PERSISTENT3D: OPEN-VOCABULARY PERSISTENT OBJECT SPATIAL PERCEPTION FROM MONOCULAR VIDEO VIA CROSS-FRAME INSTANCE ASSOCIATION

Anonymous authors
Paper under double-blind review

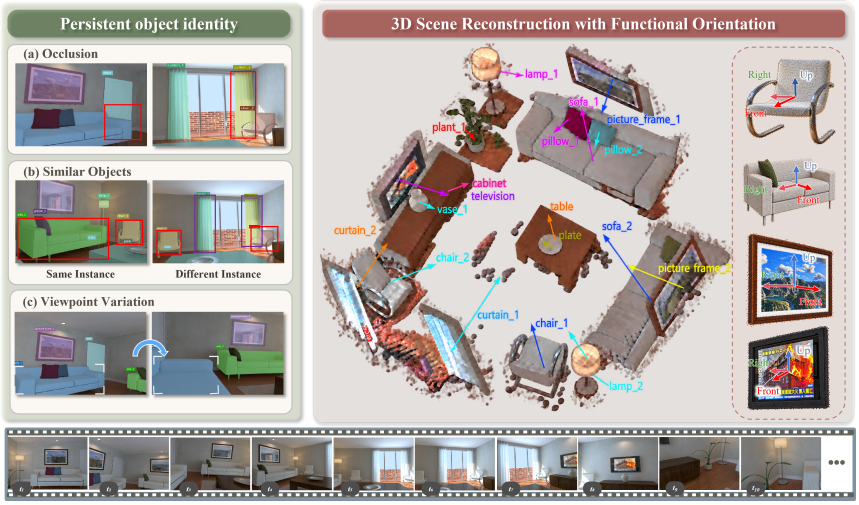


Figure 1: Persistent3D preserves identities and reconstructs functional instance-level 3D geometry.

ABSTRACT

Spatial perception for embodied intelligence requires persistently recognizing and understanding scene instances during dynamic 3D reconstruction. However, cross-frame instance association fails to maintain stable object identities under occlusion, visually similar instances, and substantial viewpoint changes. This prevents consistent aggregation of 3D information across views. To address this issue, we propose Persistent3D, an open-vocabulary framework for persistent object spatial perception from monocular video. Persistent3D first recovers 3D scene geometry and camera poses from video, and combines VLM-based open-vocabulary recognition with instance segmentation to construct per-frame object observations. Starting from frame-level instances, bidirectional temporal propagation and one-to-one mask matching are performed within local temporal windows to alleviate association failures caused by short-term occlusion and viewpoint variation. Meanwhile, a same-frame mutual exclusion constraint is introduced to maintain local cross-frame temporal associations and suppress erroneous merging between visually similar instances. For local identity fragmentation caused by long-term occlusion or viewpoint changes, instance reconnection is performed by leveraging scene-point index overlap, color distributions, and frame-set mutual exclusion constraints across different identity fragments in the 3D scene. During persistent spatial perception, a sequence-wide instance association graph is dynamically constructed to maintain stable persistent identities across frames. Building on this, Persistent3D fuses multi-view vision-language evidence to estimate functional object orientations and fits semantically oriented bounding boxes. The final unified 3D object representation integrates persistent identity, open-vocabulary semantics, 3D geometry, and functional orientation. Extensive experiments demon-

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.

strate consistent and effective improvements in cross-frame instance association, 3D information aggregation, and functional orientation estimation. Project page: https://anonymous.4open.science/r/Persistent3D-7A65/

1 INTRODUCTION

Embodied intelligence requires agents to perceive and understand the surrounding three-dimensional space through continuous interaction with the environment(Savva et al., 2019). Establishing an accurate and persistent understanding of scene structure and the objects within it is therefore fundamental to environmental comprehension and subsequent decision-making(Wang et al., 2024a; Majumdar et al., 2024). In real-world environments, agents are confronted with a wide variety of object categories. With recent advances in open-vocabulary visual perception and multimodal understanding, visual systems have increasingly moved beyond the constraints of predefined category sets, enabling more flexible object recognition and semantic understanding in open-world scenes(Peng et al., 2023; Ding et al., 2023; Liu et al., 2024). Meanwhile, progress in 3D vision has gradually reduced the reliance of scene perception on specialized sensors such as depth cameras(Wang et al., 2025) and LiDAR(Wei et al., 2023), making it increasingly feasible to recover camera motion and three-dimensional scene geometry using monocular RGB video alone(Teed & Deng, 2021; Liu et al., 2025).

Building on these advances, 3D scene representations have gradually evolved from holistic geometric reconstruction toward more fine-grained instance-level representations(Rosinol et al., 2020; Wu et al., 2021). Rather than merely describing scene surfaces and spatial structures, instance-level representations further organize individual objects within a scene, enabling embodied agents to develop a more structured and continuously updatable understanding of the environment(Gu et al., 2024). For objects with explicit functional orientations in continuous videos(Bahl et al., 2023; Li et al., 2025), stable object identities are a prerequisite for aggregating multi-view orientation evidence along the cross-frame trajectory of the same instance. Such representations therefore require not only recognizing objects at each moment, but also maintaining continuous instance identities under agent motion, viewpoint changes, and object occlusions(Wen & Bekris, 2021; Wen et al., 2023). Consequently, maintaining stable object identities over long-term spatial perception has become an important challenge in video-driven 3D scene understanding and persistent spatial perception.

Based on these observations, we propose an open-vocabulary framework for persistent object spatial perception from monocular video, as illustrated in Fig. 1. The framework first addresses persistent object identity maintenance across long video sequences by associating observations across frames and reconnecting fragmented tracks into stable physical instances. Conditioned on these persistent identities, it then aggregates multi-view 3D geometry and further estimates semantically meaningful functional orientations for individual objects.

Our main contributions are summarized as follows:

- We propose Persistent3D, an open-vocabulary framework for persistent object spatial perception from monocular video, which constructs a unified object 3D representation integrating persistent identities, open-vocabulary semantics, 3D geometry, and functional orientations.
- We introduce a persistent identity inference method that combines local cross-frame association with 3D instance-level reconnection, mitigating identity fragmentation caused by occlusion, similar instances, and substantial viewpoint changes. Building on the recovered identities, multi-view vision-language evidence is further fused to recover object functional orientations.
- Extensive experiments on real-world scene data validate the effectiveness of Persistent3D in maintaining long-term identity consistency, integrating 3D information across views, and recovering object functional orientations.

2 Related Work

Persistent Object Identity Association. Temporal propagation can alleviate the degradation of information over time, yet the most challenging cases of identity recovery often arise after temporal continuity has already broken down(Yang et al., 2021; Yang & Yang, 2022; Cheng et al., 2023). Instance representation matching essentially relies on appearance similarity to determine identity, whereas appearance similarity does not necessarily imply that two observations correspond to the same instance(Wu et al., 2022; Huang et al., 2022; Li et al., 2024). Methods that match historical features with current observations(Cheng & Schwing, 2022; Zhou et al., 2024; Lee et al., 2025a) may inherently lack a unique correspondence. In such cases, increasing the memory length merely preserves more historical information, but does not guarantee that the retained information is sufficient to resolve identity ambiguity among instances. More critically, once erroneous observations are written into long-term memory, continued propagation may instead reinforce incorrect identities with stronger temporal consistency. Subsequent studies have therefore gradually moved beyond local association paradigms toward multi-cue joint discrimination. For example, CAVIS(Lee et al., 2025b) exploits contextual information to alleviate matching ambiguity, while FDTA(Shao et al., 2026) improves instance-level discriminability through spatial, temporal, and identity representations. SAM2Long(Ding et al., 2025), in contrast, focuses more explicitly on error control in long-term memory, attributing the major long-term limitation of SAM2(Ravi et al., 2025) to the cascading accumulation of erroneous masks and employing multi-path memory search to prevent greedy errors along a single path from continuously propagating. Nevertheless, these methods still primarily determine whether instances across frames belong to the same trajectory within the 2D observation space. In other words, existing methods mainly optimize frame-to-frame or track-to-detection association. In contrast, object 3D representation requires solving the stronger problem of observation-to-entity partitioning, where all observations belonging to the same instance throughout the video must form a unique and mutually consistent set. 3D spatial perception provides an alternative route to overcoming this limitation(Zhao et al., 2024; Bhalgat et al., 2024). Therefore, long-term identity association should no longer be viewed merely as a video feature matching problem, but rather as an instance association problem jointly constrained by appearance, temporal continuity, and physical space.

Semantic Object Orientation. Beyond stable identity, an object's intrinsic orientation is also an important component of embodied spatial understanding. Traditional 6D pose estimation methods mainly describe object geometry through known models or canonical coordinate spaces, while category-level approaches further leverage shared canonical spaces to align the poses of different instances(Wang et al., 2019; Chen et al., 2020; Lin et al., 2021). However, such representations primarily focus on geometric rotation and do not directly capture the functional directions associated with actual use and interaction. In recent years, related research has begun to extend from geometric pose toward functional and interaction semantics. For example, Orient Anything(Wang et al., 2024c) estimates semantic front directions for open-vocabulary objects, while SoFar(Qi et al., 2026) further represents object functional directions using natural language, enabling interaction-relevant orientations such as insertion directions and handle directions to support spatial reasoning and robotic manipulation. Nevertheless, existing methods are largely based on single-view orientation estimation and are therefore susceptible to occlusion, viewpoint changes, and structural ambiguity, while reliable fusion of multi-view semantic evidence further depends on stable and consistent instance identities. Therefore, Persistent3D builds upon persistent identities to fuse multi-view vision-language evidence within the same instance, and combines camera poses with 3D geometry to recover functionally meaningful orientations that remain consistent across viewpoints.

