SEGA: MULTI-VIEW 3D INSTANCE SEGMENTATION WITH GEOMETRY-GUIDED ASSOCIATION FOR INDOOR ENVIRONMENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

We present **SEGA**, a geometry-guided framework for 3D instance segmentation of indoor scenes from multi-view image sequences. Existing scene reconstruction and understanding methods typically prioritize either geometric reconstruction or object-level perception, struggling to maintain both globally consistent geometry and coherent instance identities across hundreds to thousands of views. Our key insight is that these objectives are mutually beneficial: geometry provides a robust basis for cross-view instance association, while instance-level perception supplies constraints for refining geometry. SEGA decomposes a long sequence into overlapping clusters, reconstructs local geometry, and predicts 2D instance masks within each cluster. We introduce a 3D-Aware Alignment module that aligns these local predictions with a global proxy geometry and associates instances across clusters, producing temporally coherent video segmentation with globally consistent instance identities. We then apply a post processing step Instance-Aware Bundle Adjustment, which uses instance-consistent correspondences to refine the scene geometry. Finally, the predicted video masks are lifted into 3D instance proposals using the reconstructed geometry. We evaluate SEGA on ScanNet200 and ScanNet++v2 across two tasks: class-agnostic 3D instance segmentation and panoptic lifting for novel-view rendering. SEGA achieves state-of-the-art performance, improving Average Precision and Panoptic Quality over the strongest baselines. These results demonstrate the benefits of jointly modeling geometry and instance-level perception for robust indoor scene understanding from long multi-view sequences.

1 INTRODUCTION

Jointly reconstructing indoor scenes and decomposing them into individual 3D object instances is a long-standing goal in computer vision, with broad applications in robotics (Xia et al., 2026; Siddiqui et al., 2026; Maninis et al., 2022; Xu et al., 2024; Büchner et al., 2026; Boudjoghra et al., 2025), simulation (Delitzas et al., 2024; Huang et al., 2026), and embodied AI (Nguyen et al., 2025; Xu et al., 2024; Hu et al., 2025; Cen et al., 2025; Wang et al., 2024). Nevertheless, these objectives have traditionally been studied in isolation: reconstruction methods focus on recovering accurate camera poses and scene geometry, whereas 3D understanding methods focus on identifying and segmenting object instances.

Recent multi-view indoor geometric foundation models have begun to bridge this gap by extending feed-forward indoor 3D reconstruction with semantic and instance-level predictions (Zust et al., 2025; Li et al., 2025; Koch et al., 2025; Qu et al., 2026). However, two key limitations prevent these methods from producing coherent, scene-scale 3D instance segmentations. First, the interaction between geometry and segmentation is predominantly one-way: predicted geometry is used to lift or associate 2D masks, but is subsequently treated as fixed. Errors in camera poses or local geometry can therefore fragment observations of the same object or incorrectly merge nearby objects, while instance-consistent correspondences that could correct these errors remain unused. Second, these models are computationally expensive and typically operate on a limited number of input views – fewer than 50 frames, restricting their spatial coverage and their ability to maintain consistent object identities across longer indoor sequences.

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.

A natural strategy for scaling these models is to partition the input sequence into smaller, overlapping clusters and process each cluster independently (Cheng et al., 2026; Deng et al., 2025). Although computationally tractable, this strategy shifts the central challenge from local inference to global consolidation. Each cluster produces a partial reconstruction in a local coordinate frame, and imperfect alignment between clusters leads to geometric drift and discontinuities. Likewise, instance identities are defined only locally: the same object may receive different identities across clusters, while distinct but similar objects may be incorrectly merged. Naively combining these local outputs therefore produces fragmented geometry and inconsistent instance identities rather than a coherent scene-level decomposition.

To address these challenges, we introduce **SEGA**, a geometry-guided framework for indoor 3D instance segmentation from a multi-view image sequence. SEGA first partitions the long image sequence into overlapping clusters, reconstructs local geometry, and predicts 2D instance masks within each cluster. Then, it registers the local reconstructions to a shared global proxy geometry and associates instances across clusters. The resulting instance-consistent correspondences then guide the post-processing step to refine the scene geometry, after which the masks are lifted through the refined geometry to obtain scene-level 3D instance proposals.

The core contributions of this paper are:

- We present a **3D-Aware Alignment** module that aligns local reconstructions with a shared global proxy geometry, establishes dense 2D–3D correspondences, and reconciles instance identities across clusters to produce globally ID-consistent video segmentation.
- A post-processing **Instance-Aware Bundle Adjustment** step leverages globally consistent object identities obtained from proxy geometry to filter and reweight multi-view feature correspondences, suppressing cross-instance matches and improving scene geometry.
- We provide a comprehensive evaluation against strong baselines on two well-known indoor datasets: ScanNet200 (Rozenberszki et al., 2022), ScanNet++v2 (Yeshwanth et al., 2023) across two benchmarks: 3D instance segmentation under predicted- and oracle-geometry settings, downstream Panoptic Lifting for novel-view panoptic rendering. The results demonstrate the effectiveness and versatility of SEGA on challenging indoor sequences containing hundreds to thousands of views.

The key algorithmic principle of SEGA is treating geometry and identity association as *mutually constraining variables* rather than sequential stages, achieving better 3D instance segmentation of indoor scenes from a long multi-view image sequence.

2 Related work

3D scene reconstruction aims to recover a globally consistent scene representation along with camera poses and depth given a sequence of RGB. Prior work primarily differs in how long and complex sequences are processed, and can be broadly categorized into two groups. **The first group** is the streaming reconstruction methods (Zhuo et al., 2025; Chen et al., 2026; Shen et al., 2025), which sequentially and incrementally updates a global scene representation as new observations arrive. This strategy is memory efficient and naturally applicable to long videos. However, since the scene is built progressively, errors may accumulate over time, leading to drift and limited global consistency, especially in complex environments with long-range dependencies or loop closures. Another key challenge is memory retention (Chen et al., 2025; Wang et al., 2025b): as the sequence grows longer, the model must decide how much past information to preserve, compress, or discard, making it difficult to maintain long-range geometric and object-level consistency. **The second group** is chunk-based reconstruction methods (Deng et al., 2025; Zhang et al., 2026; Cheng et al., 2026), dividing the input sequence into smaller, often overlapping subsets. Each chunk is reconstructed independently using a multi-view geometry estimation model, such as VGGT (Wang et al., 2025a), MUSt3R (Cabon et al., 2025), or π^3 (Wang et al., 2025c), and the resulting local reconstructions are then aligned into a shared global coordinate system. This strategy enables the use of stronger but more memory-intensive reconstruction models and often yields higher-quality local geometry. However, it introduces a new challenge: consistently aligning scale, pose, and geometry across independently reconstructed chunks.

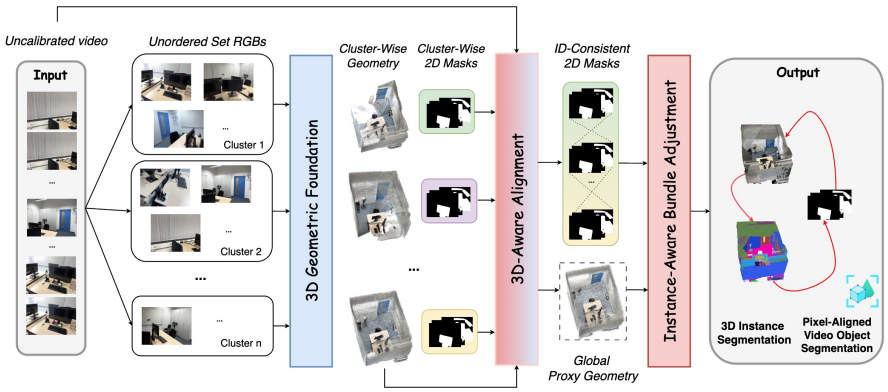

Figure 1: **Overview of SEGA**. We decompose a given RGB video sequence into overlapping frame clusters and apply a geometric foundation model to obtain cluster-wise geometries and 2D segmentation masks. Our 3D-Aware Alignment module then aligns the fragmented cluster-wise predictions into a unified global proxy geometry and temporally consistent video object segmentation. These consistent 2D priors, together with the proxy geometry, are further used to refine the scene representation through Instance-Aware Bundle Adjustment post-processing step. The final output is a globally consistent 3D reconstruction with temporally coherent, ID-consistent object segmentation, which are used to obtain 3D instance segmentation.

3D instance segmentation's goal is to parse a 3D environment into semantic categories and object-level instances, enabling structured reasoning over the captured scenes. Existing methods can be broadly grouped according to whether 3D geometry is provided as input. **Geometry-assisted methods** assume the access to precomputed geometry, such as point clouds, meshes, or depth maps with known camera poses. These methods (Nguyen et al., 2024; Yin et al., 2024; Yang et al., 2023; Nguyen et al., 2025) typically apply 2D semantic or instance segmentation independently to each frame using strong visual foundation models (Liu et al., 2024; Ke et al., 2023; Zhou et al., 2022), and then lift and fuse the frame-wise predictions into 3D object proposals. While this paradigm benefits from mature 2D priors and avoids jointly solving for geometry, its performance is inherently bounded by the quality of the input reconstruction (Felzenszwalb & Huttenlocher, 2004). Errors in camera poses, depth, or cross-view alignment can directly propagate to the final 3D predictions. Moreover, because the 2D predictions are produced frame by frame (Ke et al., 2023), these methods often lack explicit temporal consistency, leading to fragmented or inconsistent object identities across long video sequences. **Geometry-free methods**, in contrast, do not rely on precomputed geometry. Instead, they infer 3D structure and instance segmentation jointly from RGB observations in a feed-forward manner (Zust et al., 2025; Li et al., 2025; Koch et al., 2025; Li et al., 2026; Qu et al., 2026). By directly predicting geometry together with semantic or instance-level labels, these methods provide a unified formulation that can couple reconstruction and understanding more tightly. But scaling such models to long, temporally complex videos remains challenging, as they must maintain both global geometric consistency and coherent object-level understanding over frames.

3 Approach

We briefly review the key geometric components of SEGA in Appendix A.1, before presenting the full method.

Problem Definition. Given a sequence of N RGB images $\{\mathbf{I}_i\}_{i=1}^N$, where each image $\mathbf{I}_i \in \mathbb{R}^{H\times W\times 3}$, our SEGA pipeline $f(\cdot)$ aims to jointly recover a unified 3D reconstruction and scene understanding representation:

$$f(\{\mathbf{I}_i\}_{i=1}^N) = \{\mathbf{D}_i, \mathbf{C}_i, \boldsymbol{\kappa}_i, [\mathbf{R}_i \mid \mathbf{t}_i], \mathbf{M}_i\}_{i=1}^N, \tag{1}$$

where $\mathbf{D}_i$ denotes the estimated depth map for the i-th view, $\mathbf{C}_i$ is the corresponding depth confidence map, $\boldsymbol{\kappa}_i$ represents the camera intrinsics, $[\mathbf{R}_i \mid \mathbf{t}_i]$ denotes the camera extrinsics , and $\mathbf{M}_i$ denotes the predicted video segmentation masks. We obtain the 3D point cloud instance segmentation by unprojecting the depth maps. Together, these outputs provide both geometric reconstruction and instance-level understanding across the input sequence.

3.1 Overview of Our Proposed SEGA

Our method is described in Fig. 1, designed to process multiview image collections and jointly produce scene-level 3D reconstruction and instance understanding from sequences containing up to thousands of frames.

