)	41.16	32.24	75.20	0.72	13.20	0.71	0.40	57.96	**36.60**	73.59	0.60
MD+SAM3D$_{(CVPR'26)}$	4(random)	45.83	37.66	74.24	0.69	13.09	0.71	0.42	60.01	38.89	73.00	0.57
	4(coverage)	40.98	32.06	74.86	0.72	13.19	0.71	0.40	58.51	37.57	74.07	0.60
Ours	4(coverage)	**39.37**	**30.65**	75.17	**0.73**	**13.20**	**0.71**	**0.40**	**57.36**	36.94	**74.50**	**0.61**



Figure 4: **Instance-level qualitative results on the Replica and ScanNet++ datasets.**

4 Experiments

Setup. We evaluate our method on both the synthetic Replica dataset (Straub et al., 2019) and the real-world ScanNet++ dataset (Yeshwanth et al., 2023). Implementation details are in the Appendix B.

4.1 Main Results

Quantitative Results. *At the instance level*, Table 1 shows that occupancy-coverage view selection improves reconstruction quality in both single-view and multi-view settings. Replacing random views with coverage-selected views reduces SAM3D CD by 32.9% on Replica and 27.2% on ScanNet++. Interestingly, existing multi-view methods such as InstantMesh and ReconViaGen do not consistently surpass strong completion-oriented single-view models like Amodal3R and SAM3D, highlighting the importance of amodal priors under severe occlusion. By contrast, our fusion strategy helps SAM3D better exploit complementary views, achieving the best CD and F@0.05 on both datasets. Under the same four-view coverage setting, our method further outperforms MD+SAM3D, reducing CD by

Table 2: **Scene-level quantitative comparison on Replica.** BP denotes direct backprojection, ICP denotes refinement from the BP using masked depth points, and GS denotes our proposed refinement.

			Geometry			*Reconstruction*			
Method	**View**	**Align**	CD↓	NC↑	F@0.05↑	IoU↑	PSNR↑	SSIM↑	LPIPS↓
SAM3D	1(random)	BP	10.94	0.58	0.40	0.27	10.03	0.18	0.46
SAM3D	1(random)	ICP	9.49	0.62	0.52	0.29	10.03	0.19	0.45
MD+SAM3D	4(random)	BP	11.01	0.58	0.39	0.27	10.19	0.19	0.46
MD+SAM3D	4(random)	ICP	9.03	0.62	0.52	0.32	10.11	0.19	0.45
MD+SAM3D	4(coverage)	BP	10.72	0.59	0.42	0.46	10.55	0.24	0.46
MD+SAM3D	4(coverage)	ICP	8.83	0.63	0.54	0.49	10.60	0.24	0.45
Ours	4(coverage)	BP	10.24	0.59	0.41	0.48	10.67	0.25	0.46
Ours	4(coverage)	ICP	8.64	0.63	0.53	0.49	10.67	0.24	0.45
Ours	4(coverage)	GS	**4.86**	**0.73**	**0.73**	**0.69**	**12.81**	**0.38**	**0.34**



Figure 5: **Scene-level qualitative comparison with other methods on the Replica dataset.**

3.9% and Comp error by 4.4% on Replica, and reducing CD by 2.0% while improving F@0.05 by 1.7% on ScanNet++. These results show that Scene-SAM3D benefits from both informative view selection and conflict-aware multi-view fusion.

At the scene level, Table 2 and Table 6 compare different layout alignment strategies. Under the same four coverage-selected views and fusion strategy, replacing ICP with GS consistently improves both geometry and reconstruction quality. Compared with ICP, GS reduces CD by 43.8%/30.9%, improves IoU by 40.8%/42.5%, and lowers LPIPS by 24.4%/15.9% on Replica/ScanNet++, respectively. These results show that directly backprojecting SAM3D-predicted RST parameters is insufficient for accurate indoor scene layout recovery. By contrast, our rendering-based Gaussian layout refinement fully exploits scene-level multi-view cues, including RGB, depth, masks, and other view-level consistency constraints, leading to more coherent scene geometry and higher appearance fidelity.

Qualitative Results. All results use views selected by our strategy. At the instance level, Fig. 4 shows that single-view methods are vulnerable to occlusion and truncation: Trellis and Amodal3R often produce chairs and tables with missing or incomplete structures. Multi-view methods use additional observations but may still suffer from cross-view inconsistencies; ReconViaGen yields distorted reconstructions of the chair and fire extinguisher, while MultiDiffusion introduces redundant structures in the bin and table cases. By contrast, Scene-SAM3D produces more complete geometry while suppressing duplicated, misaligned, and hallucinated structures, and can even recover parts missing or fragmented in the ground-truth mesh.

Table 3: Ablation of anchor masks.

Mask	CD↓	Comp↓	NC↑	F@0.05↑
$\mathcal{A}$	41.37	32.39	74.85	0.72
$\mathcal{B}$	40.38	32.13	75.24	0.72
Ours	**39.37**	**30.65**	**75.17**	**0.73**

Table 4: Efficiency study.

Method	# View	TFLOPs	Time (s)
MD +SAM3D	10	3,120	43.18
Ours	**10**	**2,520**	**34.83**

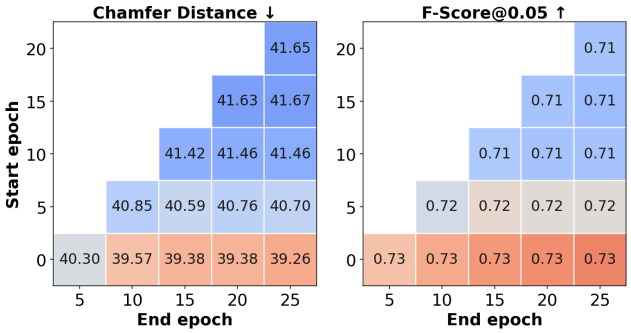


Figure 6: Ablation study on the multiview fusion steps.