3 Method

Our proposed Persistent3D follows a progressive pipeline that lifts frame-level 2D observations into instance-level 3D representations, as illustrated in Fig. 2. Given a monocular RGB video, the front-end first employs MASt3R-SLAM(Murai et al., 2025) to establish a unified scene coordinate system and recover camera poses, depth maps, a global point cloud, and point-level provenance. In parallel, Qwen3.6(Qwen Team, 2026) and SAM2.1(Ravi et al., 2025) are used to generate open-vocabulary instance observations for each frame. Building upon these observations, the persistent object identity inference module treats each frame-level instance mask as an independent seed and performs forward and backward propagation within a limited temporal window. One-to-one correspondences

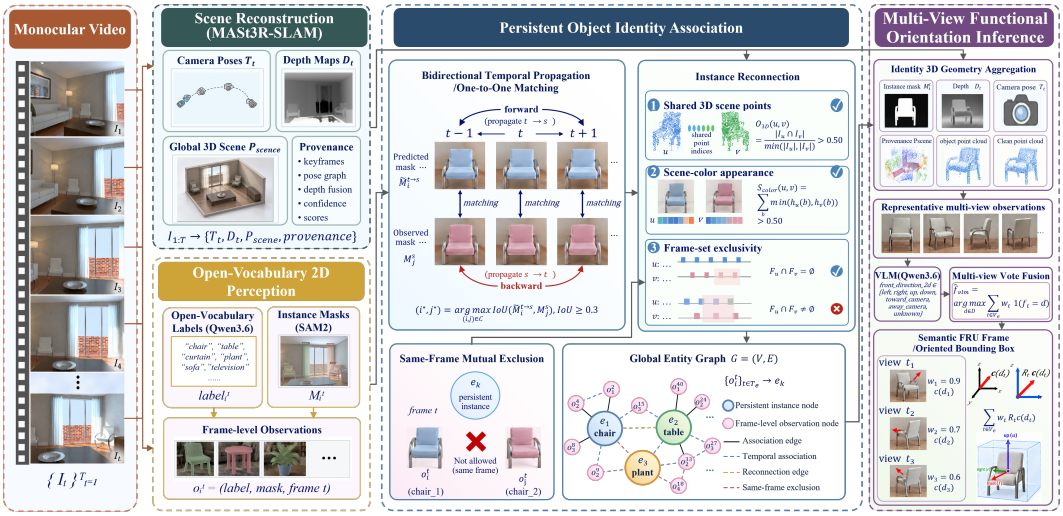


Figure 2: Persistent3D pipeline. Monocular video is converted into scene geometry and open-vocabulary frame-level observations, which are associated into persistent object identities before identity-conditioned 3D aggregation and multi-view functional-orientation inference.

are then established according to mask overlap. All candidate correspondences are subsequently integrated into a unified association structure, where weak associations that would introduce identity conflicts are removed through a same-frame mutual-exclusion rule. For identity fragments that remain disconnected, their semantic categories, shared scene-point indices, scene color distributions, and covered frame sets are further compared, and only fragments satisfying the instance consistency criteria are merged. The subsequent identity-conditioned 3D geometry construction operates directly on the associated instances. Reliable scene points are retrieved from the provenance information, while additional geometric observations are incorporated using instance masks, Z-buffer-based visibility reasoning, and 3D neighborhood expansion. The resulting observations are then refined through point-cloud cleaning to obtain instance-level 3D geometry. In multi-view functional orientation estimation, Qwen3.6(Qwen Team, 2026) determines the semantic front of an object from several representative views, performs confidence-weighted voting across the views, and then transforms the valid directions into the world coordinate system using the corresponding camera poses for fusion. Finally, geometry-constrained semantic coordinate frame recovery takes the fused functional front direction as its primary reference. Category-specific 3D geometric constraints are imposed on objects with explicit structural priors, while PCA within the plane orthogonal to the Front axis, together with a gravity-aligned world-up direction, is used to determine the remaining axes. The semantic OBB is then refitted accordingly, yielding an instance representation that jointly encodes persistent identity, semantic category, 3D geometry, and functional coordinate frame.

3.1 Persistent Object Identity Inference

Let the monocular video be $\{I_t\}_{t=1}^T$. Scene reconstruction provides the camera pose T_t, depth map D_t, global scene point cloud P_{scene}, and point provenance, while open-vocabulary 2D perception produces $o_i^t = (\ell_i^t, M_i^t, t)$. Here, o_i^t is the i-th frame-level observation at time t, ℓ_i^t is its semantic label, and M_i^t is its instance mask.

For temporal association, each M_i^t is propagated both forward and backward within a local window. We denote the mask propagated from frame t to frame s by $\widetilde{M}_i^{t\to s}$. For the remaining candidate pairs $\mathcal{C}_{t\to s}$, one-to-one matching uses the directional overlap score

$$
\begin{aligned}
A_{ij}^{t\to s} &:= \mathrm{IoU}\left(\widetilde{M}_i^{t\to s}, M_j^s\right), \\
(i^*, j^*) &= \underset{(i,j)\in\mathcal{C}_{t\to s}}{\arg\max}\, A_{ij}^{t\to s}, \\
\text{s.t.} \quad & A_{i^*j^*}^{t\to s} \geq \tau_{\text{iou}} = 0.30.
\end{aligned}
\tag{1}
$$

Here, $A_{ij}^{t\to s}$ is the propagated-mask IoU in the figure. After accepting a pair, all candidates containing the same source index i^* or target index j^* are removed. The same operation is performed for $s \to t$, which gives the bidirectional temporal propagation shown in Fig. 2. Thus, $\mathcal{C}_{t\to s}$ is the current candidate set, rather than an additional feature representation.

All accepted temporal correspondences are inserted into a global entity graph

$$\begin{aligned} \mathcal{G} = (\mathcal{V}, \mathcal{E}), \quad \mathcal{V} = \mathcal{V}_{\text{obs}} \cup \mathcal{V}_{\text{id}}, \\ \mathcal{V}_{\text{obs}} := \{o_i^t\}_{i,t}, \quad \mathcal{V}_{\text{id}} := \{e_k\}_{k=1}^{K}, \\ \mathcal{O}(e_k) := \{o_i^t : o_i^t \mapsto e_k\}. \end{aligned} \tag{2}$$

Here, e_k denotes one persistent physical object, and $\mathcal{O}(e_k)$ is the set of all observations assigned to it. Temporal edges and later reconnection edges are positive association edges; same-frame pairs are stored as cannot-link relations. In particular, with $\mathcal{O}_t := \{o_i^t\}_i$, every valid entity must satisfy

$$|\mathcal{O}(e_k) \cap \mathcal{O}_t| \le 1, \qquad \forall e_k,\ \forall t. \tag{3}$$

If a connected component violates Eq. equation 3, its lowest-confidence internal association edge is removed and connectivity is recomputed until no entity contains two observations from the same frame.

Long occlusion can leave two graph fragments corresponding to the same object disconnected. Let u and v denote two such entity fragments, ℓ_u, ℓ_v their semantic labels, $\mathcal{I}_u, \mathcal{I}_v$ their sets of global scene-point indices, $h_u(b), h_v(b)$ the b-th bins of their normalized scene-color histograms, and $\mathcal{F}_u, \mathcal{F}_v$ their covered frame sets. The three reconnection tests in the figure are

$$\begin{aligned} O_{\text{3D}}(u,v) &:= \frac{|\mathcal{I}_u \cap \mathcal{I}_v|}{\min(|\mathcal{I}_u|, |\mathcal{I}_v|)}, \\ S_{\text{color}}(u,v) &:= \sum_{b=1}^{B} \min(h_u(b), h_v(b)), \\ u \sim v \iff & \begin{cases} \ell_u = \ell_v, \\ O_{\text{3D}}(u,v) > 0.50, \\ S_{\text{color}}(u,v) > 0.50, \\ \mathcal{F}_u \cap \mathcal{F}_v = \emptyset. \end{cases} \end{aligned} \tag{4}$$

Thus O_{3D} measures shared reconstructed support, S_{color} measures scene-color appearance agreement, and frame-set exclusivity prevents reconnection from reintroducing a physical identity conflict.

3.2 Identity-Conditioned 3D Geometry

The entity index established above is also the key used for 3D aggregation. For every $o_i^t \in \mathcal{O}(e_k)$, the inputs are exactly the quantities shown in the figure: instance mask M_i^t, depth map D_t, camera pose T_t, global cloud P_{scene}, and its provenance. Let $\bar{M}_i^t$ be the eroded mask and $P_{i,\text{seed}}^t \subseteq P_{\text{scene}}$ the scene points whose provenance pixels lie inside $\bar{M}_i^t$. Each pixel is back-projected to the world frame using its depth in D_t, camera intrinsics K, and pose T_t. We write $\mathcal{Z}_i^t$ for the back-projected points inside $\bar{M}_i^t$ that survive the depth-map Z-buffer test. Using a radius-ρ neighborhood around the provenance seeds, the cleaned point cloud is

$$\begin{aligned} P_i^t &= P_{i,\text{seed}}^t \cup \left(\mathcal{Z}_i^t \cap \mathcal{N}_\rho\big(P_{i,\text{seed}}^t\big)\right), \\ P_k &= \text{Clean}\left(\bigcup_{o_i^t \in \mathcal{O}(e_k)} P_i^t\right). \end{aligned} \tag{5}$$

Here, P_i^t is the evidence contributed by one observation and P_k is the cleaned point cloud of entity e_k; $\mathcal{N}_\rho(S)$ denotes points within distance ρ of set S. Restricting supplementary points to a neighborhood of the provenance seeds prevents nearby background geometry from entering through 2D mask overlap.

3.3 MULTI-VIEW FUNCTIONAL ORIENTATION

For persistent entity e_k, let V_{e_k} denote the selected set of representative frame indices. Qwen3.6 predicts in each view $t \in V_{e_k}$ a categorical image-space front f_t from the finite set of left, right, up, down, toward-camera, away-camera, and unknown, together with confidence $w_t \in [0, 1]$. The categorical multi-view vote shown in the figure is

$$\widehat{f}_{\mathrm{vlm}}(e_k) = \arg\max_{d \in \mathcal{D}} \sum_{t \in V_{e_k}} w_t \, \mathbf{1}[f_t = d]. \tag{6}$$

Thus, d is a candidate direction label, f_t is the label predicted in view t, w_t is that view's confidence, and $\mathbf{1}[\cdot]$ is the indicator function. Unknown or low-confidence votes are excluded from the valid view set used below.