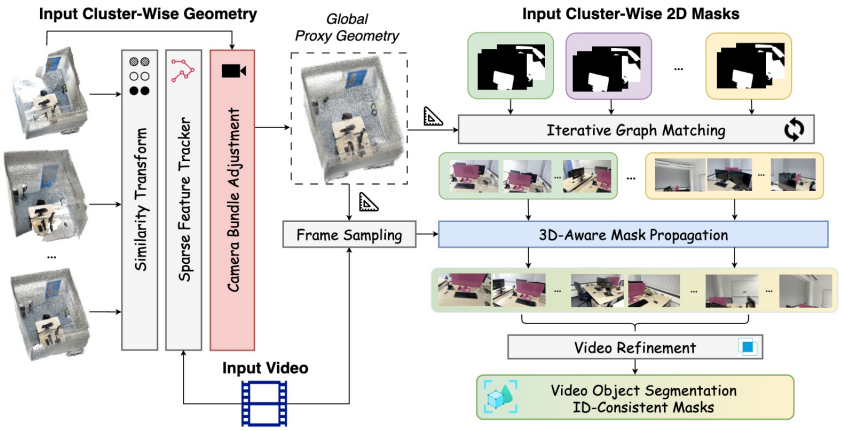

Figure 2: **3D-Aware Alignment**. Given cluster-wise geometry and 2D masks, we first align local reconstructions into a shared proxy geometry through similarity transformation, sparse feature tracking, and camera bundle adjustment. Using this proxy geometry, we iteratively associate cluster-wise 2D masks via graph matching, sample object-aware frames for propagation, and refine the resulting masks to obtain temporally coherent, ID-consistent video object segmentation.

We follow Cheng et al. (2026) to decompose the input video into smaller K overlapping clusters before applying geometric foundation models. Specifically, we construct a visual similarity matrix over all image pairs and sample approximate maximum Hamiltonian paths to form compact image clusters with strong viewpoint diversity. We additionally enforce temporal overlap between neighboring frames, allowing local reconstructions to share common frames for subsequent alignment and merging. This partitioning strategy makes inference memory and compute efficient while preserving global scene coverage across the full input sequence.

For each frame cluster $k \leq K$, we construct an initial local 3D representation using an off-the-shelf multi-view 3D geometric foundation model (Zust et al., 2025). This produces cluster-wise geometric predictions, including depth maps $\mathbf{D}^k$, confidence maps $\mathbf{C}^k$, camera intrinsics $\boldsymbol{\kappa}^k$, camera extrinsics $[\mathbf{R}^k \mid \mathbf{t}^k]$, and cluster-wise 2D instance segmentation masks $\mathbf{M}^k$.

Our **3D-Aware Alignment** module, described in Section 3.2, aligns cluster-wise geometries into a shared coordinate system; aggregates and propagates local masks into globally ID-consistent object tracks. By using the proxy geometry and camera correspondences as common priors, this module associates masks across large viewpoint changes and partial observations where purely 2D matching is unreliable. Its output is a globally aligned proxy geometry together with video-level object masks.

Then a post-processing step **Instance-Aware Bundle Adjustment**, described in Section 3.3, refines the geometry using the recovered instance-level consistency. We observe that ID-consistent objects act as stable 3D entities whose spatial support should remain coherent across views. Therefore, we use the globally consistent masks obtained from the proxy geometry to construct instance-aware correspondences, which regularize bundle adjustment beyond sparse feature tracks. This refinement improves local geometric consistency while preserving the global structure established by the alignment module.

Together, these two modules transform fragmented cluster-wise predictions into a unified scene-level representation with globally consistent geometry and ID-consistent video object segmentation.

3.2 3D-Aware Alignment

The goal of this module, as described in Fig. 2, is to transform a set of independently reconstructed cluster-wise geometries into a globally consistent representation while simultaneously establishing identity-consistent object masks across the full video. Concretely, given cluster-wise predictions $\{\mathbf{D}^k, \mathbf{C}^k, \boldsymbol{\kappa}^k, [\mathbf{R}^k|\mathbf{t}^k], \mathbf{M}^k\}_{k=1}^K$ and the input video, we produce (i) a unified proxy geometry in a shared coordinate system and (ii) temporally consistent video object segmentations with globally aligned instance identities.

Camera Bundle Adjustment. For each cluster k, we back-project depth maps into 3D point maps using π^{-1} from Eq. (10), and filter unreliable points with confidence maps $\mathbf{C}^k$. Since clusters are reconstructed in independent coordinate frames, we align neighboring clusters $(k, k+1)$ by estimating a similarity transform $\mathbf{T}_{k\to k+1} \in \mathrm{Sim}(3)$ via the robust objective in Eq. (12). Chaining

these transforms maps all cluster cameras into a shared global coordinate system, producing globally aligned but still imperfect geometry due to partial overlap and reconstruction noise.

To address these misalignments, we construct multi-view feature tracks using the sparse feature tracker described in Appendix A.1, and perform differentiable gradient-based bundle adjustment to jointly optimize camera poses and 3D structure. Each track $\mathcal{M} \in \mathcal{T}_{\text{sparse}}$ is associated with a 3D anchor point $\mathbf{X}_p$ and a set of 2D observations $\{u_{i,p}\}$ in the frames where it is visible. Given these anchors and their cross-frame observations, we minimize a confidence-aware robust reprojection objective:

$$\mathcal{L}_{\text{BA}} = \sum_{(i,p)\in\mathcal{T}_{\text{sparse}}} w_p \cdot \rho\big(\pi(\mathbf{X}_p; \boldsymbol{\kappa}_i, \mathbf{R}_i, \mathbf{t}_i) - u_{i,p}\big), \tag{2}$$

where $\rho(\cdot)$ denotes a Huber loss, and w_p encodes the reliability of each track derived from depth and tracking confidence. We parameterize camera rotations using unit quaternions and optimize translations, camera intrinsics and 3D point tracks in a fully differentiable framework. Optimization is performed using Adam with a cosine annealing schedule, which provides stable convergence under noisy and partially overlapping observations.

In this stage, we optimize only the camera parameters and point anchors, while the depth and confidence maps are carried over unchanged. The resulting optimized cameras provide a globally consistent geometric scaffold that serves as the foundation for subsequent proxy geometry initialization.

Global Proxy Geometry. Using the optimized cameras from bundle adjustment, we define the global proxy geometry as $\mathcal{P} = \{\mathbf{D}_\mathcal{P}, \mathbf{C}_\mathcal{P}, \boldsymbol{\kappa}_\mathcal{P}, [\mathbf{R}_\mathcal{P}|\mathbf{t}_\mathcal{P}]\}$. This representation lets us build dense 2D–3D correspondences between the shared geometry and all video frames. Specifically, we back-project each depth map into a globally aligned point map using π^{-1}. For a pixel u in frame i, its reprojection into frame j through the proxy geometry is:

$$\tilde{u}_{j\leftarrow i}(u) = \pi\left(\pi^{-1}\left(u, \mathbf{D}_{\mathcal{P},i}(u); \boldsymbol{\kappa}_{\mathcal{P},i}, \mathbf{R}_{\mathcal{P},i}, \mathbf{t}_{\mathcal{P},i}\right); \boldsymbol{\kappa}_{\mathcal{P},j}, \mathbf{R}_{\mathcal{P},j}, \mathbf{t}_{\mathcal{P},j}\right), \tag{3}$$

where $\tilde{u}_{j\leftarrow i}(u)$ denotes the projection of pixel u from source frame i to target frame j via the shared 3D geometry. We keep only geometrically valid correspondences whose projected pixels lie within image bounds, have positive depth, and satisfy:

$$|z_{j\leftarrow i}(u) - \mathbf{D}_{\mathcal{P},j}(\tilde{u}_{j\leftarrow i}(u))| \le \tau_{\text{depth}}, \tag{4}$$

where $z_{j\leftarrow i}(u)$ is the depth of the back-projected 3D point viewed from frame j, and τ_{depth} is a predefined threshold.

This yields a geometric correspondence field Ω linking the proxy geometry to each image. The resulting dense pixel–point relationships allow 2D mask proposals to be lifted into a common 3D coordinate system for robust cross-frame association, and later serve as dense constraints for instance-aware refinement.

Iterative Graph Matching. Mask predictions are locally coherent within each cluster, but independently processed clusters may assign different IDs to the same physical object. Since purely 2D matching is unreliable under large viewpoint changes, partial overlap, and occlusion, we use the geometric correspondence field Ω from the proxy geometry $\mathcal{P}$ to associate masks in 3D.

At iteration t, we match the current cluster to the set of previously accumulated object groups. Here, an *object group* $\mathcal{G}_a$ denotes a collection of masks from earlier clusters that have been assigned to the same physical object, while a *candidate mask* m_b is an instance mask predicted in the current cluster. We construct a bipartite graph between existing object groups $\mathcal{V}_L = \{\mathcal{G}_a\}$ and candidate masks $\mathcal{V}_R = \{m_b\}$, treating masks within the current cluster as separate nodes because their identities are already locally consistent.

Each node is represented by its lifted 3D support. For a candidate mask $m_b \in \mathcal{V}_R$ in frame i, we define its 3D support:

$$\Omega_i^{2D\to 3D}(m_b) = \left\{\mathbf{X} \in \mathcal{P} \;\middle|\; \Omega_i^{3D\to 2D}(\mathbf{X}) \in m_b\right\}. \tag{5}$$

For an object group $\mathcal{G}_a$, we aggregate the supports of its assigned masks and compute its 3D IoU affinity with m_b:

$$\Omega^{2D\to 3D}(\mathcal{G}_a) = \bigcup_{m\in\mathcal{G}_a} \Omega^{2D\to 3D}(m), \quad w_{a,b} = \frac{\left|\Omega^{2D\to 3D}(\mathcal{G}_a) \cap \Omega_i^{2D\to 3D}(m_b)\right|}{\left|\Omega^{2D\to 3D}(\mathcal{G}_a) \cup \Omega_i^{2D\to 3D}(m_b)\right|}. \tag{6}$$

The lifting and 3D IoU computation are implemented as batched tensor operations. We then solve a maximum-weight bipartite matching problem using the Hungarian algorithm:

$$\mathbf{H}^\star = \arg\max_{\mathbf{H}} \sum_{a,b} \mathbf{H}_{a,b} w_{a,b}, \quad \text{s.t. } \sum_b \mathbf{H}_{a,b} \le 1,\ \sum_a \mathbf{H}_{a,b} \le 1, \tag{7}$$

where $\mathbf{H}_{a,b} \in \{0,1\}$ indicates whether object group $\mathcal{G}_a$ is matched to candidate mask m_b. Matches with $w_{a,b} \ge \tau_{\text{3D-IoU}}$ are merged into the corresponding object group, while unmatched masks initialize new groups. Repeating this process across clusters progressively aligns instance identities and yields globally consistent object groups.

