SGOP3D: A SEMANTIC–GEOMETRIC ANCHORING AND OBSERVATION-PRESERVING RECONSTRUCTION ALGORITHM FOR MULTI-OBJECT ROBOTIC MANIPULATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-object 3D reconstruction for robotic manipulation requires consistent object identity across target admission, measured-surface fusion, and missing-geometry recovery, together with explicit rules governing the use of geometry from different sources in manipulation. We propose SGOP3D (Semantic–Geometric Anchoring and Observation-Preserving 3D Reconstruction), a semantically anchored algorithm for observation-preserving 3D reconstruction from two orthogonal views. The method combines our semantic and geometric detection networks, using class consistency, bidirectional top–front reprojection, and world-coordinate constraints to establish trusted object-level anchors. The same anchor governs measured-surface extraction and completion with our GenCompletionNet, converting complete-shape predictions into missing-geometry residuals that satisfy surface-protection and dual-view free-space constraints. Per-point provenance labels restrict contact localization to measured-source points and limit generated geometry to shape estimation and occupancy refinement. In a system-level comparison on 205 synchronized top–front observation pairs involving 20 objects arranged in two groups of 10, SGOP3D exceeds AMB3R by 36.45 and 21.10 percentage points in strict instance separation success and F@$10\,\mathrm{mm}$, respectively, under a shared scoring protocol. Across 400 controlled grasp–lift–hold trials with fixed detection, planning, and control, SGOP3D achieves 96% success versus 79% for the strongest geometric-input control. A separate continuous sorting demonstration with dynamic replenishment reports 19/20 first-attempt completions and 20/21 transfer-and-release completions. These results show that SGOP3D maintains object-consistency and observational-evidence constraints throughout geometric recovery and manipulation, providing object-level 3D geometry with traceable provenance for contact planning, grasping, and continuous sorting in multi-object robotic scenes.

Keywords: multi-object 3D reconstruction; cross-view instance association; semantic–geometric consensus; observation-preserving reconstruction; residual point cloud completion; robotic manipulation.

1 INTRODUCTION

In multi-object robotic scenes, language-specified targets must remain instance-consistent across localization, reconstruction, and contact planning. Top and front RGB-D views provide complementary surfaces, but similarity, proximity, and occlusion can cause cross-object fusion and contaminate completion inputs. **Instance admission must therefore precede fusion** and continue to constrain recovery.

Vision–language models support language-conditioned perception (Alayrac et al., 2022; Li et al., 2023; Bai et al., 2025), aided by LoRA (Hu et al., 2022), but require calibrated geometry to verify instance correspondence. Completion methods generally assume single-object inputs (Yuan et al., 2018; Yu et al., 2021; Xiang et al., 2021; Zhou et al., 2022); PoinTr also preserves observations

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.

through residual prediction and concatenation (Yu et al., 2021). SGOP3D couples instance admission, observation-preserving recovery, and source-dependent geometric use under a shared trusted anchor.

The anchor constrains measured-surface extraction, completion inputs, and residual acceptance under surface-protection and free-space constraints. Measured points support contact localization; generated geometry supports shape estimation and occupancy refinement.

In system-level comparisons under shared scoring, SGOP3D improves strict separation and F@10 mm over AMB3R by 36.45 and 21.10 percentage points. Under fixed detection, planning, and control, it achieves 96% grasp–lift–hold success, 17 percentage points above the strongest geometric-input control.

Our main contributions are as follows:

1. **Cross-view semantic–geometric instance anchoring.** We design semantic and geometric detectors whose complementary candidates undergo one-to-one association and admission through class consistency, bidirectional reprojection, and world-coordinate constraints, yielding anchors for object-level reconstruction.

2. **Anchor-driven observation-preserving residual reconstruction.** We design GenCompletionNet and its training objective for anchor-centered shape prediction. Surface-protection and dual-view free-space constraints extract missing residuals, while provenance labels govern the use of measured, generated, and auxiliary geometry in manipulation.

3. **Evaluation from reconstruction to robotic manipulation.** Experiments include 205 observation pairs of 20 objects, comparisons with six public reconstruction methods, a three-object challenge set, admission and source ablations, 400 comparative grasping trials, and a 20-object continuous sorting demonstration.