To obtain a physical direction, write $d_t := f_t$ for each accepted vote and let $\mathbf{c}(d_t)$ be the corresponding unit vector in the camera coordinate frame. If R_t is the rotational block of the camera-to-world pose T_t, then

$$\begin{aligned} \mathbf{g}_t &= R_t \mathbf{c}(d_t), \\ \widetilde{\mathbf{f}}^{\mathrm{sem}}_{e_k} &= \sum_{t \in V_{e_k}} w_t \mathbf{g}_t \\ &= \sum_{t \in V_{e_k}} w_t R_t \mathbf{c}(d_t), \\ \mathbf{f}^{\mathrm{sem}}_{e_k} &= \frac{\widetilde{\mathbf{f}}^{\mathrm{sem}}_{e_k}}{\|\widetilde{\mathbf{f}}^{\mathrm{sem}}_{e_k}\|_2}. \end{aligned} \tag{7}$$

Here, $\mathbf{g}_t$ is one camera-motion-compensated world-frame vote, $\widetilde{\mathbf{f}}^{\mathrm{sem}}_{e_k}$ is the unnormalized weighted sum depicted in Fig. 2, and $\mathbf{f}^{\mathrm{sem}}_{e_k}$ is its unit semantic-front direction. The weighted resultant length can be used as a confidence diagnostic, but it is not needed to define the fused orientation.

3.4 GEOMETRY-CONSTRAINED SEMANTIC COORDINATE SYSTEM

The multi-view estimate supplies the functional sign of the Front axis, whereas the clean entity cloud P_k determines which axes are geometrically admissible. Let ℓ_k be the semantic category inherited by persistent entity e_k, and let $\mathcal{A}_{\ell_k}(P_k) \subseteq \mathbb{S}^2$ be the category-dependent set of feasible unit front directions. We select

$$\mathbf{f}^*_{e_k} = \arg\max_{\mathbf{f} \in \mathcal{A}_{\ell_k}(P_k)} \mathbf{f}^\top \mathbf{f}^{\mathrm{sem}}_{e_k}. \tag{8}$$

For categories with explicit structure, $\mathcal{A}_{\ell_k}$ contains candidates such as plane normals, tabletop axes, or backrest-derived directions; the semantic vote resolves their otherwise ambiguous sign.

Let $\mathbf{u}_w$ be gravity-aligned world Up. The letters $\mathbf{f}$, $\mathbf{r}$, and $\mathbf{u}$ in the figure denote Front, Right, and Up, respectively. We construct the right-handed semantic FRU frame using $\mathrm{nrm}(\mathbf{x}) := \mathbf{x}/\|\mathbf{x}\|_2$ and $P^{\perp}_{e_k} := \mathbf{I} - \mathbf{f}^*_{e_k} \mathbf{f}^{*\top}_{e_k}$:

$$\begin{aligned} \mathbf{u}_{e_k} &= \mathrm{nrm}\big(P^{\perp}_{e_k} \mathbf{u}_w\big), \\ \mathbf{r}_{e_k} &= \mathrm{nrm}\big(\mathbf{u}_{e_k} \times \mathbf{f}^*_{e_k}\big), \\ \mathbf{u}_{e_k} &\leftarrow \mathrm{nrm}\big(\mathbf{f}^*_{e_k} \times \mathbf{r}_{e_k}\big), \\ R^{\mathrm{sem}}_{e_k} &= \big[\mathbf{f}^*_{e_k}, \mathbf{r}_{e_k}, \mathbf{u}_{e_k}\big] \in \mathrm{SO}(3). \end{aligned} \tag{9}$$

The remaining in-plane axis is stabilized by PCA in the plane orthogonal to $\mathbf{f}^*_{e_k}$, while its sign is chosen to agree with $\mathbf{u}_w$. Finally, each cleaned point $\mathbf{x} \in P_k$ is expressed as $\boldsymbol{\alpha}(\mathbf{x}) = (R^{\mathrm{sem}}_{e_k})^\top (\mathbf{x} - \bar{\mathbf{x}}_k)$. Coordinate-wise minima and maxima of $\boldsymbol{\alpha}$ determine the semantic oriented bounding box; its axes are therefore exactly Front, Right, and Up rather than arbitrary PCA directions. The resulting entity representation couples persistent identity e_k, category ℓ_k, clean 3D geometry P_k, and semantic frame $R^{\mathrm{sem}}_{e_k}$.

4 EXPERIMENTS

Table 1: Class-agnostic cross-frame instance association on ScanNet. IDSW is lower-is-better; all other metrics are higher-is-better. HOTA is the primary metric.

Method	IDF1 ↑	IDSW ↓	sMOTSA ↑	LocA ↑	DetA ↑	AssA ↑	**HOTA ↑**
RGB video							
DEVA(Cheng et al., 2023)	31.8	16.8	15.2	83.5	26.8	37.0	31.3
MASA(Li et al., 2024)	30.4	**1.6**	10.9	82.4	19.7	42.5	28.3
DVIS++(Zhang et al., 2025)	39.9	9.0	23.7	86.8	38.6	35.4	36.7
EntitySAM(Ye et al., 2025)	41.7	11.2	24.8	86.9	38.9	40.7	39.6
LOMM(Lee et al., 2025a)	42.9	10.1	25.6	86.5	39.4	42.2	40.7
CAVIS(Lee et al., 2025b)	45.3	8.4	27.3	87.1	41.8	45.1	43.3
VidEoMT(Cavagnero et al., 2026)	40.4	13.2	17.7	87.7	41.1	36.6	38.3
RGB + external 2D predictions							
XMem(Cheng & Schwing, 2022)	45.8	10.7	27.5	87.8	42.7	45.3	43.8
DeAOT(Yang & Yang, 2022)	48.2	9.3	29.1	88.1	44.5	47.6	46.0
Cutie(Cheng et al., 2024)	52.4	7.7	32.4	88.8	47.6	51.4	49.4
RMem(Zhou et al., 2024)	53.1	7.1	32.8	88.6	47.2	53.2	50.0
SAM2Long(Ding et al., 2025)	58.7	5.8	36.9	89.6	50.6	57.8	54.0
SAM2(Ravi et al., 2025)	54.2	7.4	34.1	89.1	48.9	53.7	51.1
RGB + external 2D predictions + 3D constraints							
IT3DEgo(Zhao et al., 2024)	55.0	7.0	31.4	87.4	45.3	54.1	49.4
EgoSeg3D(Bhalgat et al., 2024)	58.9	5.6	35.8	89.0	48.5	58.2	53.0
SAM2Object(Zhao et al., 2025)	61.8	4.8	38.6	89.7	50.5	59.7	54.9
Whareformer(Chalk et al., 2026)	64.0	4.1	40.7	**90.1**	51.5	58.0	54.6
Persistent3D (Ours)	**67.4**	6.4	**46.0**	86.6	**53.4**	**60.6**	**56.5**

We evaluate Persistent3D on benchmark datasets and real-world video sequences, focusing on three aspects of persistent spatial perception: cross-frame instance association, instance-level 3D information aggregation, and functional orientation estimation. In addition, we conduct an ablation study to analyze the contributions of the core components.

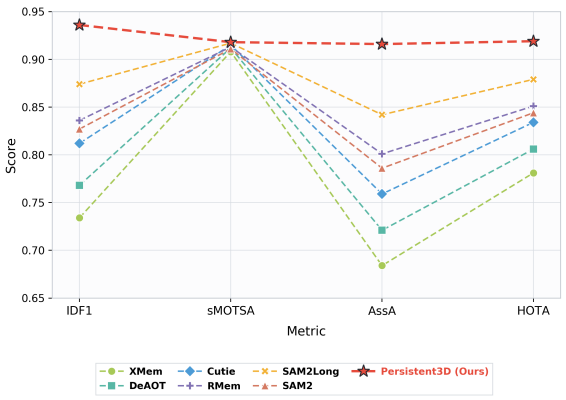


Figure 3: Cross-frame instance association results under unified video frames and pre-generated instance masks. Persistent3D achieves the strongest identity-association performance while all methods retain similar segmentation quality.

4.1 Dataset

We evaluate Persistent3D on video sequences from ScanNet(Dai et al., 2017), ICL-NUIM, self-collected videos, and publicly available online videos, covering challenging scenarios such as long-term occlusion, object re-entry, substantial viewpoint changes, and multiple instances of the same category. For ScanNet, we use the official data to obtain category labels and persistent identities, filter out structural background categories, and automatically refine the official instance masks based on RGB edges. For the other video data, we manually extract video frames and annotate instance segmentation masks, category labels, and cross-frame persistent identities. Details of the data sources are provided in Appendix B.1.

4.2 Cross-Frame Instance Association

We categorize the compared methods into three groups according to the information available during cross-frame association: RGB video, RGB + external 2D predictions, and RGB + external 2D predictions + 3D constraints. This grouping more directly distinguishes the contribution of different information sources to cross-frame identity association and avoids conflating the benefits of

Table 2: Instance-level 3D information aggregation on five ScanNet scenes. Accuracy (Acc.), completeness (Comp.), and Chamfer distance (Cham.) are in centimeters; F@5 denotes the F-score at 5 cm in percent. Bold values indicate the best result for each scene and metric.