3D-Aware Mask Propagation. Although iterative graph matching resolves cross-cluster identity inconsistencies, the matched masks remain limited by the quality and coverage of cluster-wise predictions. In particular, lifted 3D IoU may miss complete object extents when different views observe disjoint surfaces, e.g., the front and back of a monitor (Fig. 2). Thus, matched object groups serve as reliable but incomplete 3D anchors, rather than final video-level masks. To obtain complete and temporally consistent masks, we use each matched object group as a 3D-aware prompt for SAM3 video mask propagation (Carion et al., 2025). For an object group $\mathcal{G}_a$, we collect its accumulated 3D support $\mathcal{Q}_a = \Omega^{2D\to 3D}(\mathcal{G}_a)$ and identify geometrically visible frames by $\nu_i(\mathcal{Q}_a) = |\mathcal{Q}_a \cap \mathcal{V}_i|/|\mathcal{Q}_a|$, where $\mathcal{V}_i$ is the set of proxy geometry points visible in frame i according to $\Omega_i^{3D\to 2D}$. **Object-aware Frames sampling** is employed with $\nu_i(\mathcal{Q}_a) > 0$ to form the object-specific propagation sequence.

For each selected frame, we use an existing mask in $\mathcal{G}_a$ as a positive prompt when available. Otherwise, we project the visible support $\mathcal{Q}_a$ into the frame via $\Omega_i^{3D\to 2D}$ and sample $\rho = 3$ representative point prompts using Furthest Point Sampling. This object-aware sampling restricts propagation to frames with visible 3D support, substantially reducing computational cost, as shown in Table 4 and avoiding irrelevant views while preserving the observations needed to complete the mask. During propagation, we further enforce 3D consistency with negative prompts. Given a propagated mask $\hat{m}_i$, we define its 3D-supported region as

$$\hat{m}_i^{3D} = \left\{ u \in \hat{m}_i \;\middle|\; \exists \mathbf{X} \in \mathcal{Q}_a, \Omega_i^{3D\to 2D}(\mathbf{X}) = u \right\}. \tag{8}$$

If $\hat{m}_i$ extends beyond this supported region, or if the supported region is weak or empty, we treat it as geometrically inconsistent and sample negative points from the inconsistent regions. These negative prompts discourage unsupported masks, reducing outliers and propagation drift. This stage expands sparse 3D object anchors into dense, ID-consistent video masks across the full sequence. Since each object is propagated independently, the procedure is fully parallelizable and can efficiently utilize multiple GPUs for scalability.

Video Refinement. Although 3D-aware propagation produces temporally consistent masks, propagation noise and partial observations can still create duplicate or fragmented identities. We therefore refine the propagated results by consolidating spatio-temporal object tracks across views. Specifically, we build an overlap graph whose nodes are object tracks, and connect two tracks if their frame-wise mask IoU exceeds $\tau_{\text{2D-IoU}}$ in at least λ shared frames. We then compute connected components of this graph, treating each component as a single physical object, and merge its member tracks by taking the union of their masks at each frame. This multi-view refinement removes duplicate or fragmented tracks, yielding cleaner and more complete video-level object segmentation masks $\{\mathbf{M}_i\}_{i=1}^N$.

3.3 Instance-Aware Bundle Adjustment (IA-BA)

While the initial camera bundle adjustment relies on sparse appearance-based feature tracks, such tracks may contain unreliable correspondences, particularly near object boundaries, under occlusion, or between visually similar objects. After obtaining globally ID-consistent instance masks from our 3D-Aware Alignment module, we use these masks obtained from the proxy geometry to filter unreliable feature tracks.

Concretely, let $\mathcal{T}_{\text{sparse}}$ denote the sparse feature tracks used in the initial bundle adjustment, where $u_{i,p}$ is the observed location of track p in frame i. For each cross-view correspondence $(u_{i,p}, u_{j,p})$, we retain the track only when the two observations belong to the same propagated instance, i.e., $\mathbf{M}_i(u_{i,p}) = \mathbf{M}_j(u_{j,p})$. The retained tracks form the instance-consistent track set $\mathcal{T}_{\text{inst}} \subseteq \mathcal{T}_{\text{sparse}}$.

We then perform bundle adjustment using the same robust reprojection objective as in Eq. (2), replacing the original sparse tracks with the filtered instance-consistent tracks:

$$\mathcal{L}_{\text{IA-BA}} = \sum_{(i,p)\in\mathcal{T}_{\text{inst}}} w_{i,p}\, \rho\left(\pi(X_p; \kappa_i, R_i, t_i) - u_{i,p}\right). \tag{9}$$

Here, $w_{i,p}$ combines feature reliability, geometric confidence, and instance consistency. Compared with standard sparse-track bundle adjustment, IA-BA suppresses correspondences that cross instance boundaries or associate different physical objects, yielding more reliable object-consistent constraints for refining scene geometry.

4 EXPERIMENTS

Dataset. We evaluate SEGA on challenging indoor benchmarks: ScanNet200 (Rozenberszki et al., 2022) and ScanNet++v2 (Yeshwanth et al., 2023). Both provide calibrated RGB-D sequences, reconstructed 3D meshes, and dense segmentation annotations. ScanNet200 includes 312 validation scenes over 198 instance categories, with about 2K views and 200K reconstructed points per scene. ScanNet++v2 includes 50 validation scenes over 100 instance categories, with sharper images, less motion blur, and denser reconstructions, averaging about 5K views and 4M points per scene. *We focus these two indoor benchmarks because their frequent viewpoint revisits and dense annotations provide a controlled setting to rigorously evaluate complex ID consistency across the sequence*. Additional details are provided in Appendix A.4.

Implementation details. Each cluster processed by the 3D geometric foundation model PanSt3R (Zust et al., 2025) contains at most 50 frames. For SAM3 (Carion et al., 2025), we use the publicly released checkpoint[1]. Bundle adjustment in both stages is optimized with a learning rate of 1×10^{-4} for 3,000 iterations. We set the depth-consistency threshold for the correspondence field Ω to $\tau_{\text{depth}} = 0.1$, the 3D IoU threshold for iterative graph matching to $\tau_{\text{3D-IoU}} = 0.4$, and the video refinement thresholds to $\tau_{\text{2D-IoU}} = 0.95$ and $\lambda = 5$. All experiments are run on a node with 8 NVIDIA L40s GPUs. We provide a comprehensive hyperparameter analysis in Appendix A.2.

4.1 3D INSTANCE SEGMENTATION

We evaluate SEGA by considering the challenging task of class-agnostic 3D point cloud instance segmentation. We report results under two settings: the traditional 3D segmentation protocol, which requires ground-truth geometry (✓) (Schult et al., 2023; Yeshwanth et al., 2023; Rozenberszki et al., 2022), and our *pixel-aligned evaluation protocol*, which does not (×). We also adapt several well-known baselines (Felzenszwalb & Huttenlocher, 2004; Carion et al., 2025; Nguyen et al., 2024; Zust et al., 2025; Li et al., 2025) for fair comparison. We report the Average Precision $\mathbf{AP}$ and Average Recall $\mathbf{AR}$ averaged over IoU thresholds from 50% to 95% in 5% increments, along with $\mathbf{AP}_{50}$/$\mathbf{RC}_{50}$ and $\mathbf{AP}_{25}$/$\mathbf{RC}_{25}$ at IoU thresholds of 50% and 25%, respectively.

Baselines. For Direct Lifting, SAM3D (Yang et al., 2023), and Open3DIS (Nguyen et al., 2024), we first obtain per-frame 2D masks using SAM3 (Carion et al., 2025), and then lift them into 3D using different grouping strategies: no IoU-based merging for Direct Lifting, iterative IoU-based merging for SAM3D, and agglomerative IoU-based merging for Open3DIS. We also compare with PanSt3R (Zust et al., 2025) by applying it to chunked frames and concatenating its predictions over the full video. In the setting with ground-truth geometry, for SEGA, we directly lift our temporally consistent 2D proposals to 3D using superpoints, without any IoU-based merging. For Open3DIS, we adopt superpoint-based lifting with agglomerative merging (Table 1, line 9), while for SAM3D, we directly lift 2D masks to the 3D point cloud without superpoint supervision (Table 1, line 8). We also provided some controlled comparison with IGGT (Li et al., 2025) in the Appendix A.3.3.

Results. Without ground-truth geometry, SEGA substantially outperforms prior methods, achieving 17.1 AP on ScanNet200 and 18.1 AP on ScanNet++v2, compared to 7.9 and 7.7 AP from Open3DIS. Under the oracle-geometry setting, SEGA further improves to 31.5 and 28.4 AP, outperforming Open3DIS by 2.4 and 6.6 AP, respectively. These gains demonstrate the importance of globally ID-consistent masks from 3D-aware alignment and propagation, beyond improved geometry alone. Qualitative results are shown in Fig. 3b-c, with further analysis and baseline details in Appendix A.3.3.

[1] https://huggingface.co/facebook/sam3

Method	GT Geometry	ScanNet200						ScanNet++v2					
		AP↑	AP_{50}↑	AP_{25}↑	AR↑	RC_{50}↑	RC_{25}↑	AP↑	AP_{50}↑	AP_{25}	AR↑	RC_{50}↑	RC_{25}↑
1. Direct Lifting	×	1.6	3.5	8.9	4.9	7.2	13.1	0.3	0.9	1.4	3.5	4.6	5.9
2. SAM3D	×	5.2	12.7	24.7	16.6	33.3	52.3	3.3	8.8	18.9	10.9	24.4	43.4
3. Open3DIS	×	7.9	18.3	34.5	17.3	32.1	47.3	7.7	16.5	29.8	14.2	25.4	37.1
4. PanSt3R	×	2.1	4.2	11.1	5.6	10.3	16.1	1.5	2.9	5.7	5.2	8.9	10.6
5. SEGA	×	**17.1**	**32.7**	**48.4**	**30.8**	**56.6**	**79.4**	**18.1**	**33.8**	**47.2**	**32.3**	**57.8**	**76.7**
6. Direct Lifting	✓	4.8	15.3	40.1	22.5	53.4	79.5	4.3	10.1	28.9	18.5	37.2	65.9
7. Felzenszwalb	✓	4.5	9.3	26.1	8.9	18.3	50.3	4.2	8.9	22.1	8.2	17.4	42.4
8. SAM3D	✓	6.7	21.7	46.4	21.6	49.9	77.6	4.7	13.0	33.0	17.6	36.8	60.6
9. Open3DIS	✓	29.1	44.0	49.1	55.5	83.2	92.1	21.8	38.9	46.5	42.3	75.1	88.9
10. PanSt3R	✓	9.9	28.4	48.1	20.1	45.2	69.9	8.2	25.3	40.1	24.7	49.9	72.1
11. SEGA	✓	**31.5**	**48.2**	**55.5**	**58.6**	**85.2**	**93.1**	**28.4**	**43.9**	**49.3**	**54.2**	**82.7**	**92.1**

Table 1: **3D Class-Agnostic Instance Segmentation** on ScanNet200 and ScanNet++v2. We compare our SEGA with prior 2D-to-3D lifting approaches under two settings: without ground-truth (GT) geometry (top) and with ground-truth geometry (bottom). The **best** and second best results are marked in **bold** and underline.

Method	PQ↑	SQ↑	RQ↑
PanSt3R (Zust et al., 2025)	16.45	66.92	23.85
SAM3 (Carion et al., 2025)	37.35	79.52	45.53
SEGA	**48.94**	**80.85**	**59.58**

(a) COLMAP initialization.

Method	Test PSNR↑	PQ↑	SQ↑	RQ↑
Ours (w/o IA-BA)	22.35	45.43	78.44	55.86
Ours (w/ IA-BA)	**23.23**	**46.86**	**78.82**	**57.33**

(b) Predicted point-cloud initialization.

Table 2: **Evaluation on panoptic lifting** on ScanNet++v2 with Gaussian Grouping.