2 Related Work

Language-conditioned perception and instance association. Vision–language models provide language-conditioned target understanding (Alayrac et al., 2022; Li et al., 2023; Bai et al., 2025), while LoRA enables efficient task adaptation (Hu et al., 2022). Embodied models further connect multimodal perception with robotic tasks and actions (Driess et al., 2023; Zitkovich et al., 2023). In our setting, language-conditioned candidates require calibrated depth and cross-view geometric evidence to establish reliable object correspondences. SGOP3D combines semantic and geometric candidates within a fixed canonical class dictionary and propagates instance admission decisions to reconstruction.

Multi-view reconstruction and point cloud completion. Classical registration and fusion (Besl & McKay, 1992; Curless & Levoy, 1996; Newcombe et al., 2011; Laurentini, 1994) and recent multi-view methods (Jin et al., 2024; Cabon et al., 2025; Yang et al., 2025; Guo et al., 2025; Wang & Agapito, 2026; Wang et al., 2026) recover consistent scene geometry. Instance-aware registration explicitly addresses interference from other objects (Yu et al., 2024), but model-to-scene registration differs from establishing object correspondences between two robot views. Point cloud completion methods recover missing geometry from partial observations (Yuan et al., 2018; Yu et al., 2021; Xiang et al., 2021; Zhou et al., 2022); subsequent work reduces complete-supervision requirements (Cui et al., 2023; Hong et al., 2023; Wu et al., 2025) or incorporates viewpoint and diffusion priors (Fu et al., 2023; Wei et al., 2025). Such recovery generally assumes that the input belongs to one object. SGOP3D first establishes trusted instance anchors, then uses the same anchors to constrain measured-surface extraction and completion inputs. Generated candidates contribute only residuals that satisfy surface-protection and dual-view free-space constraints.

Geometry for robotic manipulation. GraspNet-1Billion and AnyGrasp support grasp prediction in complex scenes (Fang et al., 2020; 2023), and shape completion can provide occluded geometry for grasp planning (Varley et al., 2017; Mohammadi et al., 2023). Manipulation also requires consistent object identity and reliable contact surfaces. SGOP3D retains per-point provenance to distinguish measured surfaces, generated residuals, and auxiliary boundary points, restricting their uses in contact localization, shape estimation, and occupancy refinement. Expanded discussion is provided in Appendix G.

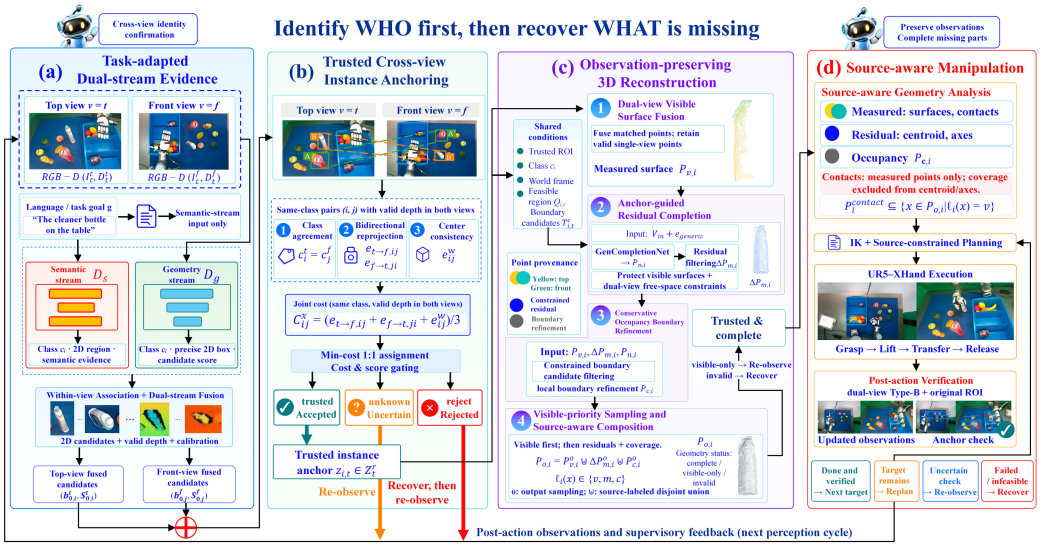

Figure 1: Overview of SGOP3D: (a) task-adapted dual-stream evidence; (b) trusted cross-view instance anchoring; (c) observation-preserving 3D reconstruction; and (d) source-aware robotic manipulation. The anchor propagates admission decisions to measured-surface extraction, observation-preserving residual completion, and source-aware composition. The resulting 3D targets support pose estimation, grasp planning, UR5–XHand execution, and post-action verification.