Fig. 5 shows qualitative scene-level comparisons. After backprojecting generated objects into the world frame, SAM3D and MD+SAM3D still exhibit noticeable layout errors, even after ICP refinement. These include low-quality instances and inaccurate object scales and orientations. For example, the office sofa in the second column is undersized, while the chairs in the first column and stools in the third show flipped or unstable orientations. Moreover, inaccurate inter-object relationships lead to floating objects or collisions, such as the lamp–cabinet arrangement in the fourth column. In contrast, our GS layout refinement produces more plausible RST and spatial relationships.

4.2 Ablation Study

We conduct ablation studies on Replica (Straub et al., 2019) to analyze the key components of Scene-SAM3D. We present ablations on anchor guidance, fusion steps, and efficiency in the main paper, and provide additional studies on view selection and fusion strategies in Appendix D.

Anchor Guidance in Velocity Fusion. We further evaluate the privilege of the anchor view during velocity fusion. Besides our default design, where the anchor is used as the reference only in its visible region $u \in \Omega_a^{\text{obs}}$, we compare two variants: $\mathcal{A}$ removes anchor privilege and lets all selected views vote at every position, while $\mathcal{B}$ assigns anchor privilege to all positions, including $u \in \Omega_a^{\text{unobs}}$. As shown in Tab. 3, using the anchor as reference only in its observed region achieves the best performance. This strategy leverages the reliable local evidence from the anchor while preventing hallucinated anchor predictions in unobserved regions from being propagated to the final geometry.

Impact of Fusion Steps. We evaluate how the fusion start and end steps affect generation quality. As shown in Fig. 6, warmer colors indicate better quality. Starting fusion from step 0 achieves the best results, suggesting that multi-view cues are most useful during early structure formation and geometry completion. In contrast, continuing fusion from the middle steps to the end brings only marginal gains and may introduce canonical-space conflicts from view-specific refinements. We therefore adopt early-stop fusion to reduce computational cost while maintaining geometric quality.

Efficiency Study in Multiview Fusion. We further quantitatively evaluate the efficiency gains brought by our step-efficient fusion strategy. Since the remaining components of the pipeline are kept unchanged, we focus on comparing the flow-model inference time and FLOPs against the MultiDiffusion fusion baseline. The main cost comes from repeated forward evaluations of the flow velocity model: MultiDiffusion fusion runs the full flow process for every selected view at every inference step, whereas our early-stop fusion only uses helper-view conditions in the early steps and switches to anchor-only inference afterwards. As shown in Table 4, under ten-view inference, our method reduces both flow-model FLOPs and flow-model sampling wall-time latency by nearly 20%.

5 Conclusion

We introduced Scene-SAM3D, a training-free framework that extends object-centric single-view 3D foundation models to calibrated multi-view scene asset generation. Compared with SAM3D, it reduces scene-level CD by 43.8% on Replica and 30.9% on ScanNet++. Existing scene datasets, however, remain limited in the availability and quality of ground-truth meshes. These meshes may omit unobserved surfaces and exhibit appearance discrepancies with the captured images. Such imperfections can cause instance-level metrics to underestimate the quality of our reconstructions, even when they are more complete and visually faithful. These scene data limitations also hinder the construction of realistic digital environments. Scene-SAM3D has the potential to help address these limitations by generating large-scale collections of geometrically complete and visually faithful scene assets, with applications in embodied intelligence, virtual reality, and augmented reality.

AI USE STATEMENT

In this work, we used generative AI tools for assisting with the implementation of our methods and the analysis of experimental results. Additionally, we used generative AI tools for editing the manuscript to improve grammar, clarity, and readability. We have reviewed all AI-assisted work. We manually checked the AI-assisted implementations, analyses, and manuscript edits to ensure their accuracy and consistency with our methods and experimental results. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Our experiments use the Replica and ScanNet++ benchmarks and publicly released pretrained models under their respective licenses and terms of use. We do not redistribute third-party datasets or pretrained weights. Scene-SAM3D generates indoor scene assets from calibrated multi-view observations. Applications to private environments should respect consent and protect sensitive visual and spatial information. Generated assets may contain inaccurate geometry, appearance, or spatial relationships, particularly in occluded regions. Their use in embodied simulation or downstream robotics should therefore include task-specific validation of geometric accuracy and physical plausibility.

REPRODUCIBILITY STATEMENT

The method is specified in Sections 3.1, 3.2, and 3.3, covering view selection, multi-view fusion, and scene layout refinement, respectively. Scene-SAM3D uses pretrained SAM3D without fine-tuning. Appendix B describes the benchmark datasets, implementation settings, hardware, and evaluation metrics. Appendices C and D provide additional qualitative results and ablation studies. Code will be released after the review phase.

REFERENCES

David Arthur and Sergei Vassilvitskii. k-means++: the advantages of careful seeding. In Nikhil Bansal, Kirk Pruhs, and Clifford Stein (eds.), *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2007, New Orleans, Louisiana, USA, January 7-9, 2007*, pp. 1027–1035. SIAM, 2007.

Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML*, pp. 1737–1752, 2023.

Paul J. Besl and Neil D. McKay. A method for registration of 3-d shapes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 14(2):239–256, 1992.

Rowan Border and Jonathan D. Gammell. The surface edge explorer (see): A measurement-direct approach to next best view planning. *Int. J. Robotics Res.*, 43(10):1506–1532, 2024.

Rowan Border, Jonathan D. Gammell, and Paul Newman. Surface edge explorer (see): Planning next best views directly from 3d observations. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1–8. IEEE, 2018.