# Views (N)	# of GPUs	Cluster-wise	3D Align.	IA-BA	Total (s)
50	8	25	246	253	524
100		45	425	274	744
400		145	1,222	439	1,806
800		310	2,558	658	3,526
1600		538	5,462	901	6,901
1600	12	347	2,978	901	4,226
1600	16	271	1,563	901	2,735

Table 3: **Runtime analysis** in seconds for different numbers of input views and GPUs.

4.2 Panoptic Lifting for Novel View Rendering

We further evaluate whether the predicted video object segmentation can be lifted into a 3D radiance-field representation for novel-view panoptic rendering (Siddiqui et al., 2023). We adopt Gaussian Grouping (Ye et al., 2024) as the lifting backbone with three different instance segmentation methods: concatenated PanSt3R (Zust et al., 2025), SAM3 (Carion et al., 2025) and our SEGA. We report Panoptic Quality (**PQ**), Segmentation Quality (**SQ**), and Recognition Quality (**RQ**) (Kirillov et al., 2019). PQ jointly measures instance matching and mask overlap, while SQ captures the mask quality of matched instances and RQ penalizes missed and spurious instances. We additionally report Peak Signal-to-Noise Ratio (**PSNR**) to measure the pixel-level fidelity of rendered images against the ground-truth views, where higher values indicate better reconstruction quality.

Results. As shown in Tab. 2, we offer two panoptic-lifting settings: (a) COLMAP (Schonberger & Frahm, 2016) initialization; (b) predicted cameras and reconstructed point cloud from SEGA initialization. In Tab. 2a, SEGA achieves the best PQ, SQ, and RQ among methods using predicted masks. PanSt3R suffers from cross-cluster identity fragmentation, while SAM3 improves through video-level propagation but lacks the 3D-aware alignment needed for reliable long-range identities. In contrast, SEGA maintains globally consistent object identities across the full sequence, leading to better downstream 3D panoptic lifting. In Tab. 2b, IA-BA is applied after the ID-consistent masks have been constructed, consuming these masks to refine geometry. Consequently, it benefits novel-view rendering by 0.88 PSNR and geometry-dependent downstream lifting. Qualitative results are shown in Fig. 3d, with additional details and analysis in Appendix A.3.4.

4.3 Ablation Studies

To validate the design choices of SEGA, we conduct ablation studies on the ScanNet++v2 validation set under the *pixel-aligned evaluation protocol* for 3D class-agnostic instance segmentation.

Runtime Analysis. We report the model latency and total runtime on ScanNet++v2 in Table 3. All experiments are benchmarked on a server with a total of 144 CPU cores, 512GB of RAM, and 8 NVIDIA L40s GPUs. Distributed computation is applied on the two stages: cluster-wise geom-

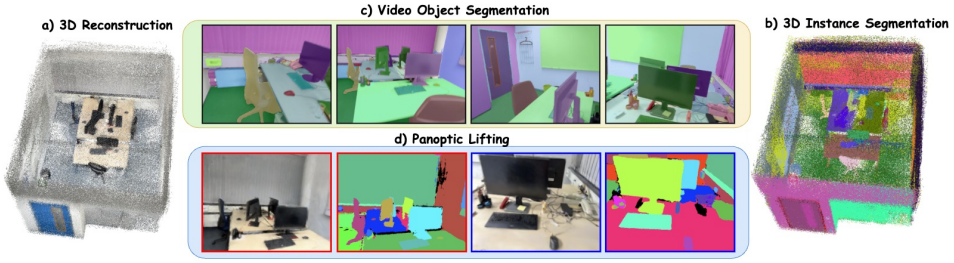

Figure 3: **Qualitative results.** SEGA produces a globally consistent 3D reconstruction (**a**), 3D instance segmentation (**b**), and temporally consistent video object segmentation (**c**) from an RGB video. (**d**) presents panoptic lifting results by lifting video object segmentation in (**c**) into the 3D radiance-field representation.

Frame samp.	Iter. graph	3D-aware prompt.	Mask prop.	Video refine.	AP	Runtime (s)
×	✓	✓	✓	✓	17.12	4,135
✓	×	✓	✓	✓	15.97	3,926
✓	✓	×	✓	✓	15.62	1,462
✓	✓	×	×	✓	6.83	595
✓	✓	✓	✓	×	16.15	1,714
✓	✓	✓	✓	✓	18.11	1,806

Table 4: **Effect of each component** measured under the setting without GT geometry on ScanNet++v2.

etry and 2D segmentation mask prediction with 3D geometric foundation model (up to 50 views per forward pass), and 3D-Aware Alignment. Instance-Aware Bundle Adjustment is performed on a single GPU. Thanks to this distributed pipeline design, SEGA is able to process long video sequences and large numbers of input views, and can further improve throughput when *more GPUs are provided*. In contrast, such behavior is difficult to achieve with most existing 3D reconstruction and understanding methods, as their computational cost grows quadratically with sequence length and the inference pipelines are *not naturally amenable to parallel execution*.

Study on the effect of each component. We ablate the main components of the 3D-Aware Alignment module in Table 4, reporting **AP** and runtime with $N = 400$ input frames. Mask propagation is the most critical component: removing it causes the largest AP drop, showing the importance of temporal consistency for 3D instance segmentation. Object-aware frame sampling and iterative graph matching are also important, as they use the geometric correspondence field Ω to guide 3D-aware sampling and cross-view association, improving consistency while reducing unnecessary computation. The 3D-aware prompting strategy and video refinement further improve overall quality.

Additional Analysis. To provide a more comprehensive analysis of SEGA, we further study open-vocabulary 3D instance segmentation, the effect of reconstruction noise, additional experiments on class-agnostic 3D instance segmentation, detailed setup of panoptic lifting, and quantitative analysis on camera-pose and point-cloud reconstruction in Appendix A.3. We also include a supplementary video to provide additional qualitative comparisons.

5 Conclusion

We introduce SEGA, a framework for indoor 3D instance segmentation from multiview images. By treating geometry and perception as mutually reinforcing signals, SEGA aligns fragmented cluster-wise predictions into a globally consistent scene representation and recovers temporally coherent, ID-consistent object segmentation.

Limitations and Future Work: Although SEGA performs well on indoor 3D instance segmentation and panoptic lifting, its effectiveness depends on the quality of the intermediate proxy geometry. Under noisy geometry caused by object motion (dynamic scenes) or severe occlusion, identity matching may become unreliable, with errors propagating to mask propagation and instance-aware refinement. Improving alignment and association under such weak geometric cues is therefore important for future work. Building on the foundation of SEGA's globally aligned geometry and persistent instance identities, incorporating explicit motion models and physical priors is a promising direction toward unified 4D scene reconstruction and understanding.

AI USE STATEMENT

Generative AI tools were used solely to refine author-written text for grammar, clarity, concision, and organization. They were not used to generate research ideas, design the method, implement code, conduct experiments, analyze data, produce results, create figures, or formulate scientific claims. All AI-assisted edits were reviewed and verified by the authors, who take full responsibility for the final content of this paper.

ETHICS STATEMENT

This work uses publicly released indoor-scene datasets in accordance with their respective licenses and does not collect new personal data or involve human-subject experiments. Because SEGA reconstructs indoor environments and identifies individual object instances, its outputs may contain geometric errors, incomplete objects, or incorrect instance associations and should not be used directly for safety-critical robotics or decision-making without independent verification. Indoor reconstruction and object-level scene understanding may also raise privacy concerns if applied to sensitive or private environments. We present SEGA for research in 3D scene reconstruction and understanding and encourage its responsible use in compliance with applicable privacy, data-use, and disclosure requirements.

REPRODUCIBILITY STATEMENT

The main paper specifies the cluster-wise reconstruction pipeline, 3D-Aware Alignment module, global proxy geometry construction, cross-cluster instance association, video-mask propagation, and Instance-Aware Bundle Adjustment. It also describes the datasets, baselines, evaluation protocols, and metrics used for 3D instance segmentation, camera-pose estimation, point-cloud reconstruction, and panoptic lifting. The appendix provides additional implementation details, hyperparameters, baseline configurations, predicted- and oracle-geometry evaluation protocols, open-vocabulary experiments, qualitative results, and runtime analyses. Code, configurations, and evaluation scripts will be released upon publication to support faithful reproduction of our results.

REFERENCES

Mohamed el amine Boudjoghra, Angela Dai, Jean Lahoud, Hisham Cholakkal, Rao Anwer, Salman Khan, and Fahad Khan. Open-yolo 3d: Towards fast and accurate open-vocabulary 3d instance segmentation. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), International Conference on Learning Representations, volume 2025, pp. 65143–65156, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/a3a661eb3308d0bb686f6a4bac521032-Paper-Conference.pdf.

Martin Büchner, Adrian Röfer, Tim Engelbracht, Tim Welschehold, Zuria Bauer, Hermann Blum, Marc Pollefeys, and Abhinav Valada. Articulated 3d scene graphs for open-world mobile manipulation. arXiv preprint arXiv:2602.16356, 2026.

Yohann Cabon, Lucas Stoffl, Leonid Antsfeld, Gabriela Csurka, Boris Chidlovskii, Jerome Revaud, and Vincent Leroy. Must3r: Multi-view network for stereo 3d reconstruction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1050–1060, 2025.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. arXiv preprint arXiv:2511.16719, 2025.

Jiazhong Cen, Jiemin Fang, Zanwei Zhou, Chen Yang, Lingxi Xie, Xiaopeng Zhang, Wei Shen, and Qi Tian. Segment anything in 3d with radiance fields. International Journal of Computer Vision, 133(8):5138–5160, 2025.

Lin-Zhuo Chen, Jian Gao, Yihang Chen, Ka Leong Cheng, Yipengjing Sun, Liangxiao Hu, Nan Xue, Xing Zhu, Yujun Shen, Yao Yao, et al. Geometric context transformer for streaming 3d reconstruction. arXiv preprint arXiv:2604.14141, 2026.

Xingyu Chen, Yue Chen, Yuliang Xiu, Andreas Geiger, and Anpei Chen. Ttt3r: 3d reconstruction as test-time training. arXiv preprint arXiv:2509.26645, 2025.

Leo Kaixuan Cheng, Abdus Shaikh, Ruofan Liang, Zhijie Wu, Yushi Guan, and Nandita Vijaykumar. Merg3r: A divide-and-conquer approach to large-scale neural visual geometry. arXiv preprint arXiv:2603.02351, 2026.

Ingrid Daubechies, Ronald DeVore, Massimo Fornasier, and C Sinan Güntürk. Iteratively reweighted least squares minimization for sparse recovery. Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences, 63(1):1–38, 2010.

Alexandros Delitzas, Ayca Takmaz, Federico Tombari, Robert Sumner, Marc Pollefeys, and Francis Engelmann. Scenefun3d: Fine-grained functionality and affordance understanding in 3d scenes. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14531–14542, 2024.

Kai Deng, Zexin Ti, Jiawei Xu, Jian Yang, and Jin Xie. Vggt-long: Chunk it, loop it, align it–pushing vggt's limits on kilometer-scale long rgb sequences. arXiv preprint arXiv:2507.16443, 2025.

Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp. 224–236, 2018.

Pedro F Felzenszwalb and Daniel P Huttenlocher. Efficient graph-based image segmentation. International journal of computer vision, 59(2):167–181, 2004.

Jie Hu, Shizun Wang, and Xinchao Wang. Pe3r: Perception-efficient 3d reconstruction. arXiv preprint arXiv:2503.07507, 2025.