3 The SGOP3D Algorithm

SGOP3D establishes instance identity before recovering missing geometry: cross-view admission produces a trusted anchor that constrains reconstruction and source-aware manipulation (Figure 1).

For view $v \in \{t, f\}$, let I_t^v, D_t^v, K^v, T_v^w denote the RGB image, depth map, intrinsics, and calibrated camera-to-world transformation. All geometric checks use the actual calibration rather than assuming an ideal 90° rotation.

Training pseudo-targets, online coverage candidates, and evaluation CAD/PLY references serve separate roles (Appendices B.2, C.3, and D.2). Normal execution requires both trusted admission and a complete reconstruction state.

3.1 Task-Adapted Dual-Stream Detection

We design and independently train a semantic detector D_s and a high-resolution geometric detector D_g for language-conditioned discovery and metric localization. Our semantic detector uses Qwen3-VL-8B-Instruct (Bai et al., 2025) as its foundation, with our image–instruction–class–box supervision for Type-A discovery and Type-B presence verification.

Online semantic inference combines full-image and local-crop discovery, candidate rechecking, and Type-B verification. Cross-view queries transfer class hypotheses without copying boxes. Repeated decoding provides instance-support and box-drift estimates for evidence fusion; association and support-counting details are given in Appendix B.1.

Type-B verification is also reused after execution. Valid depth and cross-view consensus jointly determine reconstruction eligibility; training and depth statistics are specified in Appendix B.1.

3.2 Cross-View Instance Consensus and Trusted Anchoring

Within each view, same-class semantic and geometric candidates undergo one-to-one matching with cost $1 - \mathrm{IoU}(b_s, b_g)$; fusion requires $\mathrm{IoU} \geq 0.4$.

Validation-set empirical quantiles place both streams on a common relative-evidence scale. Fusion weights decrease with semantic box drift and geometric depth dispersion, while single-stream candidates receive fixed discounts. Each fused box is backprojected to obtain a coordinate-wise

median 3D center; candidates with insufficient valid depth cannot enter cross-view assignment. Appendix C.2 gives the fusion equations and discounts.

Cross-view consistency combines bidirectional projected-center errors, normalized by box diagonals, with the scale-normalized world-center error e_{ij}^{w}. All errors use calibrated geometry; their definitions are provided in Appendix C.2.

A pair is feasible only when its classes agree and both 3D centers are valid. Its joint cost and fused admission score are

$$C_{ij}^{x} = \begin{cases} \dfrac{e_{t\to f,ij} + e_{f\to t,ji} + e_{ij}^{w}}{3}, & c_i^t = c_j^f,\ r_i^{3D,t} r_j^{3D,f} = 1, \\ +\infty, & \text{otherwise}, \end{cases} \tag{1a}$$

$$\begin{aligned} \mathcal{M}_{cv} &= \{(i,\kappa(i)) \in \mathcal{H}_{cv} \mid C_{i\kappa(i)}^{x} \le \tau_x\}, \\ \acute{S}_{0,i} &= \frac{S_{0,i}^{t} + S_{0,\kappa(i)}^{f}}{2}, \qquad e_{x,i} = C_{i\kappa(i)}^{x}, \\ S_{f,i} &= \mathrm{clip}[(1 - 0.8 e_{x,i})\acute{S}_{0,i}, 0, 1]. \end{aligned} \tag{1b}$$

Here, $\mathcal{H}_{cv}$ is the one-to-one minimum-cost assignment over finite costs, $\kappa(i)$ maps top-view candidates to front-view candidates, and $\tau_x = 0.5$.

Pairs satisfying $C_{i\kappa(i)}^{x} \le \tau_x$ and $S_{f,i} \ge 0.30$ are trusted. Matched pairs exceeding the cost threshold or having $S_{f,i} < 0.15$ are rejected; remaining low-confidence matches and unmatched candidates are unknown. Unknown and rejected candidates enter U_t^{cv} and $\mathcal{R}_t^{cv}$, respectively, and are excluded from normal reconstruction. Unknown triggers another query; rejection triggers renewed perception or safe recovery.