Jiahao Chang, Chongjie Ye, Yushuang Wu, Yuantao Chen, Yidan Zhang, Zhongjin Luo, Chenghong Li, Yihao Zhi, and Xiaoguang Han. Reconviagen: Towards accurate multi-view 3d object reconstruction via generation. In *International Conference on Learning Representations (ICLR)*, 2026.

Xiao Chen, Quanyi Li, Tai Wang, Tianfan Xue, and Jiangmiao Pang. Gennbv: Generalizable next-best-view policy for active 3d reconstruction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16436–16445. IEEE, 2024.

Xingyu Chen, Fu-Jen Chu, Pierre Gleize, Kevin J. Liang, Alexander Sax, Hao Tang, Weiyao Wang, Michelle Guo, Thibaut Hardin, Xiang Li, Aohan Lin, Jia-Wei Liu, Ziqi Ma, Anushka Sagar, Bowen Song, Xiaodong Wang, Jianing Yang, Bowen Zhang, Piotr Dollár, Georgia Gkioxari, Matt Feiszli, and Jitendra Malik. SAM 3d: 3dfy anything in images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7220–7232, 2026.

Jasmine Collins, Shubham Goel, Kenan Deng, Achleshwar Luthra, Leon Xu, Erhan Gundogdu, Xi Zhang, Tomas F. Yago Vicente, Thomas Dideriksen, Himanshu Arora, Matthieu Guillaumin, and Jitendra Malik. ABO: dataset and benchmarks for real-world 3d object understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21094–21104. IEEE, 2022.

Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, Eli VanderBilt, Aniruddha Kembhavi, Carl Vondrick, Georgia Gkioxari, Kiana Ehsani, Ludwig Schmidt, and Ali Farhadi. Objaverse-xl: A universe of 10m+ 3d objects. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Noah Frahm, Dongxu Zhao, Andrea Dunn Beltran, Ron Alterovitz, Jan-Michael Frahm, Junier Oliva, and Roni Sengupta. VIN-NBV: a view introspection network for next-best-view selection for resource-efficient 3d reconstruction. *CoRR*, abs/2505.06219, 2025.

Jiayuan Gu, Fanbo Xiang, Xuanlin Li, Zhan Ling, Xiqiang Liu, Tongzhou Mu, Yihe Tang, Stone Tao, Xinyue Wei, Yunchao Yao, Xiaodi Yuan, Pengwei Xie, Zhiao Huang, Rui Chen, and Hao Su. Maniskill2: A unified benchmark for generalizable manipulation skills. In *The Eleventh International Conference on Learning Representations (ICLR)*. OpenReview.net, 2023.

Junlin Han, Jianyuan Wang, Andrea Vedaldi, Philip Torr, and Filippos Kokkinos. Flex3d: Feed-forward 3d generation with flexible reconstruction model and input view curation. In *Forty-second International Conference on Machine Learning (ICML)*, Proceedings of Machine Learning Research. PMLR / OpenReview.net, 2025.

Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. LRM: large reconstruction model for single image to 3d. In *The Twelfth International Conference on Learning Representations (ICLR)*. OpenReview.net, 2024.

Zehuan Huang, Yuan-Chen Guo, Xingqiao An, Yunhan Yang, Yangguang Li, Zi-Xin Zou, Ding Liang, Xihui Liu, Yan-Pei Cao, and Lu Sheng. MIDI: multi-instance diffusion for single image to 3d scene generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23646–23657. Computer Vision Foundation / IEEE, 2025a.

Zhening Huang, Xiaoyang Wu, Fangcheng Zhong, Hengshuang Zhao, Matthias Nießner, and Joan Lasenby. Litereality: Graphics-ready 3d scene reconstruction from RGB-D scans. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 162794–162827, 2025b.

Zhisheng Huang, Jiahao Chen, Cheng Lin, Chenyu Hu, Hanzhuo Huang, Zhengming Yu, Mengfei Li, Yuheng Liu, Zekai Gu, Zibo Zhao, et al. Unirecgen: Unifying multi-view 3d reconstruction and generation. *arXiv preprint arXiv:2604.01479*, 2026.

Wen Jiang, Boshu Lei, and Kostas Daniilidis. Fisherrf: Active view selection and mapping with radiance fields using fisher information. In *Computer Vision - ECCV 2024*, pp. 422–440. Springer, 2024.

Brett R. Jones, Rajinder Sodhi, Michael Murdock, Ravish Mehra, Hrvoje Benko, Andrew D. Wilson, Eyal Ofek, Blair MacIntyre, Nikunj Raghuvanshi, and Lior Shapira. Roomalive: Magical experiences enabled by scalable, adaptive projector-camera units. In *The 27th Annual ACM Symposium on User Interface Software and Technology (UIST)*, pp. 637–644. ACM, 2014.

Hiroharu Kato, Yoshitaka Ushiku, and Tatsuya Harada. Neural 3d mesh renderer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3907–3916, 2018.

Mukul Khanna, Yongsen Mao, Hanxiao Jiang, Sanjay Haresh, Brennan Shacklett, Dhruv Batra, Alexander Clegg, Eric Undersander, Angel X. Chang, and Manolis Savva. Habitat synthetic scenes dataset (HSSD-200): An analysis of 3d scene scale and realism tradeoffs for objectgoal navigation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16384–16393. IEEE, 2024.

Xin Kong, Shikun Liu, Marwan Taher, and Andrew J. Davison. vmap: Vectorised object mapping for neural field SLAM. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 952–961. IEEE, 2023.