Zhao Huang, Boyang Sun, Alexandros Delitzas, Jiaqi Chen, and Marc Pollefeys. React3d: Recovering articulations for interactive physical 3d scenes. IEEE Robotics and Automation Letters, 2026.

Lei Ke, Mingqiao Ye, Martin Danelljan, Yu-Wing Tai, Chi-Keung Tang, Fisher Yu, et al. Segment anything in high quality. Advances in Neural Information Processing Systems, 36:29914–29934, 2023.

Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 6399–6408, 2019.

Sebastian Koch, Johanna Wald, Hidenobu Matsuki, Pedro Hermosilla, Timo Ropinski, and Federico Tombari. Unified semantic transformer for 3d scene understanding. arXiv preprint arXiv:2512.14364, 2025.

Changyang Li, Xueqing Huang, Shin-Fang Chng, Huangying Zhan, Qingan Yan, and Yi Xu. Fast3dis: Feed-forward anchored scene transformer for 3d instance segmentation. arXiv preprint arXiv:2603.25993, 2026.

Hao Li, Zhengyu Zou, Fangfu Liu, Xuanyang Zhang, Fangzhou Hong, Yukang Cao, Yushi Lan, Manyuan Zhang, Gang Yu, Dingwen Zhang, et al. Iggt: Instance-grounded geometry transformer for semantic 3d reconstruction. arXiv preprint arXiv:2510.22706, 2025.

Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. Lightglue: Local feature matching at light speed. In Proceedings of the IEEE/CVF international conference on computer vision, pp. 17627–17638, 2023.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In European conference on computer vision, pp. 38–55. Springer, 2024.

Kevis-Kokitsi Maninis, Stefan Popov, Matthias Nießner, and Vittorio Ferrari. Vid2cad: Cad model alignment using multi-view constraints from videos. IEEE transactions on pattern analysis and machine intelligence, 45(1):1320–1327, 2022.

Phuc Nguyen, Tuan Duc Ngo, Evangelos Kalogerakis, Chuang Gan, Anh Tran, Cuong Pham, and Khoi Nguyen. Open3dis: Open-vocabulary 3d instance segmentation with 2d mask guidance. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 4018–4028, 2024.

Phuc Nguyen, Minh Luu, Anh Tran, Cuong Pham, and Khoi Nguyen. Open-ended 3d point cloud instance segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2580–2590, 2025.

Jinyuan Qu, Hongyang Li, and Lei Zhang. Segvggt: Joint 3d reconstruction and instance segmentation from multi-view images. arXiv preprint arXiv:2603.19926, 2026.

David Rozenberszki, Or Litany, and Angela Dai. Language-grounded indoor 3d semantic segmentation in the wild. In European conference on computer vision, pp. 125–141. Springer, 2022.

Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4104–4113, 2016.

Jonas Schult, Francis Engelmann, Alexander Hermans, Or Litany, Siyu Tang, and Bastian Leibe. Mask3d: Mask transformer for 3d semantic instance segmentation. In 2023 IEEE International Conference on Robotics and Automation (ICRA), pp. 8216–8223. IEEE, 2023.

You Shen, Zhipeng Zhang, Yansong Qu, Xiawu Zheng, Jiayi Ji, Shengchuan Zhang, and Liujuan Cao. Fastvggt: Training-free acceleration of visual geometry transformer. arXiv preprint arXiv:2509.02560, 2025.

Yawar Siddiqui, Lorenzo Porzi, Samuel Rota Buló, Norman Müller, Matthias Nießner, Angela Dai, and Peter Kontschieder. Panoptic lifting for 3d scene understanding with neural fields. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9043–9052, 2023.

Yawar Siddiqui, Duncan Frost, Samir Aroudj, Armen Avetisyan, Henry Howard-Jenkins, Daniel DeTone, Pierre Moulon, Qirui Wu, Zhengqin Li, Julian Straub, et al. Shaper: Robust conditional 3d shape generation from casual captures. arXiv preprint arXiv:2601.11514, 2026.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. arXiv preprint arXiv:2508.10104, 2025.

Ayça Takmaz, Elisabetta Fedele, Robert W Sumner, Marc Pollefeys, Federico Tombari, and Francis Engelmann. Openmask3d: Open-vocabulary 3d instance segmentation. arXiv preprint arXiv:2306.13631, 2023.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 5294–5306, 2025a.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A. Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In CVPR, 2025b.

Tai Wang, Xiaohan Mao, Chenming Zhu, Runsen Xu, Ruiyuan Lyu, Peisen Li, Xiao Chen, Wenwei Zhang, Kai Chen, Tianfan Xue, et al. Embodiedscan: A holistic multi-modal 3d perception suite towards embodied ai. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 19757–19767, 2024.

Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. pi3: Permutation-equivariant visual geometry learning. arXiv preprint arXiv:2507.13347, 2025c.

Chong Xia, Kai Zhu, Zizhuo Wang, Fangfu Liu, Zhizheng Zhang, and Yueqi Duan. Simrecon: Simready compositional scene reconstruction from real videos. arXiv preprint arXiv:2603.02133, 2026.

Xiuwei Xu, Huangxing Chen, Linqing Zhao, Ziwei Wang, Jie Zhou, and Jiwen Lu. Embodiedsam: Online segment any 3d thing in real time. arXiv preprint arXiv:2408.11811, 2024.

Yunhan Yang, Xiaoyang Wu, Tong He, Hengshuang Zhao, and Xihui Liu. Sam3d: Segment anything in 3d scenes. arXiv preprint arXiv:2306.03908, 2023.

Mingqiao Ye, Martin Danelljan, Fisher Yu, and Lei Ke. Gaussian grouping: Segment and edit anything in 3d scenes. In ECCV, 2024.

Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In Proceedings of the International Conference on Computer Vision (ICCV), 2023.

Yingda Yin, Yuzheng Liu, Yang Xiao, Daniel Cohen-Or, Jingwei Huang, and Baoquan Chen. Sai3d: Segment any instance in 3d scenes. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3292–3302, 2024.

Junyi Zhang, Charles Herrmann, Junhwa Hur, Chen Sun, Ming-Hsuan Yang, Forrester Cole, Trevor Darrell, and Deqing Sun. Loger: Long-context geometric reconstruction with hybrid memory. arXiv preprint arXiv:2603.03269, 2026.

Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In European conference on computer vision, pp. 350–368. Springer, 2022.

Dong Zhuo, Wenzhao Zheng, Jiahe Guo, Yuqi Wu, Jie Zhou, and Jiwen Lu. Streaming 4d visual geometry transformer. arXiv preprint arXiv:2507.11539, 2025.

Lojze Zust, Yohann Cabon, Juliette Marrie, Leonid Antsfeld, Boris Chidlovskii, Jerome Revaud, and Gabriela Csurka. Panst3r: Multi-view consistent panoptic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 5856–5866, 2025.

A APPENDIX

A.1 BACKGROUND

In this section, we briefly review the geometric components used in our method: 2D–3D correspondence, weighted similarity-transform estimation for aligning local point maps, and sparse feature tracking for constructing reliable multi-view constraints.

2D-3D correspondences. For a pixel $\mathbf{u} = (x, y)$ with homogeneous coordinate $\tilde{\mathbf{u}} = [x, y, 1]^\top$, camera intrinsics $\boldsymbol{\kappa} \in \mathbb{R}^{3\times3}$, and camera extrinsics $(\mathbf{R}, \mathbf{t})$, we define the depth-conditioned back-projection from 2D to 3D as:

$$\mathbf{X} = \pi^{-1}(\mathbf{u}, d; \boldsymbol{\kappa}, \mathbf{R}, \mathbf{t}) = \mathbf{R}\big(d\,\boldsymbol{\kappa}^{-1}\tilde{\mathbf{u}}\big) + \mathbf{t}, \tag{10}$$

where $d \in \mathbb{R}_+$ denotes the depth of pixel $\mathbf{u}$, and $\mathbf{X} \in \mathbb{R}^3$ is the corresponding 3D point in the local frame. Conversely, given a 3D point $\mathbf{X} \in \mathbb{R}^3$ in the local frame, its projection onto the image plane is defined as:

$$\mathbf{u} = \pi(\mathbf{X}; \boldsymbol{\kappa}, \mathbf{R}, \mathbf{t}) = \left(\frac{\tilde{u}_x}{\tilde{u}_z}, \frac{\tilde{u}_y}{\tilde{u}_z}\right), \quad \begin{bmatrix}\tilde{u}_x\\ \tilde{u}_y\\ \tilde{u}_z\end{bmatrix} = \boldsymbol{\kappa}\,\mathbf{R}^\top(\mathbf{X} - \mathbf{t}) \tag{11}$$

where $\pi : \mathbb{R}^3 \to \mathbb{R}^2$ denotes the projection operator, and $\tilde{\mathbf{u}} = (\tilde{u}_x, \tilde{u}_y, \tilde{u}_z)^\top$ is the intermediate homogeneous image coordinate before perspective division. Note that π^{-1} is not a true inverse of π, but a depth-conditioned back-projection operator.

Weighted similarity-transform estimation. To merge local reconstructions into a globally consistent structure, we need to estimate a robust weighted similarity transform that aligns overlapping 3D point maps while accounting for scale, rotation, translation, and noisy correspondences. Following (Deng et al., 2025), we estimate a robust similarity transform $\mathbf{T} \in \mathrm{Sim}(3)$ that aligns overlapping point maps while accounting for scale, rotation, translation, and noisy correspondences. Given paired 3D points $\{\mathbf{X}_i^1, \mathbf{X}_i^2\}$, we solve:

$$\mathbf{T}_{1,2}^* = \arg\min_{\mathbf{s},\mathbf{R},\mathbf{t}} \sum_i c_i\, \rho\left(\left\|\mathbf{X}_i^1 - \big(\mathbf{s}\mathbf{R}\mathbf{X}_i^2 + \mathbf{t}\big)\right\|_2\right), \tag{12}$$

where $\rho(\cdot)$ is the Huber loss and c_i is a confidence weight for correspondence i. We solve this objective using Iteratively Reweighted Least Squares (IRLS) (Daubechies et al., 2010) optimization and use the estimated transform to map the second point map into the coordinate frame of the first, suppressing noisy correspondences while preserving reliable overlap.

Sparse Feature Tracker. To establish reliable geometric constraints across views, we need to construct sparse multi-view feature tracks from the input image sequence. While dense predictions provide rich per-pixel geometry, sparse tracks offer highly confident and repeatable correspondences that are crucial for camera pose estimation, scale alignment, and subsequent bundle-adjustment refinement. Given a sequence of N images, we first extract sparse keypoints and descriptors for each frame using SuperPoint (DeTone et al., 2018). Following (Cheng et al., 2026), we build a correspondence k-NN graph using DINO (Siméoni et al., 2025) features to retrieve a set of visually similar neighboring frames for each image i,. For each visually similar image pair (i, γ), we compute tentative correspondence matches using LightGlue (Lindenberger et al., 2023) as $\mathcal{M}_{i,\gamma} = \{(u_{i,j}, u_{\gamma,k})\}$, where $u_{i,j}$ and $u_{\gamma,k}$ denote the j-th and k-th keypoint observations in images i and γ, respectively. Then, we merge pairwise correspondences across all image pairs into multi-view tracks via a disjoint-set (union–find) structure. Each track forms a connected component $\mathcal{T}_{\text{sparse}}$, and all observations within the same component are treated as projections of a common 3D point. The resulting sparse tracks and associated 3D points provide a compact multi-view representation for subsequent geometric processing.