Each anchor stores dual-view regions, class, world-coordinate evidence, fused center, mean scale, score, and admission state:

$$\begin{aligned} b_i^t &= b_{0,i}^t, \qquad b_i^f = b_{0,\kappa(i)}^f, \\ z_{i,t} &= (c_i, b_i^t, b_i^f, A_i^w, p_i^w, s_i^w, S_{f,i}, \eta_i, q_i^r), \\ Z_t &= \{z_{i,t} \mid (i,\kappa(i)) \in \mathcal{H}_{cv}\}. \end{aligned} \tag{2}$$

The anchor regions constrain measured-point extraction and residual acceptance; c_i determines target consistency and completion eligibility, p_i^w defines the spatial reference, and q_i^r controls reconstruction admission.

Joint same-class NMS uses both dual-view IoU and world-center distance, preserving instances with overlapping projections but separated depths.

Appendix C.2 and Algorithm C1 provide the complete fusion, suppression, and admission rules.

3.3 Observation-Preserving Residual Reconstruction

For each trusted anchor, valid depth samples within the predicted top and front ROIs are backprojected into world coordinates. A bounded residual translation is applied only when registration-quality criteria are met. Mutual nearest-neighbor pairs are averaged, while all unmatched points from these ROI point sets are retained, yielding the sensor-derived measured component P_v (Appendix C.3). This component represents the fused measurements extracted from the admitted ROIs.

Reconstruction states are independent of admission: complete satisfies all output constraints, visible-only retains measured geometry without a complete output, and invalid denotes absent geometry or failed constraints.

The dual-view ROI frustum intersection, after removing observed free space, defines the feasible set $Q_{i,t}$. Generated points must lie within this set; auxiliary candidates $T_{i,t}^{c}$ come from its feasible-volume boundary and are not measured surfaces. Depth intervals, voxelization, ray-based filtering, and degenerate cases are specified in Appendix C.3.

Our GenCompletionNet predicts a complete-shape candidate from normalized visible points in the anchor-centered frame, followed by transformation back to world coordinates. **The frozen test configuration maps all conditioning indices to a generic embedding**, using a shared shape prior. Architecture, normalization, and training supervision are given in Appendix B.2.

Observation preservation retains measured-surface priority and provenance while adding constrained missing geometry. Let P_v, P_n, and P_c denote measured surfaces, complete-shape predictions, and auxiliary occupancy-boundary candidates, respectively. Network residuals and auxiliary points are selected as

$$\begin{aligned} \Delta P_{m,i} &= \{x \in P_{n,i} \cap Q_{i,t} \mid d(x, P_{v,i}) > \tau_v\}, \\ R_i &= P_{v,i} \cup \Delta P_{m,i}, \qquad \tau_v = 3\,\mathrm{mm}, \end{aligned} \tag{3a}$$

$$\begin{aligned} P_{c,i} = \{x \in T^c_{i,t} \cap Q_{i,t} \mid\ & d(x, P_{v,i}) > \tau_v, \\ & d(x, \Delta P_{m,i}) > \tau_c,\ d(x, P_{n,i}) \le \tau_o\}, \\ \tau_c = 2\,\mathrm{mm}, \qquad \tau_o &= 6\,\mathrm{mm}. \end{aligned} \tag{3b}$$

Sampling first satisfies the measured-surface coverage bound, then adds available residuals and coverage points within the remaining budget. Padding contributes no geometric evidence. If the coverage bound cannot be met, the output falls back to visible-only and triggers renewed perception.

$$\begin{aligned} P_{o,i} &= P^o_{v,i} \uplus \Delta P^o_{m,i} \uplus P^o_{c,i}, \\ \ell_i(x) &\in \{v, m, c\}, \qquad |P_{o,i}| = N^u_i, \end{aligned} \tag{4a}$$

$$\begin{aligned} h_v(n^*_v) &\le \tau_v, \qquad \forall x \in P_{o,i} : \\ \ell_i(x) &\ne v \ \Rightarrow\ [d(x, P_{v,i}) > \tau_v \ \wedge\ x \in Q_{i,t}], \end{aligned} \tag{4b}$$

$$P^{contact}_i \subseteq \{x \in P_{o,i} \mid \ell_i(x) = v\}. \tag{4c}$$

Here, $\uplus$ denotes a union with provenance labels, and $h_v(n)$ is the maximum nearest-neighbor distance from the fused visible surface to a subset of n unique measured points; n^*_v is the selected count.