Baicheng Li, Dong Wu, Jun Li, Shunkai Zhou, Zecui Zeng, Lusong Li, and Hongbin Zha. MV-SAM3D: adaptive multi-view fusion for layout-aware 3d generation. *CoRR*, abs/2603.11633, 2026.

Yangguang Li, Zi-Xin Zou, Zexiang Liu, Dehu Wang, Yuan Liang, Zhipeng Yu, Xingchao Liu, Yuan-Chen Guo, Ding Liang, Wanli Ouyang, et al. TripoSG: high-fidelity 3d shape synthesis using large-scale rectified flow models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Yang Zhao, Sida Peng, Hengkai Guo, Xiaowei Zhou, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth Anything 3: Recovering the visual space from any views. In *International Conference on Learning Representations (ICLR)*, 2026.

Lu Ling, Yunhao Ge, Yichen Sheng, and Aniket Bera. I-scene: 3d instance models are implicit generalizable spatial learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26974–26983, 2026.

Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T., Zexiang Xu, and Hao Su. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023a.

Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9264–9275. IEEE, 2023b.

Shichen Liu, Tianye Li, Weikai Chen, and Hao Li. Soft rasterizer: A differentiable renderer for image-based 3d reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7707–7716, 2019. doi: 10.1109/ICCV.2019.00780.

Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. In *The Twelfth International Conference on Learning Representations (ICLR)*. OpenReview.net, 2024.

Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, and Wenping Wang. Wonder3d: Single image to 3d using cross-domain diffusion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9970–9980. IEEE, 2024.

Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. SQA3D: situated question answering in 3d scenes. In *The Eleventh International Conference on Learning Representations (ICLR)*. OpenReview.net, 2023.

Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul McVay, Oleksandr Maksymets, Sergio Arnaud, Karmesh Yadav, Qiyang Li, Ben Newman, Mohit Sharma, Vincent-Pierre Berges, Shiqi Zhang, Pulkit Agrawal, Yonatan Bisk, Dhruv Batra, Mrinal Kalakrishnan, Franziska Meier, Chris Paxton, Alexander Sax, and Aravind Rajeswaran. Openeqa: Embodied question answering in the era of foundation models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16488–16498. IEEE, 2024.

Yanxu Meng, Haoning Wu, Ya Zhang, and Weidi Xie. Scenegen: Single-image 3d scene generation in one feedforward pass. In *International Conference on 3D Vision (3DV)*, pp. 543–553, 2026. doi: 10.1109/3DV69130.2026.00058.

Richard A. Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J. Davison, Pushmeet Kohli, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *10th IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 127–136. IEEE, 2011.

Junfeng Ni, Yu Liu, Ruijie Lu, Zirui Zhou, Song-Chun Zhu, Yixin Chen, and Siyuan Huang. Decompositional neural scene reconstruction with generative diffusion prior. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6022–6033. Computer Vision Foundation / IEEE, 2025.

Sergio Orts-Escolano, Christoph Rhemann, Sean Ryan Fanello, Wayne Chang, Adarsh Kowdle, Yury Degtyarev, David Kim, Philip Davidson, Sameh Khamis, Mingsong Dou, Vladimir Tankovich, Charles T. Loop, Qin Cai, Philip A. Chou, Sarah Mennicken, Julien P. C. Valentin, Vivek Pradeep, Shenlong Wang, Sing Bing Kang, Pushmeet Kohli, Yuliya Lutchyn, Cem Keskin, and Shahram Izadi. Holoportation: Virtual 3d teleportation in real-time. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology (UIST)*, pp. 741–754. ACM, 2016.

Daryl Peralta, Joel Casimiro, Aldrin Michael Nilles, Justine Aletta Aguilar, Rowel Atienza, and Rhandley Cajote. Next-best view policy for 3d reconstruction. In *Computer Vision - ECCV 2020 Workshops*, pp. 558–573. Springer, 2020.

Richard Pito. A solution to the next best view problem for automated surface acquisition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 21(10):1016–1030, 1999.

Xavier Puig, Eric Undersander, Andrew Szot, Mikael Dallaire Cote, Tsung-Yen Yang, Ruslan Partsey, Ruta Desai, Alexander Clegg, Michal Hlavac, So Yeon Min, Vladimir Vondrus, Théophile Gervet, Vincent-Pierre Berges, John M. Turner, Oleksandr Maksymets, Zsolt Kira, Mrinal Kalakrishnan, Jitendra Malik, Devendra Singh Chaplot, Unnat Jain, Dhruv Batra, Akshara Rai, and Roozbeh Mottaghi. Habitat 3.0: A co-habitat for humans, avatars, and robots. In *The Twelfth International Conference on Learning Representations (ICLR)*. OpenReview.net, 2024.

Yukai Shi, Weiyu Li, Zihao Wang, Hongyang Li, Xingyu Chen, Ping Tan, and Lei Zhang. Scenemaker: Open-set 3d scene generation with decoupled de-occlusion and pose estimation model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27146–27156, 2026.

Yawar Siddiqui, Duncan P. Frost, Samir Aroudj, Armen Avetisyan, Henry Howard-Jenkins, Daniel DeTone, Pierre Moulon, Qirui Wu, Zhengqin Li, Julian Straub, Richard A. Newcombe, and Jakob J. Engel. Shaper: Robust conditional 3d shape generation from casual captures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27157–27168, 2026.

Misha Sra, Sergio Garrido-Jurado, and Pattie Maes. Oasis: Procedurally generated social virtual spaces from 3d scanned real spaces. *IEEE Trans. Vis. Comput. Graph.*, 24(12):3174–3187, 2018.

Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Yuheng Ren, Shobhit Verma, Anton Clarkson, Mingfei Yan, Brian Budge, Yajie Yan, Xiaqing Pan, June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, Jesus Briales, Tyler Gillingham, Elias Mueggler, Luis Pesqueira, Manolis Savva, Dhruv Batra, Hauke M. Strasdat, Renzo De Nardi, Michael Goesele, Steven Lovegrove, and Richard A. Newcombe. The replica dataset: A digital replica of indoor spaces. *CoRR*, abs/1906.05797, 2019.

Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitry Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani. SV3D: novel multi-view synthesis and 3d generation from a single image using latent video diffusion. In *Computer Vision - ECCV 2024*, pp. 439–457. Springer, 2024.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotný. VGGT: visual geometry grounded transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 5294–5306. Computer Vision Foundation / IEEE, 2025.

Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2mesh: Generating 3d mesh models from single RGB images. In *Computer Vision - ECCV 2018*, pp. 55–71. Springer, 2018.

Yufei Wang, Zhou Xian, Feng Chen, Tsun-Hsuan Wang, Yian Wang, Katerina Fragkiadaki, Zackory Erickson, David Held, and Chuang Gan. Robogen: Towards unleashing infinite data for automated robot learning via generative simulation. In *Forty-first International Conference on Machine Learning (ICML)*, Proceedings of Machine Learning Research, pp. 51936–51983. PMLR / OpenReview.net, 2024a.

Zhengyi Wang, Yikai Wang, Yifei Chen, Chendong Xiang, Shuo Chen, Dajiang Yu, Chongxuan Li, Hang Su, and Jun Zhu. CRM: single image to 3d textured mesh with convolutional reconstruction model. In *Computer Vision - ECCV 2024*, pp. 57–74. Springer, 2024b.

Hao Wen, Zehuan Huang, Yaohui Wang, Xinyuan Chen, and Lu Sheng. Ouroboros3d: Image-to-3d generation via 3d-aware recursive diffusion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21631–21641. Computer Vision Foundation / IEEE, 2025.

Joey Wilson, Marcelino Almeida, Sachit Mahajan, Martin Labrie, Maani Ghaffari, Omid Ghasemalizadeh, Min Sun, Cheng-Hao Kuo, and Arnab Sen. Pop-gs: Next best view in 3d-gaussian splatting with p-optimality. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3646–3655. Computer Vision Foundation / IEEE, 2025.

Kailu Wu, Fangfu Liu, Zhihan Cai, Runjie Yan, Hanyang Wang, Yating Hu, Yueqi Duan, and Kaisheng Ma. Unique3d: High-quality and efficient 3d mesh generation from a single image. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024a.

Shuang Wu, Youtian Lin, Feihu Zhang, Yifei Zeng, Jingxi Xu, Philip Torr, Xun Cao, and Yao Yao. Direct3d: Scalable image-to-3d generation via 3d latent diffusion transformer. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024b.

Tianhao Wu, Chuanxia Zheng, Frank Guan, Andrea Vedaldi, and Tat-Jen Cham. Amodal3r: Amodal 3d reconstruction from occluded 2d images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9181–9193, 2025.

Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21469–21480, 2025.

Jiale Xu, Weihao Cheng, Yiming Gao, Xintao Wang, Shenghua Gao, and Ying Shan. Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. *CoRR*, abs/2404.07191, 2024.

Linning Xu, Vasu Agrawal, William Laney, Tony Garcia, Aayush Bansal, Changil Kim, Samuel Rota Bulò, Lorenzo Porzi, Peter Kontschieder, Aljaž Božič, et al. VR-NeRF: high-fidelity virtualized walkable spaces. In *SIGGRAPH Asia Conference Papers*, pp. 1–12, 2023.

Qiangeng Xu, Weiyue Wang, Duygu Ceylan, Radomír Mech, and Ulrich Neumann. DISN: deep implicit surface network for high-quality single-view 3d reconstruction. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 490–500, 2019.

Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12–22. IEEE, 2023.

Zibo Zhao, Zeqiang Lai, Qingxiang Lin, Yunfei Zhao, Haolin Liu, Shuhui Yang, Yifei Feng, Mingxin Yang, Sheng Zhang, Xianghui Yang, Huiwen Shi, Sicong Liu, Junta Wu, Yihang Lian, Fan Yang, Ruining Tang, Zebin He, Xinzhou Wang, Jian Liu, Xuhui Zuo, Zhuo Chen, Biwen Lei, Haohan Weng, Jing Xu, Yiling Zhu, Xinhai Liu, Lixin Xu, Changrong Hu, Tianyu Huang, Lifu Wang, Jihong Zhang, Meng Chen, Liang Dong, Yiwen Jia, Yulin Cai, Jiaao Yu, Yixuan Tang, Hao Zhang, Zheng Ye, Peng He, Runzhou Wu, Chao Zhang, Yonghao Tan, Jie Xiao, Yangyu Tao, Jianchen Zhu, Jinbao Xue, Kai Liu, Chongqing Zhao, Xinming Wu, Zhichao Hu, Lei Qin, Jianbing Peng,

Zhan Li, Minghui Chen, Xipeng Zhang, Lin Niu, Paige Wang, Yingkai Wang, Haozhao Kuang, Zhongyi Fan, Xu Zheng, Weihao Zhuang, YingPing He, Tian Liu, Yong Yang, Di Wang, Yuhong Liu, Jie Jiang, Jingwei Huang, and Chunchao Guo. Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation. *CoRR*, abs/2501.12202, 2025.

APPENDIX

In the appendix, we provide an extended discussion of related work (Section A), detailed experimental settings (Section B), visualizations of view selection and error maps on Replica and ScanNet++ (Section C), and additional ablation studies (Section D).