A.2 HYPERPARAMETER ANALYSIS

Study on the sensitivity of depth. We first study the depth-consistency threshold τ_{depth}, which filters geometrically valid correspondences in the proxy geometry. As shown in Table 5, $\tau_{\text{depth}} =$

(a) τ_{depth}

	0.05	0.1	0.2	0.3
AP	17.2	**18.1**	16.9	16.5
AP_{50}	30.1	**33.8**	29.9	27.1
AP_{25}	42.3	**47.2**	43.2	39.5

(b) $\tau_{\text{3D-IoU}}$

	0.3	0.4	0.5	0.6
AP	17.8	**18.1**	17.9	17.5
AP_{50}	31.1	**33.8**	32.1	29.1
AP_{25}	45.9	**47.2**	46.9	42.4

(c) $\tau_{\text{2D-IoU}}$

	0.8	0.85	0.9	0.95
AP	17.4	17.7	17.6	**18.1**
AP_{50}	30.9	32.3	31.2	**33.8**
AP_{25}	44.1	46.8	45.9	**47.2**

(d) λ

	2	5	10	15
AP	17.8	**18.1**	18.0	17.5
AP_{50}	31.1	**33.8**	32.9	30.1
AP_{25}	44.4	**47.2**	45.4	41.6

Table 5: **Hyperparameter analysis on mask generation thresholds.**

0.1 achieves the best performance, with 18.1 AP, 33.8 AP_{50}, and 47.2 AP_{25}. A smaller threshold is more conservative and may discard useful correspondences, while larger thresholds admit noisier cross-view matches, leading to degraded mask association.

Study on 3D IoU threshold. We then ablate the 3D IoU threshold $\tau_{\text{3D-IoU}}$ used in iterative graph matching. The method is relatively stable across different thresholds, but $\tau_{\text{3D-IoU}} = 0.4$ gives the best overall results. Lower thresholds may merge unrelated object groups, while higher thresholds make matching overly strict and leave more duplicate or fragmented identities.

Study on Video Refinement. Finally, we analyze the video refinement thresholds $\tau_{\text{2D-IoU}}$ and λ, which control track merging based on repeated 2D mask overlap. For $\tau_{\text{2D-IoU}}$, the best result is obtained at 0.95, suggesting that duplicate-track merging benefits from a strict overlap criterion. For λ, performance peaks at 5 shared frames, while larger values become too conservative and reduce the ability to merge fragmented tracks. These results show that our default refinement setting effectively removes duplicates while preserving distinct object identities.

A.3 Additional Analysis

We provide **additional qualitative demo videos** in our supplementary materials!

Compared with ScanNet200, ScanNet++v2 provides consistently higher-fidelity scene observations and geometry. Its RGB frames are typically sharper and less affected by motion blur, while its reconstructed meshes are significantly denser and more complete, resulting in cleaner rasterized 2D labels and annotated 3D geometry. By contrast, ScanNet200 often contains blurrier views and sparser reconstructions, which introduce greater ambiguity for both geometry estimation and object-level reasoning. As a result, ScanNet200 serves as a more challenging benchmark, whereas ScanNet++v2 offers higher-quality supervision and a stronger setting for evaluating methods under accurate geometric reconstruction. Accordingly, we adopt ScanNet++v2 for panoptic lifting and novel view rendering, as its superior reconstruction quality yields more faithful object lifting and more stable rendering results.

We organize the additional analyses as follows: Sec. A.3.1 evaluates SEGA for open-vocabulary 3D instance segmentation and studies the benefit of additional input views; Sec. A.3.2 isolates the effect of reconstruction noise on segmentation; Sec. A.3.3 details the class-agnostic 3D instance segmentation protocol, baselines, and additional results; Sec. A.3.4 presents the evaluation setup and qualitative analysis for panoptic lifting; and Sec. A.3.5 reports additional camera-pose and point-cloud reconstruction results.

A.3.1 Open-Vocabulary 3D Instance Segmentation

Although SEGA primarily produces class-agnostic 3D instances, its globally consistent instance masks can be combined with open-vocabulary semantic recognition. We evaluate this capability on ScanNet++v2 using a vocabulary of 100 object classes and following the OpenMask3D (Takmaz et al., 2023) multi-level crop evaluation protocol. We report average precision (**AP**) and average precision at IoU thresholds of 0.50 (**AP50**) and 0.25 (**AP25**).

We further use this setting to study whether processing long image sequences provides meaningful benefits over uniformly sampling a smaller subset of frames. We uniformly sample 50, 300, 800, or 1,600 frames from each sequence while keeping the remaining pipeline unchanged. Indoor objects are often heavily occluded, visible only briefly, or observed clearly from only a small subset of

# Frames per Scene	AP↑	AP50↑	AP25↑
50	5.3	8.9	10.1
300	10.1	18.2	21.5
800	13.5	22.1	25.6
1,600	**14.9**	**24.9**	**27.2**

Table 6: **Open-vocabulary 3D instance segmentation on ScanNet++v2.** We evaluate 100 object classes following the OpenMask3D protocol while varying the number of uniformly sampled input frames. Performance consistently improves as more views are incorporated.

Global Proxy Geometry Used for Propagation	AP↑	AP50↑	AP25↑	AR↑	RC50↑	RC25↑
Predicted Depth + Predicted Cameras	28.4	43.9	49.3	54.2	82.7	92.1
Oracle Geometry (GT Depth + GT Cameras)	**32.5**	**49.1**	**58.2**	**65.6**	**91.8**	**95.2**

Table 7: **Class-agnostic 3D instance segmentation results on ScanNet++v2 using oracle evaluation geometry.** We compare the predicted global proxy geometry with oracle geometry for cross-view instance propagation. Oracle geometry uses ground-truth depth maps and camera poses. Higher values are better for all metrics.

viewpoints. Uniformly reducing the input may therefore discard observations that are important for recovering complete instances and assigning their semantic labels.

As shown in Table 6, performance improves consistently with the number of input views. Increasing the sequence length from 50 to 1,600 frames improves AP from 5.3 to 14.9, AP50 from 8.9 to 24.9, and AP25 from 10.1 to 27.2. More importantly, compared with the 300-frame setting, processing 1,600 frames improves AP, AP50, and AP25 by 4.8, 6.7, and 5.7 points, respectively. Performance continues to increase from 800 to 1,600 frames, indicating that the additional observations remain informative rather than merely redundant. These results demonstrate that SEGA improves not only geometric coverage and instance consistency but also downstream open-vocabulary 3D instance recognition.

A.3.2 Disentangling reconstruction noise from 3D segmentation errors

To isolate how reconstruction noise affects 3D instance segmentation, we conduct a modified-oracle experiment at the global proxy geometry construction stage. In both variants, we keep the initial cluster-wise 2D mask predictions unchanged. In the default setting, we construct the global proxy geometry using SEGA's predicted depth maps and camera parameters (our paper results). In the modified-oracle setting, we instead construct it using ground-truth depth maps and ground-truth cameras.

The final video segmentation masks are not held fixed. For each geometry setting, we rerun all subsequent geometry-dependent operations, including the construction of dense 2D–3D correspondences, cross-cluster graph matching, object-aware frame sampling, 3D-aware positive and negative prompting, mask propagation, and video refinement. Therefore, the final ID-consistent video segmentation is regenerated from either the predicted or oracle global proxy geometry. The same algorithms, initial cluster-wise masks, and hyperparameters are used in both variants.

Both variants are then evaluated under the same GT-geometry 3D instance segmentation protocol by lifting their respective video segmentation results onto the ground-truth point cloud. Consequently, the performance difference specifically measures how reconstruction noise in the global proxy geometry affects cross-view association and geometry-guided video-mask propagation in Tab. 7

A.3.3 Class-Agnostic 3D Instance Segmentation evaluation

Existing protocols (Li et al., 2025; Koch et al., 2025) typically align the predicted geometry to the ground-truth using the first ground-truth pose, followed by scale normalization via bounding-box fitting. Subsequently, each predicted point is associated with a ground-truth point using k-nearest neighbors (k-NN). However, we observe that this evaluation strategy can severely distort geometric consistency, as bounding-box-based scale alignment is unreliable in complex indoor scenes. More-

Method	GT Geometry	ScanNet200						ScanNet++v2					
		AP↑	**AP**$_{50}$↑	**AP**$_{25}$↑	**AR**↑	**RC**$_{50}$↑	**RC**$_{25}$↑	**AP**↑	**AP**$_{50}$↑	**AP**$_{25}$↑	**AR**↑	**RC**$_{50}$↑	**RC**$_{25}$↑
1. PanSt3R	×	2.1	4.2	11.1	5.6	10.3	16.1	1.5	2.9	5.7	5.2	8.9	10.6
2. PanSt3R$^{\#}$	×	12.1	22.3	38.2	<u>22.9</u>	<u>39.5</u>	<u>58.3</u>	11.7	21.9	31.0	20.9	33.5	48.9
3. PanSt3R†	×	<u>13.6</u>	<u>24.9</u>	<u>40.6</u>	20.9	36.3	56.5	<u>15.0</u>	<u>23.7</u>	<u>33.0</u>	20.8	30.2	38.2
4. IGGT†	×	6.9	11.2	20.9	18.2	29.9	39.6	7.2	19.1	29.2	21.6	<u>36.7</u>	<u>50.9</u>
5. SEGA	×	**17.1**	**32.7**	**48.4**	**30.8**	**56.6**	**79.4**	**18.1**	**33.8**	**47.2**	**32.3**	**57.8**	**76.7**
6. PanSt3R	✓	9.9	28.4	48.1	20.1	45.2	69.9	8.2	25.3	40.1	24.7	49.9	72.1
7. PanSt3R$^{\#}$	✓	10.4	26.0	46.8	19.6	47.7	81.1	7.4	17.5	33.1	14.2	32.7	58.6
8. PanSt3R†	✓	<u>29.2</u>	<u>45.1</u>	<u>49.9</u>	<u>56.7</u>	<u>84.2</u>	<u>92.9</u>	<u>22.2</u>	<u>40.1</u>	<u>47.9</u>	<u>42.8</u>	<u>77.1</u>	**92.2**
9. IGGT†	✓	12.1	32.9	45.1	25.5	51.6	89.9	9.2	20.8	25.6	25.1	46.9	70.2
10. SEGA	✓	**31.5**	**48.2**	**55.5**	**58.6**	**85.2**	**93.1**	**28.4**	**43.9**	**49.3**	**54.2**	**82.7**	<u>92.1</u>

Table 8: **Additional 3D class-agnostic instance segmentation results** on ScanNet200 and ScanNet++v2. For concatenated multi-run baselines (Zust et al., 2025; Li et al., 2025), # denotes integration with SAM3D (Yang et al., 2023), while † denotes integration with Open3DIS (Nguyen et al., 2024) for 3D grouping. In contrast, SEGA *directly lifts predictions* to the 3D point cloud without any additional merging strategy. The **best** and <u>second-best</u> results are highlighted in **bold** and <u>underline</u>, respectively.

over, the k-NN matching procedure breaks the underlying 2D–3D correspondence, leading to ambiguous and potentially biased evaluation.