	DG-Recon(Ju et al., 2023)				GeoRecon(Wang et al., 2024b)				GP-Recon(Zou et al., 2024)				**Persistent3D (Ours)**			
Scene	Acc. ↓	Comp. ↓	Cham. ↓	F@5 ↑	Acc. ↓	Comp. ↓	Cham. ↓	F@5 ↑	Acc. ↓	Comp. ↓	Cham. ↓	F@5 ↑	Acc. ↓	Comp. ↓	Cham. ↓	F@5 ↑
00003	**5.32**	7.82	6.57	61.9	6.97	9.34	8.16	52.7	5.48	9.07	7.28	60.3	5.49	**6.98**	**6.24**	**64.0**
00035	4.56	6.75	5.66	68.3	6.06	8.05	7.06	59.6	**4.32**	7.81	6.07	66.7	4.86	**5.91**	**5.39**	**70.2**
00036	3.98	6.11	5.05	73.1	5.43	7.46	6.45	65.3	**3.75**	7.16	5.46	72.1	4.37	**5.32**	**4.85**	**75.7**
00061	**4.74**	7.08	5.91	66.8	6.31	8.51	7.41	58.2	4.86	8.27	6.57	65.3	5.14	**6.27**	**5.71**	**68.8**
00701	4.90	7.24	6.07	62.4	6.48	8.64	7.56	54.2	**4.34**	8.44	6.39	60.6	5.39	**6.52**	**5.96**	**63.8**
Average	4.70	7.00	5.85	66.5	6.25	8.40	7.33	58.0	**4.55**	8.15	6.35	65.0	5.05	**6.20**	**5.63**	**68.5**

additional priors with the association capability of the methods themselves. As show in Tab. 1, the three groups exhibit a relatively clear progression in performance. Methods using only RGB video generally struggle to achieve both high DetA and AssA, and their identities are more prone to fragmentation over long temporal spans. With external 2D predictions, methods such as Cutie(Cheng et al., 2024), RMem(Zhou et al., 2024), and SAM2Long(Ding et al., 2025) benefit from more stable instance masks, allowing cross-frame propagation to operate on more reliable observations. Nevertheless, 2D observations still face a key bottleneck: observations from different temporal segments may fail to be correctly associated with the same instance. This limitation becomes more evident for methods incorporating 3D constraints. Although SAM2Object(Zhao et al., 2025) and Whareformer(Chalk et al., 2026) achieve relatively high HOTA, their AssA remains lower than that of Persistent3D. Moreover, Whareformer attains the highest LocA but does not achieve the best overall association performance, indicating that more accurate local mask localization does not necessarily translate into more stable long-term identity consistency. Persistent3D achieves a HOTA of 56.5 and obtains the best results in IDF1, DetA, and AssA. Our method not only preserves more valid instance observations but also maintains identity consistency among them over longer temporal spans. In particular, its AssA reaches 60.6, further indicating that the performance gain primarily comes from improved association rather than more accurate single-frame segmentation. Overall, the improvement of Persistent3D is mainly attributable to better long-term identity continuity: even when local propagation is interrupted, the same physical instance can still be recovered through full-sequence constraints and instance reconnection, rather than relying solely on short-range matching between consecutive frames.

To compare the cross-frame identity association capability of different methods, we use unified inputs for all models: exactly the same video frames and pre-generated instance masks. As shown in Fig. 3, under this setting, Persistent3D achieves AssA, HOTA, and IDF1 scores of approximately 0.916, 0.919, and 0.936, respectively, outperforming the other methods and demonstrating more stable long-term identity consistency, overall tracking quality, and sequence-level identity preservation. Since all methods share the same 2D instance masks, their sMOTSA scores remain concentrated around 0.91, indicating that the performance differences mainly stem from cross-frame association capability rather than 2D detection or segmentation quality.

4.3 3D Information Aggregation

To evaluate the effect of persistent identities on cross-view instance geometry aggregation, we compare the instance-level 3D geometry produced by different methods on five ScanNet scenes. As shown in Tab. 2, the main advantages of Persistent3D lie in Completeness, Chamfer, and F@5, while its Accuracy is not the best. This is consistent with the objective of our method: Persistent3D does not primarily aim to further improve the local positioning accuracy of individual 3D points, but instead relies on stable persistent identities to continuously aggregate valid geometric observations of the same instance across different viewpoints, thereby obtaining a more complete instance-level 3D representation. In terms of the average results, GP-Recon(Ju et al., 2023) achieves lower Accuracy, whereas Persistent3D achieves substantially better Completeness and consequently obtains lower Chamfer Distance and higher F@5. In other words, the main benefit of persistent identities is to increase the effective coverage of the true object surface while keeping the overall geometric error low. This trend is largely consistent across different scenes, indicating that cross-view identity con-

sistency effectively improves the completeness and overall quality of instance-level 3D information aggregation.

4.4 Functional Orientation Estimation

To evaluate Persistent3D's ability to estimate the functional front direction of objects, we select five frequently occurring object categories. As shown in the table 3, Persistent3D consistently recovers functional orientations across different categories, with most predictions falling within relatively small angular errors. Performance still varies across categories. Television achieves the best results because its display surface and viewing direction provide clear and consistent front-direction cues across multiple viewpoints. Chair and Sofa also exhibit relatively distinctive structural and functional orientations, resulting in lower overall errors. In contrast, the functional fronts of Cabinet and Desk are more easily affected by structural symmetry, weak texture, and partial occlusion. As the tolerance increases to 30°, the accuracy of all categories improves noticeably and the performance gap between categories becomes smaller, indicating that although some predictions exhibit moderate angular deviations, the overall functional direction can still be recovered correctly.

Table 3: Functional orientation estimation for the five most frequent object categories. Angular errors are in degrees, and threshold accuracies are percentages. The best results are bold.

Category	Median error ↓	Acc@30° ↑	Acc@15° ↑
Chair	12.7	92.4	73.6
Sofa	11.4	94.1	78.8
Cabinet	14.5	89.8	68.1
Television	**8.6**	**97.3**	**86.7**
Desk	15.8	88.9	64.8
Mean	**12.6**	**92.5**	**74.4**

4.5 Ablation Study

To evaluate the contribution of each component in persistent identity inference, we compare several configurations: replacing bidirectional propagation with forward-only propagation; relaxing the one-to-one matching constraint to allow non-one-to-one associations; removing same-frame exclusion, 3D overlap, color similarity, and frame exclusivity individually; and comparing them with the full model. As shown in Table 4, both forward-only propagation and relaxed one-to-one matching lead to clear performance degradation, with non-one-to-one association having the largest impact on HOTA, IDF1, and IDSW, reflecting the importance of identity uniqueness during local association. Removing same-frame exclusion causes a slight decrease in overall identity consistency, indicating that same-frame conflict control is important for avoiding erroneous merges between similar instances. During long-term reconnection, 3D overlap contributes the most, and removing it leads to a noticeable drop in HOTA and IDF1; color similarity further assists in assessing consistency between identity fragments by providing additional appearance cues. Frame exclusivity mainly constrains potentially conflicting reconnections in relatively rare cases and therefore has the smallest impact.

Table 4: Leave-one-component-out ablation on ScanNet.

Variant	IDF1 ↑	IDSW ↓	LocA ↑	HOTA ↑
w/o Bidirectional Prop.	58.2	26.8	83.2	48.1
w/o One-to-One	53.5	35.1	82.1	45.7
w/o Same-Frame Excl.	64.8	10.1	85.8	54.2
w/o 3D Overlap	62.5	15.8	85.0	52.4
w/o Color Similarity	62.2	16.2	84.8	52.1
w/o Frame Exclusivity	66.1	8.0	86.2	55.3
Full	**67.4**	**6.4**	**86.6**	**56.5**

5 Conclusion

We presented Persistent3D, an open-vocabulary framework for persistent object spatial perception from monocular video. By combining local cross-frame association, same-frame mutual exclusion, and 3D instance reconnection, Persistent3D maintains persistent identities under occlusion, similar instances, and large viewpoint changes. Based on these identities, multi-view observations are further aggregated to recover instance-level 3D geometry and functional Front–Right–Up orientations.

AI USE STATEMENT

In this work, we used generative AI tools to assist with polishing the writing and improving the clarity and readability of the manuscript. We also used AI tools to assist with literature retrieval and discovery of potentially relevant related work. All AI-assisted text and literature suggestions were manually reviewed and verified by the authors, and relevant references were checked against their original sources. We take full responsibility for the final content of this work.

ETHICS STATEMENT

This research was conducted in accordance with established ethical standards for scientific work. Topics considered include, but are not limited to, the involvement of human subjects, dataset usage and release practices, potentially harmful insights, research methodologies and applications, conflicts of interest and sponsorship, discrimination/bias/fairness concerns, privacy and security issues, legal compliance, and research integrity (e.g., IRB approvals, documentation, and research ethics). Specifically, our study does not involve human subjects or personally identifiable information, and therefore no Institutional Review Board (IRB) approval was required. All datasets used are publicly available and released under appropriate licenses. Potential risks, including fairness, bias, privacy, and unintended harmful use of the findings, were carefully assessed, and steps were taken to minimize such risks. We affirm that our work complies with research integrity guidelines, including accurate reporting, transparency, and reproducibility.

REPRODUCIBILITY STATEMENT

The main paper provides detailed descriptions of the methodology and inference settings. The experimental data include public datasets, videos captured by the authors, and publicly available online videos, with their data sources and collection procedures documented in the appendix. We provide the source code, configuration files, environment specifications, data preprocessing and execution scripts, as well as the relevant pretrained-model settings, in an anonymous repository, enabling other researchers to reproduce our experimental results as completely and accurately as possible.

REFERENCES

Shikhar Bahl, Russell Mendonca, Lili Chen, Unnat Jain, and Deepak Pathak. Affordances from human videos as a versatile representation for robotics. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 01–13. IEEE, 2023.

Yash Bhalgat, Vadim Tschernezki, Iro Laina, Joao F Henriques, Andrea Vedaldi, and Andrew Zisserman. 3d-aware instance segmentation and tracking in egocentric videos. In *Asian Conference on Computer Vision*, pp. 347–364. Springer, 2024.

Niccolò Cavagnero, Narges Norouzi, Gijs Dubbelman, and Daan de Geus. Pmt: Plain mask transformer for image and video segmentation with frozen vision encoders. *arXiv preprint arXiv:2603.25398*, 2026.

Jacob Chalk, Saptarshi Sinha, Dima Damen, Yannis Kalantidis, and Diane Larlus. Whareformer: Learning to track what is where in long egocentric videos. *arXiv preprint arXiv:2607.08537*, 2026.

Dengsheng Chen, Jun Li, Zheng Wang, and Kai Xu. Learning canonical shape space for category-level 6d object pose and size estimation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11970–11979. IEEE, 2020.

Ho Kei Cheng and Alexander G Schwing. Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *European conference on computer vision*, pp. 640–658. Springer, 2022.

Ho Kei Cheng, Seoung Wug Oh, Brian Price, Alexander Schwing, and Joon-Young Lee. Tracking anything with decoupled video segmentation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1316–1326. IEEE, 2023.