A RELATED WORK

Single-image 3D Asset Generation. Early single-image 3D reconstruction methods relied on explicit geometric priors, differentiable rendering, or implicit surface modeling (Wang et al., 2018; Liu et al., 2019; Kato et al., 2018; Xu et al., 2019). Recent progress has shifted toward large generative models that either directly infer 3D representations (Hong et al., 2024; Wu et al., 2024b) or synthesize intermediate views before reconstruction (Liu et al., 2023a; Wang et al., 2024b; Long et al., 2024; Xu et al., 2024; Wu et al., 2024a). Large-scale 3D asset datasets have further enabled flow-based 3D foundation models with stronger object-level priors (Deitke et al., 2023; Collins et al., 2022; Xiang et al., 2025; Zhao et al., 2025; Li et al., 2025). Recent works extend these priors to compositional or occlusion-aware settings (Huang et al., 2025a; Meng et al., 2026; Shi et al., 2026; Ling et al., 2026; Wu et al., 2025). Most notably, SAM3D (Chen et al., 2026) marks a recent leap in natural-scene object reconstruction by jointly predicting shape, texture, and pose from arbitrary single image. We build on this strong prior, but move beyond its single-view formulation to multi-view scene asset generation, where uneven view quality, severe occlusion, and cross-view canonical conflicts must be handled explicitly.

Multi-view Fusion for 3D Generation. A common strategy in image-to-3D generation is to synthesize a small set of predefined views and reconstruct 3D geometry from them (Wang et al., 2024b; Xu et al., 2024; Wu et al., 2024a). Other methods improve cross-view consistency through feature alignment, geometry-aware fusion, video priors, or reconstruction-in-the-loop feedback (Liu et al., 2024; Long et al., 2024; Voleti et al., 2024; Wen et al., 2025). More recent approaches learn genuinely multi-view-conditioned 3D generators from large-scale 3D data (Huang et al., 2026; Siddiqui et al., 2026). Hunyuan3D 2.0 (Zhao et al., 2025) also provides a multi-view variant that conditions shape generation on up to four predefined canonical view slots—front, back, left, and right. However, it does not model arbitrary calibrated camera viewpoints, limiting its applicability to irregular multi-view observations captured in real scenes. ReconViaGen (Chang et al., 2026) further integrates VGGT (Wang et al., 2025) reconstruction priors into TRELLIS, using multi-view features as global geometry conditions for coarse structure generation and local per-view conditions for fine geometry and texture generation. Concurrent with our work, MV-SAM3D (Li et al., 2026) investigates training-free multi-view fusion in the latent space of a pre-trained SAM3D model, primarily for object-centric inputs. In contrast, Scene-SAM3D targets indoor scene asset generation, where object instances are heavily occluded, observations are highly redundant, and view selection and fusion must handle inconsistent multi-view evidence.

View Selection for 3D Generation and Reconstruction. In generative 3D, many methods still rely on a small set of fixed or pre-defined views, and explicit view selection remains largely underexplored. One-2-3-45 (Liu et al., 2023a) uses a two-stage view generation strategy, where Zero-1-to-3 (Liu et al., 2023b) expands each base view into several nearby views and a ground-truth target view is randomly sampled for supervision. This is closer to target-view sampling than explicit input-view selection. Flex3D (Han et al., 2025) is one of the few recent exceptions that explicitly performs candidate-view generation and curation, filtering views by quality and consistency before reconstruction. By contrast, active view selection or next-best-view (NBV) has been studied much more extensively in 3D *reconstruction*. Geometry-heuristic methods (Pito, 1999; Border et al., 2018; Border & Gammell, 2024) choose views according to the current geometric state, for example, by prioritizing uncovered surfaces, missing regions, or occlusion boundaries. Information-gain methods (Jiang et al., 2024; Wilson et al., 2025) instead estimate how much uncertainty a candidate view is expected to reduce, often using probabilistic map uncertainty or Fisher-information-based criteria. RL-based methods (Peralta et al., 2020; Chen et al., 2024) learn view-selection policies in continuous spaces, and prediction-based methods (Frahm et al., 2025) directly estimate the reconstruction benefit of each candidate view. In contrast, we study a more challenging problem in real scenes: selecting anchor and helper views for severely occluded instances from highly redundant observations, so that helper views can provide missing information for 3D generation.

B More Details

Setup. We evaluate our method on both the synthetic Replica dataset (Straub et al., 2019) and the real-world ScanNet++ dataset (Yeshwanth et al., 2023). Replica is a photorealistic indoor 3D scene dataset with dense mesh geometry and high-resolution textures. It contains 18 indoor scenes, including rooms and offices, and covers objects from 88 semantic categories. It supports rendering multi-view RGB images from arbitrary camera poses. ScanNet++ is a real-world indoor dataset comprising 460 high-resolution 3D scenes and over 1,000 categories of 3D assets, together with DSLR images and RGB-D frames captured using an iPhone. We follow the subset selection protocols of prior work (Ni et al., 2025; Kong et al., 2023) for Replica and ScanNet++, evaluating a total of 160 object instances across the two datasets.

Implementation Details. All experiments are conducted on a single NVIDIA H200 GPU with 140GB memory. We follow the default hyperparameter settings of SAM3D. For our proposed modules, we set the number of helper views to $M = 3$ in Occupancy-Coverage View Selection, and the midpoint to $T_{\text{mid}} = 15$ in Step-Efficient Conflict-Aware Multiview Fusion. If sign voting results in a tie during fusion, we retain all views. The balancing coefficients λ_m, λ_{rgb}, λ_d, λ_b, λ_c, and λ_p are set to 2.0, 2.5, 2.0, 0.04, 0.05, and 0.5, respectively. For view selection and instance-level generation, we use depth maps predicted by Depth Anything 3 (Lin et al., 2026). For placing the generated instances in the world coordinate and GS refinement, we use GT depth maps and camera parameters provided by the datasets. Before instance-level geometry and appearance evaluation, we align each predicted mesh to its corresponding GT mesh. Outputs from SAM3D-based and TRELLIS-based methods are already in normalized canonical coordinates. For InstantMesh and ReconViaGen, we apply bounding-box-based translation and uniform scaling, followed by rigid ICP, retaining the converged alignment with the lowest CD.