To overcome these limitations, we employ a more robust *pixel-aligned evaluation protocol* for *GT geometry-free evaluation* (×). Concretely, we unproject the interpolated ground-truth depth maps together with rasterized 2D segmentation annotations derived from the ground-truth annotated meshes, yielding pixel-aligned annotated 3D point maps. These ground-truth point maps establish one-to-one correspondence at the pixel-level. We then lift the 2D segmentation predictions to 3D via per-pixel association and aggregate them into the final 3D point cloud. This design eliminates the need for nearest-neighbor matching and preserves the point cloud resolution, keeping geometric consistency throughout evaluation. Our protocol provides a more faithful and robust assessment of 3D segmentation quality compared to prior evaluation schemes based on k-NN. Additionally, in the *oracle setup* for *GT geometry-based evaluation* (✓), where ground-truth 3D geometry, i.e., ground-truth point cloud, depth, camera poses provided. In this setup, groundtruth mesh is provided so we can use graph-based segmentation (Felzenszwalb & Huttenlocher, 2004) to obtain the privilege 3D prior – superpoints. For this setup, the evaluation follows the standard protocol of 3D point cloud instance segmentation (Schult et al., 2023; Yeshwanth et al., 2023; Rozenberszki et al., 2022) for lifting 2D segmentation masks to 3D point cloud.

Additional Baselines Setup. We also include additional variants of PanSt3R (Zust et al., 2025) and IGGT (Li et al., 2025) equipped with the grouping strategies of Open3DIS (†) and SAM3D (#), as reported in Table 8. Specifically, PanSt3R and IGGT are executed multiple times over different chunks, and the resulting predictions are concatenated across runs. The corresponding 2D masks are then merged using reprojected point maps, yielding merged pixel-aligned 3D segmentation masks for evaluation.

Quantitative results. As shown in Table 8, under the pixel-aligned evaluation setting, equipping PanSt3R and IGGT with additional merging strategies: Open3DIS (†) and SAM3D (#), substantially improves their performance. These auxiliary grouping modules help associate cluster-wise 2D segmentation masks in 3D and partially mitigate the fragmentation caused by processing long sequences in separate clusters. However, their performance remains fundamentally constrained by the quality and coverage of the underlying cluster-wise geometric foundation. In particular, when the appearance gap across views becomes large, these pure post-hoc merging strategies often fail to recover complete object instances across the entire scene. This limitation is reflected in the large gap between precision and recall. While the adapted baselines can achieve moderate precision, their recall remains significantly lower, indicating that many object instances are only partially recovered or entirely missed. Across both ScanNet200 and ScanNet++v2, this gap is often on the order of 20–30 % in AP and nearly 50 % in recall-oriented metrics, highlighting the incomplete object coverage of these approaches. For IGGT combined with Open3DIS-style grouping, its results are consistently worse than the corresponding PanSt3R variants. We attribute this to the heuristic PCA-based

grouping principle used in the dense prediction transformer head, which makes nearby instances less distinguishable and degrades the quality of subsequent 3D grouping. In contrast, SEGA directly lifts predictions to the 3D point cloud without relying on any additional merging strategy, yet still achieves the strongest overall performance, demonstrating the advantage of producing globally consistent and temporally coherent segmentations.

Under the ground-truth geometry setting, equipping PanSt3R and IGGT with IoU-based merging leads to substantial performance gains, slightly surpassing the original Open3DIS variant with SAM3 masks (Table 1 line 9). Despite not relying on any additional merging strategy, SEGA with ground-truth geometry still achieves the strongest overall results. The improvement is especially pronounced on ScanNet++v2, while the gain on ScanNet200 is smaller but remains consistent. One possible explanation is the long-tailed distribution of 200 categories in ScanNet200 and the small-sized scenes, which make further improvements inherently harder to realize.

Qualitative results. Fig. 6 presents representative class-agnostic 3D instance segmentation results on ScanNet++v2 and ScanNet200. Overall, the visualizations show that our method is able to recover spatially coherent object-level decompositions across a wide range of indoor scenes, including bedrooms, offices, kitchens, and cluttered living spaces. Distinct object instances are assigned consistent colors, indicating that the method can separate nearby objects with similar geometry while preserving clean object boundaries in 3D.

The results on ScanNet++v2 demonstrate strong performance on densely reconstructed scenes with fine-grained geometry. Our method is able to capture both large structural elements, such as walls, floors, and furniture, and smaller foreground objects, yielding a detailed decomposition of the scene. On ScanNet200, which contains more diverse object categories and often more challenging clutter, the predictions remain stable and capture the dominant object layouts effectively. Despite the increased complexity of these scenes, the method still produces compact and interpretable object partitions.

We also observe that the predicted instances are reasonably complete in 3D, rather than being limited to sparse or view-dependent fragments. This suggests that the proposed lifting and association pipeline can aggregate information across views to form coherent object representations in the reconstructed scene. Taken together, these qualitative examples highlight the robustness of our approach across different scene types and dataset characteristics.

Additionally, Fig. 7 presents a qualitative comparison of SEGA, Open3DIS, and SAM3D in the setting with ground-truth geometry. SEGA yields substantially cleaner and more coherent 3D instance segmentation results, whereas Open3DIS often produces fragmented object masks due to its reliance on independent frame-wise segmentations. SAM3D fails more severely in occluded regions, where the absence of superpoint guidance prevents reliable separation of object instances.

A.3.4 Panoptic Lifting evaluation

Gaussian Grouping (Ye et al., 2024) is leveraged as the lifting backbone, which augments 3D Gaussians with learnable identity encodings and supervises them by differentiably rendering the identity features to match 2D instance labels. This produces a 3D Gaussian representation that can render both RGB images and instance maps from novel viewpoints. We use the video object masks predicted by each method as the 2D identity supervision for Gaussian Grouping. We then evaluate the rendered instance maps on 50 held-out views against rasterized ground-truth 2D annotations on 10 test scenes in the ScanNet++v2 dataset.

We compare 3 different instance segmentation methods on the same number of images. First, *PanSt3R + Gaussian Grouping* uses cluster-wise PanSt3R predictions. Since PanSt3R is applied independently to each frame cluster, its instance IDs are only locally consistent. We therefore concatenate the cluster-level instance IDs across the video without cross-cluster identity alignment. Second, *SAM3 + Gaussian Grouping* uses masks obtained by running the video segmentation model SAM3 (Carion et al., 2025) on the full sequence. Finally, *SEGA + Gaussian Grouping* uses our globally ID-consistent video object masks, produced by 3D-aware alignment, mask propagation, and video refinement.

Method	ScanNet200			ScanNet++v2		
	RRA@30↑	RTA@30↑	AUC@30↑	RRA@30↑	RTA@30↑	AUC@30↑
TTT3R (Chen et al., 2025)	99.89	96.20	75.21	98.05	96.40	78.98
FastVGGT (Shen et al., 2025)	99.92	**96.54**	76.28	98.14	96.56	79.56
LingBot-Map (Chen et al., 2026)	**99.96**	96.27	76.23	98.43	**97.78**	**83.23**
VGGT-Long (Deng et al., 2025)	99.52	96.09	76.66	96.88	95.27	77.19
MERG3R (Cheng et al., 2026)	99.90	96.23	76.79	97.25	96.45	78.45
SEGA (w/o IA-BA)	99.92	96.23	<u>77.90</u>	**99.12**	97.24	81.31
SEGA (w IA-BA)	<u>99.93</u>	<u>96.46</u>	**78.91**	<u>99.10</u>	<u>97.49</u>	<u>82.42</u>

Table 9: **Camera pose estimation results** on ScanNet200 and ScanNet++v2. Shaded rows indicate streaming reconstruction methods that process the sequence online without chunk-based optimization. RRA@30 and RTA@30 report the percentage of frame pairs with rotation and translation errors below 30°, respectively, and AUC@30 summarizes performance over thresholds up to 30°; higher is better. SEGA has two variants with (**w IA-BA**) and without (**w/o IA-BA**) Instance-Aware Bundle Adjustment. The **best** and <u>second best</u> results are marked as **bold** and <u>underline</u>.

Method	Geometry Backbone	Acc.↓	Comp.↓	Chamfer Dist.↓	F-Score↑
FastVGGT (Shen et al., 2025)	VGGT	0.0325	0.0164	0.0490	0.8852
LingBot-Map (Chen et al., 2026)	VGGT	0.0315	**0.0141**	0.0457	0.8942
VGGT-Long (Deng et al., 2025)	VGGT	0.1653	0.0367	0.2021	0.3494
π^3-Long (Wang et al., 2025c)	π^3	0.1018	0.0268	0.1286	0.5187
SEGA (w/o IA-BA)	MUSt3R	<u>0.0255</u>	0.0155	<u>0.0410</u>	<u>0.9383</u>
SEGA (w IA-BA)	MUSt3R	**0.0250**	<u>0.0153</u>	**0.0403**	**0.9401**

Table 10: **3D point cloud reconstruction results on ScanNet++v2**, supplementing Table 9. Shaded rows indicate streaming reconstruction methods that process the sequence online without chunk-based optimization. Accuracy (Acc.), completeness (Comp.), and Chamfer distance measure geometric reconstruction errors, for which lower values are better. F-Score measures the overall reconstruction quality, for which higher values are better. The **best** and <u>second-best</u> results are marked in **bold** and <u>underline</u>, respectively.

Qualitative Results. We present qualitative comparisons in Fig. 8 by visualizing rendered test views from different methods. Our method produces more coherent and geometrically consistent 3D panoptic segmentations across views. In contrast, concatenated PanSt3R exhibits substantial noise, as instance identities are not shared across clusters, leading to fragmented 3D panoptic predictions. Compared with SAM3, while the propagated masks are often visually sharp, objects are frequently missed when propagation is applied over long sequences with many views (e.g., row 5 of Fig. 8). We attribute this limitation to the memory constraints of SAM3 in large-scale scenes. Relative to the rendered ground-truth panoptics, our method also produces a denser segmentation, yielding a qualitatively more complete decomposition of the scene.

A.3.5 3D Reconstruction evaluation

We evaluate SEGA's 3D reconstruction capability using both camera-pose accuracy and point-cloud quality. For camera poses, we report **RRA@30** and **RTA@30**, which measure the percentages of frame pairs whose relative rotation and translation-direction errors are below 30°, respectively. We additionally report **AUC@30**, computed by integrating the minimum of RRA and RTA over thresholds from 1° to 30°. To ensure a fair comparison, we compare SEGA only with state-of-the-art methods capable of processing large numbers of input views.

After reconstructing the complete sequence, we anchor the predicted trajectory to the coordinate frame of the first ground-truth camera. For every frame pair with available ground-truth poses, we compute the predicted and ground-truth relative transformations. Rotation error is measured as the geodesic distance on $\mathrm{SO}(3)$, while translation error is measured as the angular difference between the corresponding relative translation directions.

We evaluate point-cloud reconstruction using accuracy (**Acc.**), completeness (**Comp.**), Chamfer distance (**CD**), and **F-Score**. Accuracy measures the average distance from each predicted point to its nearest ground-truth point, while completeness measures the average distance from each ground-

truth point to its nearest predicted point. Chamfer distance is the sum of accuracy and completeness, and F-Score is the harmonic mean of reconstruction precision and recall at a fixed distance threshold. Lower Acc., Comp., and CD indicate better geometric reconstruction, whereas a higher F-Score is better.

We construct the point cloud by unprojecting its predicted depth maps using its predicted camera intrinsics and extrinsics, and evaluate all outputs using the same alignment and sampling protocol.