Ho Kei Cheng, Seoung Wug Oh, Brian Price, Joon-Young Lee, and Alexander Schwing. Putting the object back into video object segmentation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3151–3161. IEEE, 2024.

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

Runyu Ding, Jihan Yang, Chuhui Xue, Wenqing Zhang, Song Bai, and Xiaojuan Qi. Pla: Language-driven open-vocabulary 3d scene understanding. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7010–7019. IEEE, 2023.

Shuangrui Ding, Rui Qian, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Yuwei Guo, Dahua Lin, and Jiaqi Wang. Sam2long: Enhancing sam 2 for long video segmentation with a training-free memory tree. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13614–13624. IEEE, 2025.

Qiao Gu, Ali Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Rama Chellappa, et al. Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5021–5028. IEEE, 2024.

De-An Huang, Zhiding Yu, and Anima Anandkumar. Minvis: A minimal video instance segmentation framework without video-based training. *Advances in Neural Information Processing Systems*, 35:31265–31277, 2022.

Jihong Ju, Ching Wei Tseng, Oleksandr Bailo, Georgi Dikov, and Mohsen Ghafoorian. Dg-recon: Depth-guided neural 3d scene reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18184–18194, 2023.

Seunghun Lee, Jiwan Seo, Minwoo Choi, Kiljoon Han, Jahoon Jeong, Zane Durante, Ehsan Adeli, Sang Hyun Park, and Sunghoon Im. Lomm: Latest object memory management for temporally consistent video instance segmentation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13719–13729. IEEE, 2025a.

Seunghun Lee, Jiwan Seo, Kiljoon Han, Minwoo Choi, and Sunghoon Im. Cavis: Context-aware video instance segmentation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4507–4517. IEEE, 2025b.

Gen Li, Nikolaos Tsagkas, Jifei Song, Ruaridh Mon-Williams, Sethu Vijayakumar, Kun Shao, and Laura Sevilla-Lara. Learning precise affordances from egocentric videos for robotic manipulation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10581–10591. IEEE, 2025.

Siyuan Li, Lei Ke, Martin Danelljan, Luigi Piccinelli, Mattia Segu, Luc Van Gool, and Fisher Yu. Matching anything by segmenting anything. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18963–18973. IEEE, 2024.

Jiehong Lin, Zewei Wei, Zhihao Li, Songcen Xu, Kui Jia, and Yuanqing Li. Dualposenet: Category-level 6d object pose and size estimation using dual pose network with refined learning of pose consistency. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3540–3549. IEEE, 2021.

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.

Yuzheng Liu, Siyan Dong, Shuzhe Wang, Yingda Yin, Yanchao Yang, Qingnan Fan, and Baoquan Chen. Slam3r: Real-time dense scene reconstruction from monocular rgb videos. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16651–16662. IEEE, 2025.

Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, et al. Openeqa: Embodied question answering in the era of foundation models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16488–16498. IEEE, 2024.

Riku Murai, Eric Dexheimer, and Andrew J Davison. Mast3r-slam: Real-time dense slam with 3d reconstruction priors. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16695–16705. IEEE, 2025.

Songyou Peng, Kyle Genova, Chiyu Jiang, Andrea Tagliasacchi, Marc Pollefeys, and Thomas Funkhouser. Openscene: 3d scene understanding with open vocabularies. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 815–824. IEEE, 2023.

Zekun Qi, Wenyao Zhang, Yufei Ding, Runpei Dong, Xinqiang Yu, Jingwen Li, Lingyun Xu, Baoyu Li, Xialin He, Guofan Fan, et al. Sofar: Language-grounded orientation bridges spatial reasoning and object manipulation. *Advances in neural information processing systems*, 38:76367–76412, 2026.

Qwen Team. Qwen3.6-27B: Flagship-level coding in a 27B dense model, April 2026. URL `https://qwen.ai/blog?id=qwen3.6-27b`.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. In *International Conference on Learning Representations*, volume 2025, pp. 28085–28128, 2025.

Antoni Rosinol, Arjun Gupta, Marcus Abate, Jingnan Shi, and Luca Carlone. 3d dynamic scene graphs: Actionable spatial perception with places, objects, and humans. *arXiv preprint arXiv:2002.06289*, 2020.

Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9338–9346. IEEE, 2019.

Yuqing Shao, Yuchen Yang, Rui Yu, Weilong Li, Xu Guo, Huaicheng Yan, Wei Wang, and Xiao Sun. From detection to association: Learning discriminative object embeddings for multi-object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6878–6888, 2026.

Zachary Teed and Jia Deng. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. *Advances in neural information processing systems*, 34:16558–16569, 2021.

He Wang, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, and Leonidas J Guibas. Normalized object coordinate space for category-level 6d object pose and size estimation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2637–2646. IEEE, 2019.

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

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19757–19767. IEEE, 2024a.

Yanmei Wang, Fupeng Chu, Zhi Han, and Yandong Tang. Georecon: Geometric coherence for online 3d scene reconstruction from monocular video. *IEEE Robotics and Automation Letters*, 10 (1):500–507, 2024b.

Zehan Wang, Ziang Zhang, Tianyu Pang, Chao Du, Hengshuang Zhao, and Zhou Zhao. Orient anything: Learning robust object orientation estimation from rendering 3d models. *arXiv preprint arXiv:2412.18605*, 2024c.

Yi Wei, Linqing Zhao, Wenzhao Zheng, Zheng Zhu, Jie Zhou, and Jiwen Lu. Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21672–21683. IEEE, 2023.

Bowen Wen and Kostas Bekris. Bundletrack: 6d pose tracking for novel objects without instance or category-level 3d models. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 8067–8074. IEEE, 2021.

Bowen Wen, Jonathan Tremblay, Valts Blukis, Stephen Tyree, Thomas Müller, Alex Evans, Dieter Fox, Jan Kautz, and Stan Birchfield. Bundlesdf: Neural 6-dof tracking and 3d reconstruction of unknown objects. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 606–617. IEEE, 2023.

Junfeng Wu, Qihao Liu, Yi Jiang, Song Bai, Alan Yuille, and Xiang Bai. In defense of online models for video instance segmentation. In *European Conference on Computer Vision*, pp. 588–605. Springer, 2022.

Shun-Cheng Wu, Johanna Wald, Keisuke Tateno, Nassir Navab, and Federico Tombari. Scenegraphfusion: Incremental 3d scene graph prediction from rgb-d sequences. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7511–7521. IEEE, 2021.

Zongxin Yang and Yi Yang. Decoupling features in hierarchical propagation for video object segmentation. *Advances in Neural Information Processing Systems*, 35:36324–36336, 2022.

Zongxin Yang, Yunchao Wei, and Yi Yang. Associating objects with transformers for video object segmentation. *Advances in neural information processing systems*, 34:2491–2502, 2021.

Mingqiao Ye, Seoung Wug Oh, Lei Ke, and Joon-Young Lee. Entitysam: Segment everything in video. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24234–24243. IEEE, 2025.

Tao Zhang, Xingye Tian, Yikang Zhou, Shunping Ji, Xuebo Wang, Xin Tao, Yuan Zhang, Pengfei Wan, Zhongyuan Wang, and Yu Wu. Dvis++: Improved decoupled framework for universal video segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(7): 5918–5929, 2025.

Jihuai Zhao, Junbao Zhuo, Jiansheng Chen, and Huimin Ma. Sam2object: Consolidating view consistency via sam2 for zero-shot 3d instance segmentation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19325–19334. IEEE, 2025.

Yunhan Zhao, Haoyu Ma, Shu Kong, and Charless Fowlkes. Instance tracking in 3d scenes from egocentric videos. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21933–21944. IEEE, 2024.

Junbao Zhou, Ziqi Pang, and Yu-Xiong Wang. Rmem: Restricted memory banks improve video object segmentation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18602–18611. IEEE, 2024.

Zi-Xin Zou, Shi-Sheng Huang, Yan-Pei Cao, Tai-Jiang Mu, Ying Shan, Hongbo Fu, and Song-Hai Zhang. Gp-recon: Online monocular neural 3d reconstruction with geometric prior. *IEEE Transactions on Visualization and Computer Graphics*, 31(9):4790–4804, 2024.

A Additional Methodological Details

This section specifies the reproducible implementation details of the Persistent3D pipeline: notation, association thresholds, 2D-to-3D lifting, point-cloud cleaning, semantic OBB refitting, gravity alignment, and geometric predicates. The presentation follows the observation-to-entity notation used in Fig. 2.

A.1 Notation and Coordinate Conventions

We use the symbols in Fig. 2 throughout this appendix. For RGB frames $\{I_t\}_{t=1}^{T}$, reconstruction returns camera poses T_t, depth maps D_t, the global cloud P_{scene}, and point provenance; 2D perception returns $o_i^t = (\ell_i^t, M_i^t, t)$. The persistent graph is $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, with observation nodes o_i^t and entity nodes e_k. The associated observation set and clean cloud are $\mathcal{O}(e_k) = \{o_i^t : o_i^t \mapsto e_k\}$ and $P_k = \text{Clean}(\bigcup_{o_i^t \in \mathcal{O}(e_k)} P_i^t)$. After gravity alignment, entity e_k is represented by

$$\mathcal{B}(e_k) = \big(\mathbf{c}_k, R_{e_k}^{\text{sem}}, \boldsymbol{\ell}_k\big), \qquad R_{e_k}^{\text{sem}} = [\mathbf{f}_{e_k}^*, \mathbf{r}_{e_k}, \mathbf{u}_{e_k}] \in \text{SO}(3), \tag{A1}$$

where $\mathbf{c}_k$ is the OBB center, $\boldsymbol{\ell}_k$ contains its three side lengths, and the columns are semantic Front, Right, and Up. A relation "A relative to B" uses the columns of $R_{e_B}^{\text{sem}}$ as its reference axes.

A.2 Association Implementation Settings

The association objectives and constraints are given in Eqs. equation 1–equation 3. In the implementation, each detected mask is propagated in both directions within a 20-frame window. Candidate pairs with mask IoU below 0.30 are discarded; the remaining pairs are sorted by confidence and accepted greedily subject to one-to-one source and target usage. During global conflict repair, ties are resolved deterministically by temporal offset and instance identifier.