Metrics. Following prior 3D object generation and scene reconstruction methods, we evaluate geometry quality at both object and scene levels using Chamfer Distance (CD), F-Score at a threshold of 0.05, and Normal Consistency (NC). All instance-level Chamfer-based distances, including CD and $\text{CD}_{\text{GT}\to\text{Pred}}$, are reported in millimeters, while scene-level CD is reported in centimeters. To further assess scene-level spatial arrangement, we report Mask IoU results. For appearance quality, we evaluate rendered reconstructions using PSNR, SSIM, and LPIPS at both object and scene levels. Since indoor scene annotations often provide incomplete instance meshes due to limited visibility and occlusion, symmetric CD may over-penalize plausible geometry completed in unobserved regions. We therefore additionally report the one-sided Chamfer Distance from ground truth to prediction, denoted as $\text{CD}_{\text{GT}\to\text{Pred}}$ or Completeness, which measures the coverage of observed ground-truth surfaces by the generated object mesh. For experiments with random view selection, we report the mean evaluation metrics over five runs using different seeds.

Baselines. We compare Scene-SAM3D with both single-view and multi-view 3D generation baselines. For single-view reconstruction, we include Trellis (Xiang et al., 2025), Amodal3R (Wu et al., 2025), and SAM3D (Chen et al., 2026). For multi-view reconstruction, we compare with InstantMesh (Xu et al., 2024), ReconViaGen (Chang et al., 2026), MV-SAM3D (Li et al., 2026), and a reproduced MultiDiffusion-based baseline, denoted as MD + SAM3D, which applies MultiDiffusion fusion (Bar-Tal et al., 2023) on SAM3D (Chen et al., 2026). For scene-level evaluation, we further compare different alignment strategies, including direct backprojection (BP) and ICP refinement (Besl & McKay, 1992), against our proposed Gaussian layout refinement. All methods are evaluated under the same protocol, and multi-view methods use the same number of selected views for fair comparison.

C More Visualization

We visualize the selected views by comparing our Occupancy-Coverage View Selection strategy with random view selection. As shown in Fig. 7, random selection often produces incomplete or redundant observations, whereas our method selects a compact anchor-helper view set with more reliable and complementary 3D evidence. Specifically, the anchor view maximizes world-space occupancy coverage, favoring views that explain a broader 3D extent of the target object, while the helper views further complement anchor-unobserved regions or provide additional local details. For example, in the first-row table case, the randomly selected views observe the object only partially, whereas our selected views jointly cover the table more completely. Similarly, in the third-row sofa



Figure 7: Selected views visualization on Replica.

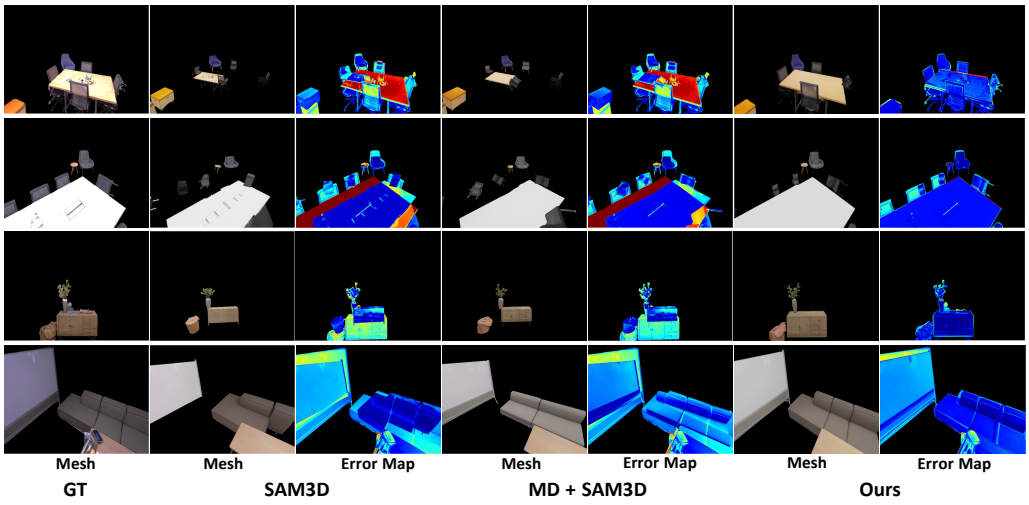


Figure 8: Qualitative visualization on Replica. In the error maps, warmer colors indicate larger alignment errors.