Quantitative results. As shown in Table 9, SEGA achieves competitive camera pose accuracy despite using MUSt3R (Cabon et al., 2025) rather than the stronger VGGT backbone used by most recent baselines. On ScanNet200, SEGA with IA-BA achieves the best AUC@30 of 78.91, outperforming the strongest cluster-based baseline, MERG3R, by 2.12 points. On ScanNet++v2, it achieves an AUC@30 of 82.42, ranking second overall and surpassing MERG3R by 3.97 points. IA-BA consistently improves RTA@30 from 96.23 to 96.46 and AUC@30 from 77.90 to 78.91 on ScanNet200, and improves RTA@30 from 97.24 to 97.49 and AUC@30 from 81.31 to 82.42 on ScanNet++v2. These results demonstrate that instance-consistent correspondences provide effective constraints for refining camera poses beyond the initial geometry alignment.

The improved poses also translate into more accurate 3D geometry. As shown in Table 10, SEGA with IA-BA achieves the best accuracy error (0.0250), Chamfer distance (0.0403), and F-Score (0.9401), together with the second-best completeness error (0.0153), on ScanNet++v2. Compared with SEGA without IA-BA, the refinement reduces the accuracy, completeness, and Chamfer errors from 0.0255, 0.0155, and 0.0410 to 0.0250, 0.0153, and 0.0403, respectively, while increasing the F-Score from 0.9383 to 0.9401. Compared with LingBot-Map, SEGA reduces the accuracy error by 20.6% and the Chamfer distance by 11.8%, while improving the F-Score by 4.59 percentage points. Together, these results confirm that IA-BA improves both camera pose estimation and the resulting 3D reconstruction.

Qualitative results. We showcase the performance of SEGA across the two datasets in Fig. 5. These instance-aware constraints allow our post-processing step IA-BA to correct residual pose errors and local geometric inconsistencies introduced by the preceding reconstruction and alignment stages. In turn, this reduces blur, improves structural sharpness, and produces more coherent object geometry. The effect is especially pronounced around thin structures, object boundaries, and partially occluded regions, where sparse constraints are typically too weak to prevent drift. As highlighted in the figure, SEGA leads to visibly clean reconstructions and accurate alignment, demonstrating the importance of identity-consistent correspondence for high-quality 3D refinement.

A.4 Dataset details

We have chosen the two ScanNet200 and ScanNet++v2 indoor datasets, because they provide high-quality geometry, dense annotations, detailed scans, and mesh-based segmentation labels over diverse and cluttered scenes with complex object layouts. These properties enable reliable evaluation of reconstruction quality, object-level understanding, novel-view synthesis, and cross-view consistency. The visualization of some examples in the two datasets is shown in Fig. 4.

Rasterization. We derive rasterized 2D instance ground truth by rendering the annotated 3D mesh into each camera view using the ground-truth camera intrinsics and poses. Each mesh vertex is assigned a panoptic label that encodes both semantic class and instance identity, and these labels are stored as vertex colors for rendering. After undistortion and resizing, the labeled mesh is rasterized from each viewpoint to obtain a dense panoptic image aligned with the input frame. We further mask out invalid regions introduced by image preprocessing and remove tiny projected components. This procedure yields dense frame-wise 2D instance masks with exact correspondence to the annotated 3D scene geometry.

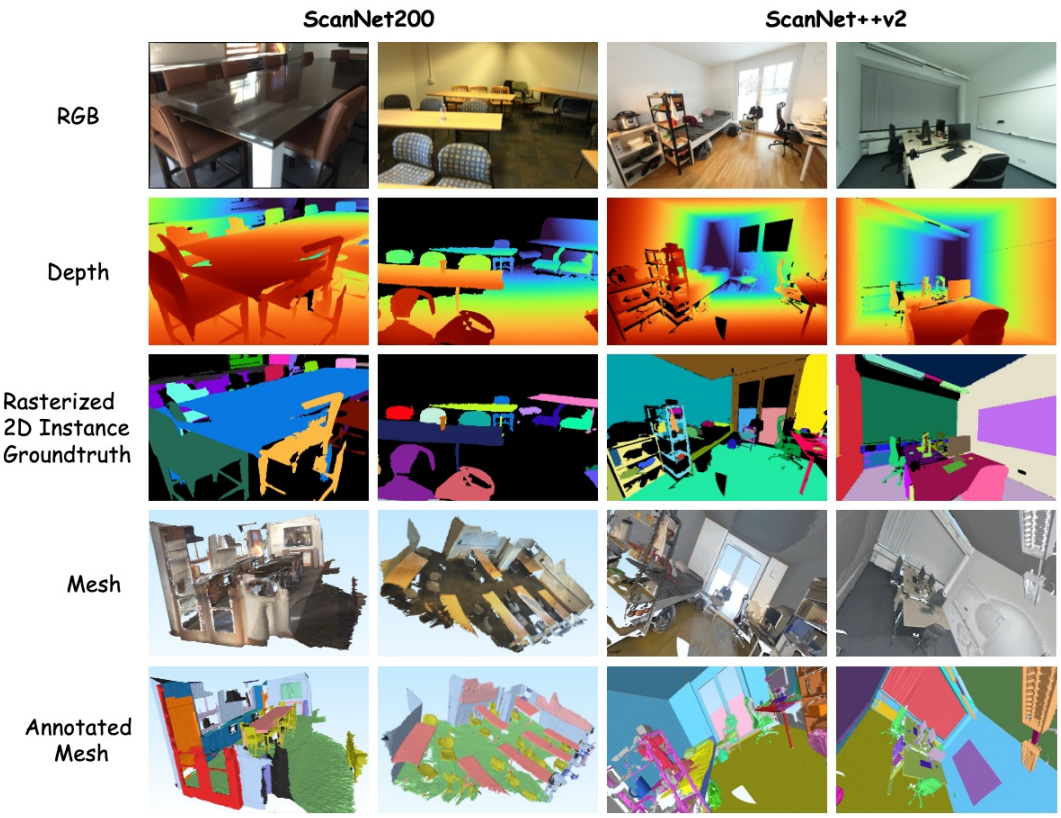

Figure 4: **Datasets Visualization.** Comparison of representative samples from ScanNet200 and ScanNet++v2. For each dataset, we show the input RGB image, the corresponding depth map, rasterized 2D instance ground-truth labels, the reconstructed mesh, and the annotated mesh, highlighting the multi-view indoor scenes, geometric structure, and semantic/instance-level annotations available in both datasets.



Figure 5: **Qualitative examples of 3D Reconstruction**. The top two rows show scenes from ScanNet++v2, which feature denser and more complete geometry. The bottom two rows show scenes from ScanNet200, its reconstructions are generally sparser and less complete.



Figure 6: **Qualitative examples of class-agnostic 3D instance segmentation** rendered under pixel-aligned setting. We visualize instance segmentation results for queried objects on representative scenes from ScanNet++v2 and ScanNet200. Each color denotes a distinct predicted object instance. The top two rows show examples from ScanNet++v2, while the bottom two rows show examples from ScanNet200.



Figure 7: **Qualitative examples of class-agnostic 3D instance segmentation using groundtruth geometry.** Our SEGA produces the cleanest and most coherent 3D instance segmentation. Its object regions are more spatially complete, and the predicted instances are less fragmented. In contrast, both Open3DIS and SAM3D frequently break a single object into multiple disconnected parts or assign inconsistent instance labels to nearby surfaces. For better visualization, we sample 500K of points from the groundtruth mesh.

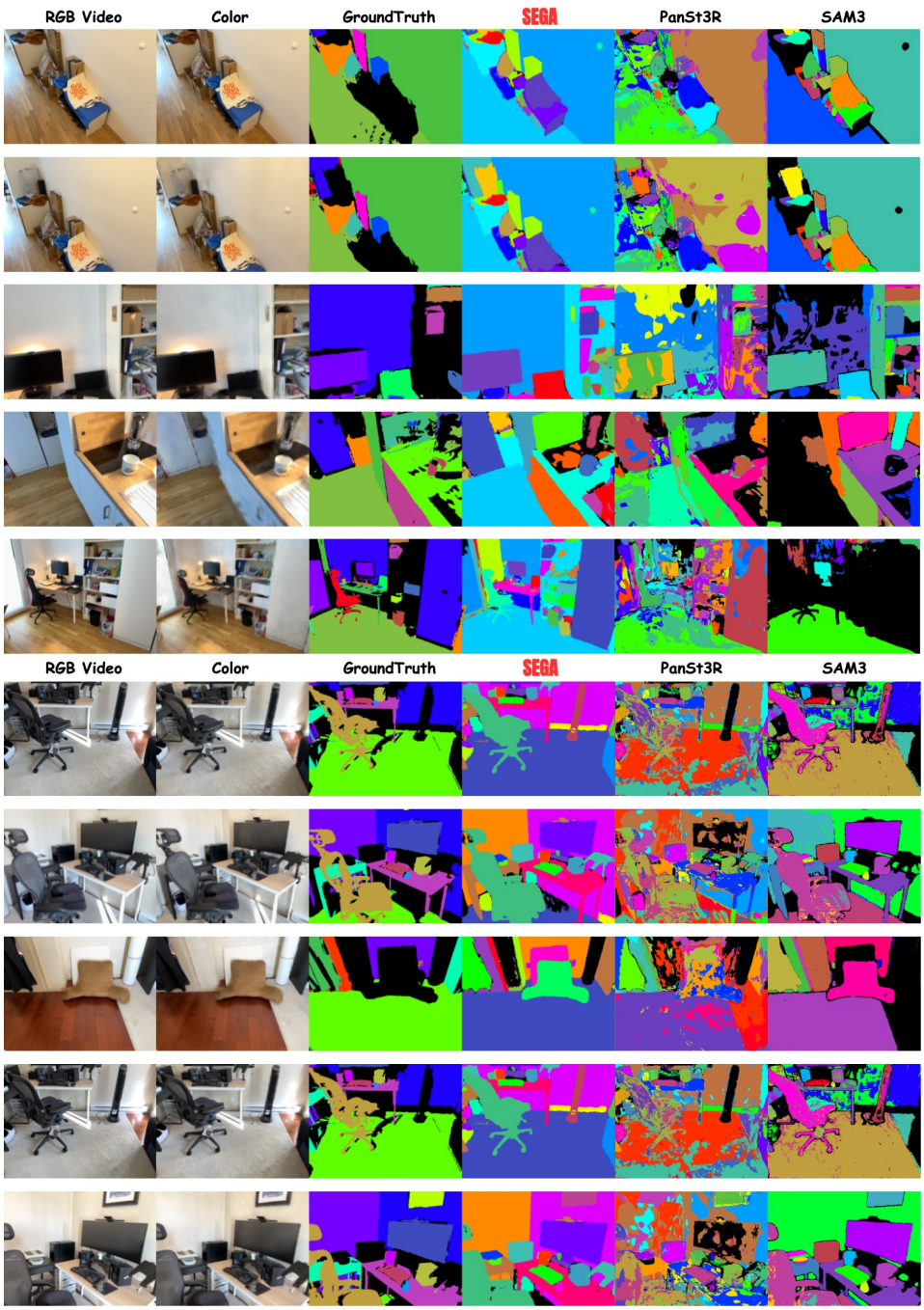

Figure 8: **Qualitative examples of Panoptic Lifting.** We compare the novel test view rendering using Gaussian Grouping on different 2D instance segmentation methods. PanSt3R concatenated the results from multi-run predictions while SAM3 does segmentation and tracking across the video sequence.