For the fragment reconnection rule in Eq. equation 4, both the scene-point overlap threshold and RGB-histogram-intersection threshold are 0.50. Complete frame sets must be disjoint. This condition is re-evaluated after every union so that a transitive merge cannot reintroduce a same-frame conflict.

A.3 Lifting 2D Masks to 3D Points

For a reconstruction-world point $\widetilde{\mathbf{p}}_n = [\mathbf{p}_n^\top, 1]^\top$, we use the same camera-to-world pose T_t as in the main method. Its camera-frame projection is computed by

$$\begin{aligned} \widetilde{\mathbf{p}}_n^{c,t} &= T_t^{-1} \widetilde{\mathbf{p}}_n, \\ u_n = f_x \frac{x_n^c}{z_n^c} + c_x, &\qquad v_n = f_y \frac{y_n^c}{z_n^c} + c_y. \end{aligned} \tag{A2}$$

Only points with positive finite depth in $[10^{-6}, 20]$ meters and projected coordinates inside the image are retained. When multiple points round to the same pixel, a z-buffer keeps the point with minimum camera depth.

The exporter supplies a source keyframe and source pixel for each scene point. We erode each mask by three pixels and use provenance-consistent points inside the eroded mask as seeds. The neighborhood radius in Eq. equation 5 is $\rho = 0.05$ m; provenance is used for identity and projection only for local geometric completion, reducing leakage from nearby surfaces.

A.4 Object-Point Cleaning

For entity e_k, let a_k, $\mathbf{m}_k$, and q_k denote mask area, 3D point centroid, and maximum axis-aligned extent. Entities with fewer than 20 points are invalid. In strict cleaning mode, the robust centroid-distance threshold is

$$\begin{aligned} \widetilde{\mathbf{m}} &= \text{median}_k(\mathbf{m}_k), \qquad d_k = \|\mathbf{m}_k - \widetilde{\mathbf{m}}\|_2, \\ \tau_c &= \min\big(1.0, \text{median}(d_k) + 3.5 \max(\text{MAD}(d_k), 10^{-6})\big)\,. \end{aligned} \tag{A3}$$

An instance is rejected if $d_k > \tau_c$, if its area lies outside a factor of three of the median area, or if its maximum extent exceeds three times the median extent. If too few instances remain, those closest to the median centroid are restored. The default *preserve* mode applies only the minimum-point validity check; the stricter consistency test is an optional ablation setting.

Duplicate points are removed before voxelization. With voxel width $v = 0.02$ m, all points whose index $\lfloor \mathbf{p}/v \rfloor$ is identical are replaced by their mean. Strict mode additionally applies statistical and

radius outlier removal. If $\bar{d}_k$ is the mean distance from point k to its 16 nearest neighbors, statistical filtering retains

$$\bar{d}_k \leq \mu_{\bar{d}} + 1.5\sigma_{\bar{d}}. \tag{A4}$$

The radius filter then requires at least four other points within 0.05 m. Fallbacks to earlier cleaning stages prevent an object from being reduced below 20 points.

A.5 Semantic OBB Estimation

Geometric initialization. For the cleaned cloud $P_k = \{\mathbf{x}_n\}_{n=1}^{N_k}$ of entity e_k, a PCA OBB is initialized from

$$C_k = \frac{1}{N_k} \sum_n (\mathbf{x}_n - \bar{\mathbf{x}}_k)(\mathbf{x}_n - \bar{\mathbf{x}}_k)^\top. \tag{A5}$$

The eigenvectors of C_k, ordered by decreasing eigenvalue and orthonormalized into a right-handed basis, form the initial rotation. In this basis, the coordinate-wise minimum and maximum give the OBB lengths, and their midpoint is transformed back to obtain the OBB center. Objects with fewer than four valid points or a degenerate floor footprint are not used for spatial supervision.

Multi-view semantic-front settings. The voting formulation is given in Eq. equation 7. Up to ten representative frames are selected by descending instance support. Votes marked unknown, votes below confidence 0.35, and invalid transformed vectors are discarded. Category geometry can further constrain the front axis: cabinet, curtain, picture, and picture frame use the normal of the smallest variance plane; a desk uses a co-visible chair direction when available and otherwise the tabletop short axis; and a sofa uses the horizontal normal from its seat/support region toward its fitted backrest. The multi-view vote is used to resolve an otherwise ambiguous sign where appropriate.

Front-constrained OBB refitting. The semantic front is fixed rather than merely relabeling an arbitrary PCA axis. Let $P_\perp = I - \mathbf{f}^*_{e_k} \mathbf{f}^{*\top}_{e_k}$. We project centered points into the plane orthogonal to $\mathbf{f}^*_{e_k}$ and eigendecompose

$$C_\perp = \frac{1}{N_k} \sum_n P_\perp (\mathbf{x}_n - \bar{\mathbf{x}}_k)(\mathbf{x}_n - \bar{\mathbf{x}}_k)^\top P_\perp. \tag{A6}$$

Among the in-plane eigenvectors, the one most aligned with world up is chosen as $\mathbf{u}_{e_k}$ and its sign is made upward. We then set $\mathbf{r}_{e_k} \propto \mathbf{u}_{e_k} \times \mathbf{f}^*_{e_k}$ and recompute $\mathbf{u}_{e_k} \propto \mathbf{f}^*_{e_k} \times \mathbf{r}_{e_k}$. Projecting all points onto $R^{\text{sem}}_{e_k} = [\mathbf{f}^*_{e_k}, \mathbf{r}_{e_k}, \mathbf{u}_{e_k}]$ and taking coordinate-wise minima and maxima yields a geometry-consistent semantic OBB.

A.6 Gravity-Aligned World Frame

The scene point cloud is aligned before relation generation. Candidate ground planes are generated by deterministic RANSAC. A hypothesis $\mathbf{n}^\top \mathbf{x} = b$ is retained only when $|\mathbf{n}^\top \mathbf{u}_{\text{cam}}| \geq 0.60$; its inliers satisfy

$$|\mathbf{n}^\top \mathbf{x} - b| \leq \tau_g, \qquad \tau_g = \min(0.05, \max(0.01, 0.003 D_{\text{scene}})). \tag{A7}$$

The lowest plane below the first camera is selected among candidates with at least 20% of the maximum supported inlier count. Its normal is refined as the smallest-eigenvalue eigenvector of the inlier covariance and oriented toward the first camera's up axis.

Let $\mathbf{u}$ be the refined ground normal and $\mathbf{o}$ the first camera optical axis. The horizontal world-front and left directions are

$$\mathbf{f} = \frac{\mathbf{o} - (\mathbf{o}^\top \mathbf{u})\mathbf{u}}{\|\mathbf{o} - (\mathbf{o}^\top \mathbf{u})\mathbf{u}\|_2}, \qquad \mathbf{l} = \frac{\mathbf{u} \times \mathbf{f}}{\|\mathbf{u} \times \mathbf{f}\|_2}. \tag{A8}$$

With $\mathbf{r} = -\mathbf{l}$, the raw-to-aligned rotation is $R_{w\leftarrow r} = [\mathbf{r}, \mathbf{f}, \mathbf{u}]^\top$. The origin is the orthogonal projection of the first camera center onto the selected ground plane, and any raw point is aligned by

$$\mathbf{x}^w = R_{w\leftarrow r}(\mathbf{x}^r - \mathbf{o}_g^r). \tag{A9}$$

A.7 Pairwise 3D Spatial Relations

Directional predicates. For an axis $\mathbf{a}$, the lenient rule uses the signed center offset

$$\delta_{A,B}(\mathbf{a}) = (\mathbf{c}_A - \mathbf{c}_B)^\top \mathbf{a}. \tag{A10}$$

The positive or negative relation is assigned when δ is greater than 10^{-4} or less than -10^{-4}, respectively. The optional strict rule instead projects all OBB corners onto $\mathbf{a}$ and requires the two intervals to be separated by the same tolerance. World-centric predicates use the aligned axes, while object-centric predicates use the semantic axes of the reference object. Object-centric predicates are omitted when that orientation is unreliable.

Separation and topology. We use the separating axis theorem (SAT) with the three axes of each OBB and the nine pairwise cross products. For a unit candidate axis $\mathbf{a}$, the projected radius of object i is

$$\rho_i(\mathbf{a}) = \sum_{k=1}^{3} \frac{\ell_{i,k}}{2} |\mathbf{r}_{i,k}^\top \mathbf{a}|, \tag{A11}$$

and the signed SAT gap is

$$g(A, B) = \max_{\mathbf{a}\in\mathcal{A}} \left(|(\mathbf{c}_B - \mathbf{c}_A)^\top \mathbf{a}| - \rho_A(\mathbf{a}) - \rho_B(\mathbf{a}) \right). \tag{A12}$$

Thus $g > 0$ denotes separation and $g < 0$ penetration. Containment is verified by transforming all eight corners of the inner OBB into the outer OBB frame and checking them against the outer half-extents with 0.01 m tolerance. To avoid contradictory labels, every directed pair receives at most one topological relation, using the following priority: inside/contains, overlapping when $g < -0.03$ m, touching when $|g| \le 0.03$ m, and near when $\max(0, g) \le \tau_n$. The scene-adaptive near threshold is

$$\tau_n = \max(0.05, 0.05 D_{\text{scene}}). \tag{A13}$$

If two nearly identical boxes contain one another within tolerance, the relation is reported as overlapping rather than assigning an arbitrary containment direction.

Scalar geometry. The OBB volume is $V_i = \prod_k \ell_{i,k}$. Its orthographic footprint on the world XY plane is computed without axis alignment as

$$A_i^{xy} = \ell_x \ell_y |(\mathbf{r}_x \times \mathbf{r}_y)_z| + \ell_x \ell_z |(\mathbf{r}_x \times \mathbf{r}_z)_z| + \ell_y \ell_z |(\mathbf{r}_y \times \mathbf{r}_z)_z|. \tag{A14}$$

These quantities, center distance, and nonnegative surface separation $\max(0, g)$ support metric comparisons and spatial predicates.