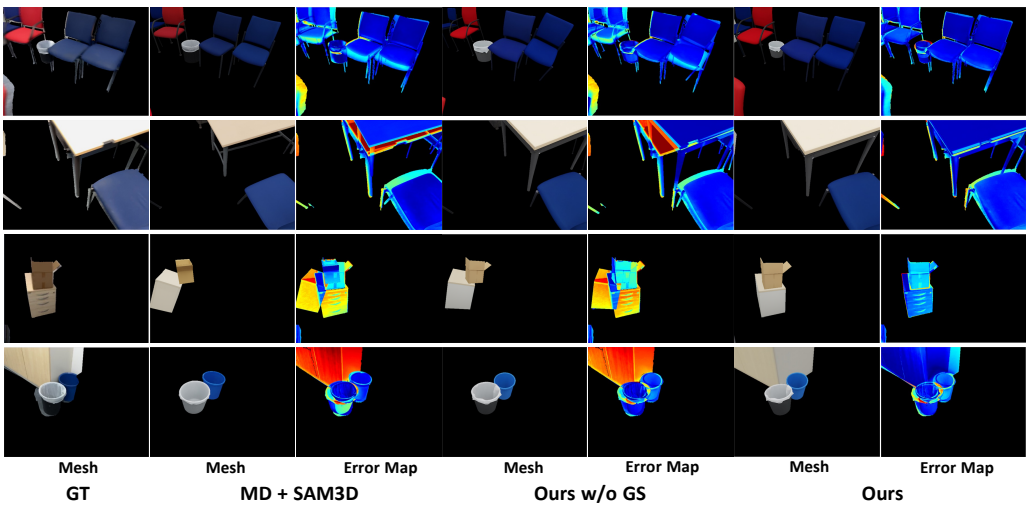


Figure 9: Qualitative visualization on ScanNet++. In the error maps, warmer colors indicate larger alignment errors.

case, random views lack sufficient observation of the sofa back, while our anchor and helper views complement each other to cover both the front and back sides.

We also provide additional qualitative results on Replica and ScanNet++ in Fig. 8 and Fig. 9, respectively, to better analyze geometry fidelity in the world coordinate system. For each scene, we place the reconstructed objects back into the world coordinate system and render them from the same viewpoints as the ground truth. We visualize both the rendered meshes and the corresponding surface error maps, where warmer colors indicate larger geometry errors. The visualizations show that our method improves both object-level reconstruction quality and scene-level layout consistency. First, our method produces more accurate object shapes and appearances, resulting in visibly lower surface errors on reconstructed objects. For example, in Fig. 8, our method better preserves the geometry of the laundry basket in the third row and the sofa structure in the fourth row. Similarly, in Fig. 9, the bins in the fourth row are reconstructed with more faithful shape and appearance, leading to smaller error regions compared with the baselines. Second, the error maps reveal that methods without our GS refinement often suffer from noticeable spatial offsets after placing objects back into the world coordinate system. These errors include inaccurate object positions, rotations, and relative object-object arrangements. For instance, in Fig. 8, the reconstructed table is placed too far from the surrounding objects, producing large errors in the scene layout. In Fig. 9, the cabinet and cardboard box in the third row exhibit an incorrect collision relationship before GS refinement. After applying GS refinement, the reconstructed scene layout becomes much better aligned with the ground truth, as indicated by the reduced warm-colored regions in the error maps.

D More Ablation Studies

View Selection and Fusion Strategy. To evaluate our multi-view selection and fusion designs, we compare Scene-SAM3D with alternative view selection and fusion strategies. For view selection, we compare our occupancy-coverage strategy with random selection and a k-means++ feature-diversity baseline (Arthur & Vassilvitskii, 2007). Specifically, the k-means++ baseline performs view selection in the multi-view feature space by favoring feature-diverse views, a criterion commonly used to reduce redundancy among candidate observations. For fusion, we compare with stochastic view conditioning and MultiDiffusion fusion, both of which can be adapted to the 3D multi-view generation setting. As shown in Tab. 5, our method achieves the best overall performance in both view selection and fusion. The only exception is SSIM, where different strategies show marginal differences. This is because SSIM mainly measures image-space structural similarity and is less sensitive to localized 3D geometric artifacts, such as duplicated legs, discontinuous surfaces, or small missing parts. Therefore, geometry-related metrics better reflect the advantages of our view selection and fusion strategy.

Table 5: Ablation on view selection and fusion strategy.

		Geometry				*Reconstruction*		
Selection	**Fusion**	CD$^{\downarrow}$	Comp$^{\downarrow}$	NC$^{\uparrow}$	F@0.05$^{\uparrow}$	PSNR$^{\uparrow}$	SSIM$^{\uparrow}$	LPIPS$^{\downarrow}$
Random	Stochastic	46.35	38.34	74.10	0.68	13.06	0.71	0.43
Random	MultiDiffusion (MD)	45.83	37.66	74.24	0.69	13.09	0.71	0.42
k-means++	Stochastic	42.45	33.56	74.76	0.71	13.07	0.71	0.41
k-means++	MultiDiffusion (MD)	47.45	38.05	73.91	0.67	12.92	0.71	0.43
Coverage	MultiDiffusion	40.98	32.06	74.86	0.72	13.19	0.71	0.40
Coverage	Ours	**39.37**	**30.65**	**75.17**	**0.73**	**13.20**	0.71	**0.40**

Table 6: Scene-level quantitative comparison on ScanNet++. As the 4-view Coverage strategy has been shown on Replica to consistently outperform both single-view inference and random view selection, we restrict the ScanNet++ comparison to the strongest baseline variants.

		Geometry			*Reconstruction*			
Method	**Align**	CD$^{\downarrow}$	NC$^{\uparrow}$	F@0.05$^{\uparrow}$	IoU$^{\uparrow}$	PSNR$^{\uparrow}$	SSIM$^{\uparrow}$	LPIPS$^{\downarrow}$
$\text{MD}_{\text{(ICML'23)}}$+$\text{SAM3D}_{\text{(CVPR'26)}}$	BP	8.05	0.53	0.58	0.36	9.36	0.21	0.46
$\text{MD}_{\text{(ICML'23)}}$+$\text{SAM3D}_{\text{(CVPR'26)}}$	ICP	7.18	0.61	0.60	0.39	9.59	0.22	0.45
Ours	ICP	6.89	0.63	0.61	0.40	9.65	0.22	0.44
Ours	GS	**4.76**	**0.74**	**0.69**	**0.57**	**11.07**	**0.31**	**0.37**