A.8 VISIBILITY, OCCLUSION, AND VIEWPOINT CHANGE

Identity-preserving visibility. Cleaned object points are projected into each real frame. A point is counted as visible only if its rounded pixel lies inside the associated current-frame instance mask. Among points landing on the same mask pixel, only the nearest point is kept to account for self-occlusion. If $\mathcal{V}_i^t$ is this set, the visibility ratio is

$$r_i^t = \frac{|\mathcal{V}_i^t|}{|\mathcal{P}_i|}, \tag{A15}$$

where the denominator is the complete unique cleaned cloud of object i and is fixed across frames. At least three verified points are required (or all points for a smaller cloud). Geometry-only OBB projections may localize an off-screen candidate but never establish instance identity.

For cross-frame change, we use the symmetric relative difference

$$\Delta_r = \frac{r_i^{t_2} - r_i^{t_1}}{\max(r_i^{t_1}, r_i^{t_2})}. \tag{A16}$$

A change is material when $|\Delta_r| \geq 0.10$. We emit the label only when its direction agrees with the change in normalized instance-mask area; a cross-modal disagreement is treated as uncertainty.

Depth-aware labeled occlusion. For each projected support pixel p, object depth $z_i(p)$ is compared with measured scene depth $z_s(p)$. It is depth-consistent when

$$z_i(p) \leq z_s(p) + 0.05 \text{ m}. \tag{A17}$$

Depth-disagreeing pixels are attributed to another labeled object only when current-frame identity masks, local boundary adjacency, and depth ordering jointly support that attribution. The basic acceptance gate requires at least 12 pixels and target coverage of at least 0.01; direct area-overlap evidence uses the stricter coverage 0.05. For disjoint but adjacent masks, the candidate must be locally closer at the boundary, with depth margin $\max(0.05 \text{ m}, 0.05 z_{\text{target}})$. Non-horizontal OBB-face samples provide conservative amodal evidence when the reconstructed cloud contains only previously visible surfaces. Support-contact configurations, ambiguous evidence, and reciprocal occlusion claims are suppressed. The primary occluder is the accepted object with maximum coverage.

The per-mask camera distance used for depth-order decisions is the fifth percentile of valid scene depth inside the identity mask rather than the minimum, which reduces sensitivity to isolated near-depth outliers.

Camera and image-space transformations. For frames s and t, camera translation is $\Delta\mathbf{t} = \mathbf{t}_t - \mathbf{t}_s$. With $R_{s\to t} = R_s^\top R_t$, the rotation magnitude is

$$\theta = \arccos\left(\text{clip}\left(\frac{\text{tr}(R_{s\to t}) - 1}{2}, -1, 1\right)\right). \tag{A18}$$

Translations below 0.03 m and rotations below 5° are treated as insignificant. Object pixel motion uses the projection of one fixed complete-cloud centroid in both frames, and both endpoints must fall inside the correct instance mask. For image size $W \times H$, its normalized magnitude is

$$m_i^{s,t} = \left\| \left(\frac{u_i^t - u_i^s}{W}, \frac{v_i^t - v_i^s}{H} \right) \right\|_2. \tag{A19}$$

Values below 0.03 are unchanged. Otherwise the image-plane displacement is mapped through the target camera rotation and assigned to the nearest aligned world direction. Projected-scale change uses the ratio of clipped projected bounding-box areas and a 0.15 relative dead zone. All chronological frame pairs are used by default; when a limit is requested, pairs ordered by temporal gap are sampled at evenly spaced indices to retain both short- and long-baseline comparisons.

B Additional Experimental Details

B.1 Datasets and Data Sources

The evaluation uses complementary sources to test both controlled indoor reconstruction and unconstrained monocular-video generalization. ScanNet and ICL-NUIM provide standard indoor sequences, while self-collected and publicly available online videos provide changes in viewpoint, occlusion, and object appearance. The table records each source's role and licensing status; original third-party data are not redistributed. This study uses complementary data sources to evaluate both 3D reconstruction in controlled settings and generalization to unconstrained monocular videos. ScanNet and ICL-NUIM provide standard scene sequences, while self-collected and publicly available online videos introduce more challenging conditions, including viewpoint changes, occlusion, and variations in object appearance. Table 5 summarizes the role of each data source in our experiments and its licensing status.

Table 5: Datasets and video sources used in the experiments.

Data Source	Data Type	Acquisition / Source	Usage in This Work	License and Redistribution
ScanNet Dai et al. (2017)	Public RGB-D benchmark dataset	ScanNet official dataset: https://www.scan-net.org/	Used for quantitative or qualitative evaluation in indoor scenes.	Used in accordance with the official ScanNet Terms of Use. We do not redistribute the original ScanNet data.
ICL-NUIM RGB-D Benchmark Dataset ?	Public RGB-D benchmark dataset	Official webpage: https://www.doc.ic.ac.uk/~ahanda/VaFRIC/iclnuim.html	Used as an additional benchmark sequence to assess the performance of our method in standard indoor scenes.	The dataset is used in accordance with its official license. We provide only the official source link and do not redistribute the original data.
Self-collected videos	Self-collected monocular RGB videos	Captured by the authors using an RGB camera in real indoor environments.	Used to evaluate our method on self-collected real-world video sequences.	The videos were collected by the authors and are used solely for research experiments. Unless otherwise specified, the original videos are not publicly redistributed.
Online Video 1	Publicly accessible online video	Xiaohongshu (RED): http://xhslink.com/o/5uqI5dED1mi	Used to further assess the generalization ability of our method in unconstrained real-world scenes.	The video is publicly accessible online. No explicit open reuse license has been verified. Copyright remains with the original creator or rights holder. We do not redistribute the original video.
Online Video 2	Publicly accessible online video	Xiaohongshu (RED): http://xhslink.com/o/3FTWl3cXnp7	Used to further assess the generalization ability of our method in unconstrained real-world scenes.	The video is publicly accessible online. No explicit open reuse license has been verified. Copyright remains with the original creator or rights holder. We do not redistribute the original video.

Continued on the next page

Table 5 continued from the previous page

Data Source	Data Type	Acquisition Source	/ Usage in This Work	License and Redistribution
Online Video 3	Publicly accessible online video	Xiaohongshu (RED): http://xhslink.com/o/1eGzxiVhl0x	Used to further assess the generalization ability of our method in unconstrained real-world scenes.	The video is publicly accessible online. No explicit open reuse license has been verified. Copyright remains with the original creator or rights holder. We do not redistribute the original video.

B.2 Evaluation Metrics

This section defines every metric reported in the main-paper tables. Ratio metrics are multiplied by 100 when reported as percentages. Unless stated otherwise, sequence-level scores are macro-averaged so that long sequences do not dominate the result.

Tracking notation. At mask-IoU threshold α, predicted and ground-truth masks are matched one-to-one. Let TP_α denote the matched detections, and let FP_α and FN_α denote unmatched predictions and missed ground-truth detections, respectively. We use the strict threshold $\alpha_0 = 0.5$ for IDF1, IDSW, sMOTSA, and LocA.

IDF1.

$$\mathrm{IDF1} = \frac{2\,\mathrm{IDTP}}{2\,\mathrm{IDTP} + \mathrm{IDFP} + \mathrm{IDFN}}. \tag{B1}$$

Here IDTP counts detections assigned to the correct persistent identity after global trajectory matching. IDFP counts predicted identity assignments that do not correspond to the matched ground-truth identity, and IDFN counts ground-truth identity observations that are missing from the matched prediction. IDF1 is the harmonic mean of identity precision and identity recall; a higher value means that an object keeps the same identity over a larger fraction of its lifetime.

Identity switches (IDSW).

$$\mathrm{IDSW} = \sum_{g,t} \chi_g^t. \tag{B2}$$

The indicator χ_g^t equals one when ground-truth trajectory g is successfully matched at frame t but its assigned predicted identity differs from the most recent matched assignment; otherwise it is zero. IDSW therefore counts explicit identity changes rather than missed detections. It is the only tracking metric in our tables for which a lower value is better. We first count switches per sequence and then report their mean across sequences, which allows the final value to be non-integer.

Segmentation-aware MOTSA (sMOTSA). Let $\mathcal{M}_{\alpha_0}$ be the matched mask pairs and define $S_{\alpha_0} = \sum_{(g,p)\in\mathcal{M}_{\alpha_0}} \mathrm{IoU}(g,p)$. Then

$$\mathrm{sMOTSA} = \frac{S_{\alpha_0} - \mathrm{FP}_{\alpha_0} - \mathrm{IDSW}}{N_{\mathrm{GT}}}. \tag{B3}$$

N_{GT} is the total number of ground-truth masks. The numerator rewards accurate mask overlap and subtracts one unit for each false positive and identity switch. Consequently, sMOTSA jointly reflects segmentation, detection, and temporal identity quality; higher is better and severely poor predictions can produce a negative score.

Localization accuracy (LocA).

$$\mathrm{LocA}_\alpha = \frac{1}{|\,\mathrm{TP}_\alpha\,|} \sum_{c\in\mathrm{TP}_\alpha} \mathrm{IoU}(c). \tag{B4}$$

For every true-positive match c, $\mathrm{IoU}(c)$ is the intersection over union between the predicted and ground-truth masks. LocA averages these overlaps only over successfully matched detections, thereby isolating spatial mask precision from missed detections and identity association. Higher LocA indicates more accurately localized instance masks.

Detection accuracy (DetA).

$$\mathrm{DetA}_\alpha = \frac{|\,\mathrm{TP}_\alpha\,|}{|\,\mathrm{TP}_\alpha\,| + \frac{1}{2}(|\,\mathrm{FP}_\alpha\,| + |\,\mathrm{FN}_\alpha\,|)}. \tag{B5}$$

DetA rewards matched detections and symmetrically penalizes false positives and false negatives with half weight. It does not use predicted identity labels, so it measures whether objects are detected consistently rather than whether their identities remain correct. A higher value corresponds to a better balance between detection precision and recall.

Association accuracy (AssA). For a matched detection c, let $\mathrm{TPA}(c)$ be the number of correct trajectory-level associations and let $\mathrm{FPA}(c)$ and $\mathrm{FNA}(c)$ be the incorrect and missed associations. We define

$$A(c) = \frac{\mathrm{TPA}(c)}{\mathrm{TPA}(c) + \mathrm{FPA}(c) + \mathrm{FNA}(c)}, \qquad \mathrm{AssA}_\alpha = \frac{1}{|\,\mathrm{TP}_\alpha\,|} \sum_{c \in \mathrm{TP}_\alpha} A(c). \tag{B6}$$

Thus, AssA measures how consistently each matched observation is connected to the correct trajectory over time. False merges increase FPA, while fragmented or missing links increase FNA. Higher AssA indicates more persistent associations.

Higher-order tracking accuracy (HOTA).

$$\mathrm{HOTA} = \frac{1}{|\mathcal{T}|} \sum_{\alpha \in \mathcal{T}} \sqrt{\mathrm{DetA}_\alpha\, \mathrm{AssA}_\alpha}, \qquad \mathcal{T} = \{0.05, 0.10, \ldots, 0.95\}. \tag{B7}$$

At each IoU threshold, HOTA takes the geometric mean of detection and association accuracy and then averages over thresholds. The geometric mean prevents a method from obtaining a high score by excelling at detection while failing at association, or vice versa. We therefore use HOTA as the primary tracking metric.

3D geometry notation. For an aligned scene, let P be the aggregated predicted object points and G the sampled ground-truth object-surface points. The nearest-neighbor distance from point $\mathbf{x}$ to set S is $d(\mathbf{x}, S) = \min_{\mathbf{y} \in S} \|\mathbf{x} - \mathbf{y}\|_2$.

Accuracy.

$$\mathrm{Accuracy}(P, G) = \frac{1}{|P|} \sum_{\mathbf{p} \in P} d(\mathbf{p}, G). \tag{B8}$$

Accuracy measures how far reconstructed points lie from the closest ground-truth surface. It primarily penalizes spurious or geometrically misplaced reconstruction. We report the distance in centimeters, and lower is better.

Completeness.

$$\mathrm{Completeness}(P, G) = \frac{1}{|G|} \sum_{\mathbf{g} \in G} d(\mathbf{g}, P). \tag{B9}$$

Completeness reverses the nearest-neighbor direction and measures how much of the ground-truth object surface is covered by the reconstruction. Missing object regions increase this value. It is also reported in centimeters, and lower is better.

Chamfer distance.

$$\mathrm{Chamfer}(P, G) = \tfrac{1}{2}\big(\mathrm{Accuracy}(P, G) + \mathrm{Completeness}(P, G)\big). \tag{B10}$$

Chamfer distance symmetrizes the two one-way surface distances. It penalizes both unsupported predicted points and uncovered ground-truth regions, so a low value requires the reconstruction to be simultaneously accurate and complete.

F-score at 5 cm. For distance threshold τ, geometric precision and recall are

$$\mathrm{Prec}_\tau = \frac{1}{|P|}\sum_{\mathbf{p}\in P}\mathbf{1}[d(\mathbf{p},G)\le\tau], \qquad \mathrm{Rec}_\tau = \frac{1}{|G|}\sum_{\mathbf{g}\in G}\mathbf{1}[d(\mathbf{g},P)\le\tau]. \tag{B11}$$

Precision is the fraction of reconstructed points supported by the ground truth, whereas recall is the fraction of ground-truth points recovered by the prediction. Their harmonic mean is

$$\mathrm{F}_\tau = \frac{2\,\mathrm{Prec}_\tau\,\mathrm{Rec}_\tau}{\mathrm{Prec}_\tau + \mathrm{Rec}_\tau}. \tag{B12}$$

Table 2 uses $\tau = 5$ cm and reports $100\,\mathrm{F}_\tau$. A high F-score requires both accurate predicted geometry and broad ground-truth coverage.

3D intersection over union. Let V_P and V_G be the occupied voxels of the predicted and ground-truth instance geometry under the same voxelization. Then

$$\mathrm{IoU}_{3D} = \frac{|V_P \cap V_G|}{|V_P \cup V_G|}. \tag{B13}$$

This metric measures volumetric agreement and penalizes both missing and extraneous occupied space. Higher is better. Although the compact main table focuses on point-based geometry metrics, we use the same definition whenever 3D IoU is discussed in the text.

Functional-orientation error. For instance i, let $\mathbf{f}_i$ and $\mathbf{f}_i^{\mathrm{gt}}$ be the predicted and ground-truth front directions. Both directions are projected onto the plane orthogonal to world up and normalized, yielding $\tilde{\mathbf{f}}_i$ and $\tilde{\mathbf{f}}_i^{\mathrm{gt}}$. The angular error is

$$e_i = \frac{180}{\pi}\arccos\Big(\mathrm{clip}\Big(\tilde{\mathbf{f}}_i^{\top}\tilde{\mathbf{f}}_i^{\mathrm{gt}}, -1, 1\Big)\Big). \tag{B14}$$

The clipping operation prevents numerical round-off from producing an invalid inverse-cosine argument. Projecting onto the gravity-orthogonal plane makes the metric evaluate functional heading rather than reconstruction-induced pitch or roll.

Median angular error.

$$\mathrm{MedErr} = \mathrm{median}_{i=1}^{N}\, e_i. \tag{B15}$$

The median is robust to a small number of severely occluded or ambiguous instances. A lower value indicates more accurate functional-front estimates. The Mean row of Table 3 is the arithmetic macro-average of the five reported category-level values.

Angular-threshold accuracy.

$$\mathrm{Acc@}\,\theta = \frac{100}{N}\sum_{i=1}^{N}\mathbf{1}[e_i\le\theta], \qquad \theta\in\{15^\circ, 30^\circ\}. \tag{B16}$$

Acc@15° measures strict orientation correctness, while Acc@30° allows moderate angular deviation. Both are percentages for which higher is better. As with the median error, the table's Mean row macro-averages the five category-level accuracies.

B.3 Cross-Frame Instance Association Visualization

We select representative frames from two data scenes to qualitatively compare cross-frame instance association results, as shown in Fig.4. The first and third columns show the ground-truth instance annotations, while the second and fourth columns show the model predictions. As can be seen, under viewpoint changes, partial occlusions, and the simultaneous presence of similar instances, the predicted results are able to maintain consistent identities for the same objects across frames and largely agree with the ground truth in terms of instance separation and identity correspondence, demonstrating the stable cross-frame instance association capability of our method.

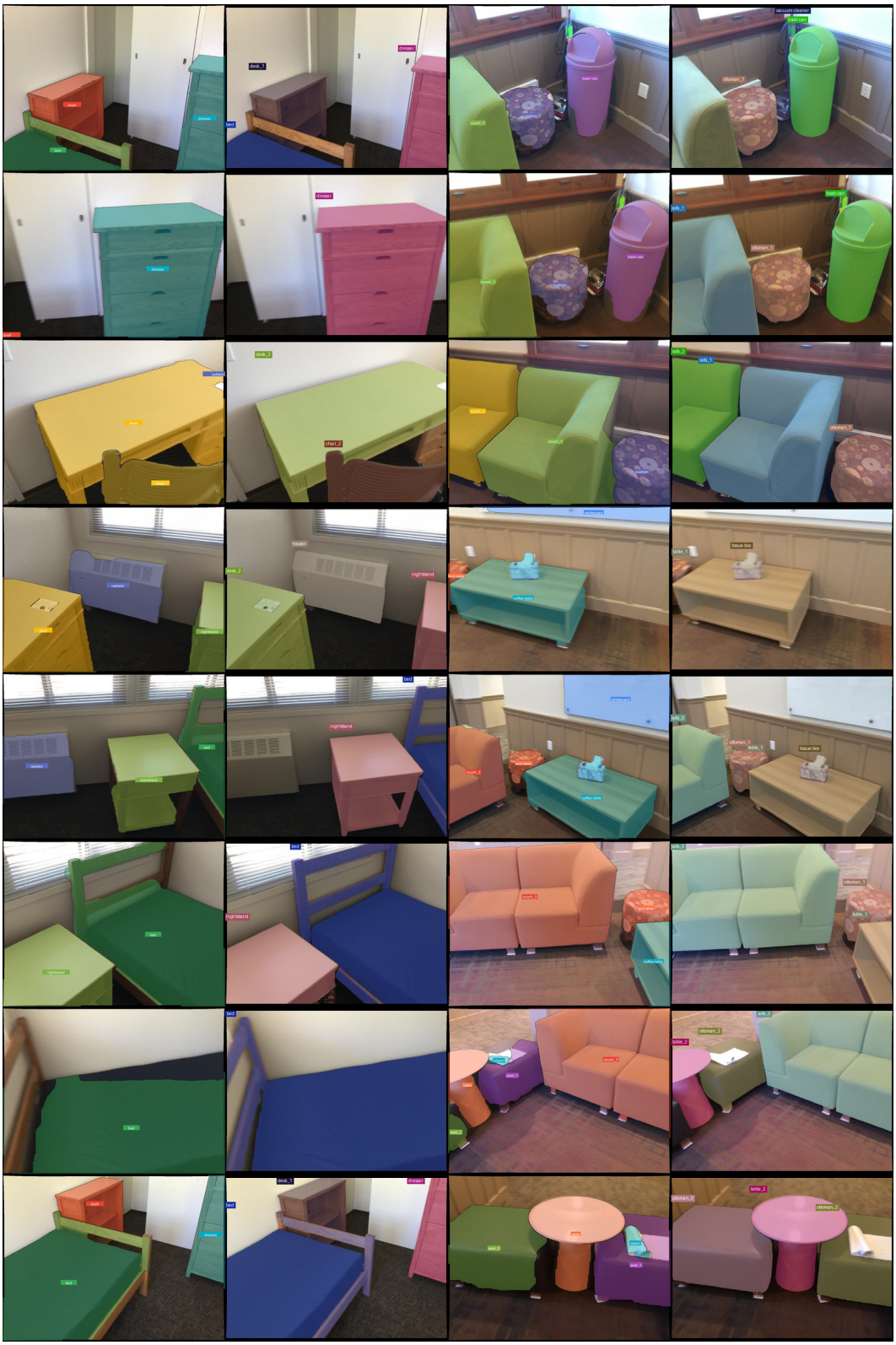


Figure 4: Qualitative visualization of cross-frame instance association across challenging temporal observations.