Only measured-source points are eligible for contact localization and local-surface estimation. Equation (4c) expresses this source restriction; it does not by itself establish target-instance membership or contact reliability. Correct target contact also depends on the extracted geometry and downstream feasibility checks. Network residuals additionally support centroid, principal-axis, and occluded-boundary estimation; auxiliary coverage points contribute only to occupied extent and collision envelopes.

Appendix C.3 specifies source quotas, deduplication, valid-point masks, and fallback handling.

3.4 Source-Constrained Robotic Closed Loop

Normal execution requires a trusted anchor and complete reconstruction. Feasible grasps use measured-source contacts and satisfy inverse-kinematics and full-path collision constraints; synchronized post-action verification determines whether to advance, replan, reobserve, or recover. Appendix C.4 and Algorithms C2–C3 give the complete procedure.

4 Experiments

4.1 Experimental Setup and Statistical Conventions

This section evaluates overall reconstruction, admission and source composition, and gains in robotic manipulation, in that order. Appendix D.1 provides the full experimental scale and statistical units.

The completion training, validation, and test sets contain 31, 10, and 20 objects and 3100, 1000, and 600 object-level samples, respectively. The splits are disjoint in physical object identity, acquisition sequence, and file hash (Appendix A.1).

Throughout this paper, unseen objects refer to physical objects and their acquisition sequences that were not used for completion training.

Ground-truth instance masks, CAD poses, and object IDs are used only for offline scoring. Strict separation requires Recall@10 mm of at least 50% and cross-instance contamination of at most 10%. All methods process 205/205 observation pairs and share one fixed set of 1808 CAD-based geometric records. SGOP3D provides outputs for 1700 records (94.03%) and has 108 missing predictions (5.97%). All 1808 records enter separation, contamination, and F@10 mm; missing predictions count as separation failures. CD-L1 and Completeness are conditional means over each method's records with outputs. The scoring set is linked to the separate class-query evaluation by scene and target class; 54 target-present queries lack a frozen reference pose because their saved annotated visible-point sets are empty (Appendix D.3).

Reference poses are fixed using measured visible points and are not re-registered to predictions. The three-object challenge-set configuration, complete stratification, and qualitative analysis are provided in Appendix E.2.

4.2 Object-Level Reconstruction in Ten-Object Scenes

We compare DUSt3R (Wang et al., 2024), MUSt3R (Cabon et al., 2025), Fast3R (Yang et al., 2025), AMB3R (Wang & Agapito, 2026), VGGT (Wang et al., 2025), and MVSplat (Chen et al., 2024) on shared synchronized top–front observations with at most two views. Baselines use official released weights; SGOP3D uses our trained semantic, geometric, and completion networks with frozen test-time weights. All pipelines share frozen target records, references and poses, the 8192-point evaluation sampling budget, and metric and missing-prediction rules. This comparison measures complete frozen pipelines with method-specific sensor use, adaptation, and architectures; no test-time weight updates are allowed.

Table 1: Overall dual-view reconstruction in ten-object scenes: 20 objects in two groups and 205 observation pairs.

Method	Separation (%)↑	Contam. (%)↓	CD-L1 (mm)↓	Complete. (mm)↓	F@10 mm (%)↑
DUSt3R	0.94	22.09	70.70	40.69	19.03
MUSt3R	0.50	28.18	112.95	62.97	13.25
Fast3R	0.44	37.70	125.65	63.42	11.64
AMB3R	13.55	**7.90**	30.38	19.20	42.40
VGGT	6.03	16.19	52.81	31.63	26.11
MVSplat	5.25	34.24	320.02	48.98	22.21
SGOP3D (Ours)	**50.00**	13.04	**20.71**	**7.83**	**63.50**

Note. Separation: instance separation success; Contam.: cross-instance contamination; Complete.: Completeness. Bold indicates the best value in each metric column.

Overall results. SGOP3D achieves **50.00%** strict separation, exceeding AMB3R's 13.55% by **36.45 percentage points**, and F@10 mm of 63.50% (Table 1). Conditional CD-L1 and Completeness are 20.71 mm and 7.83 mm. Cross-instance contamination is 13.04%, higher than AMB3R's 7.90%; this full-output metric does not separately measure measured-source contamination or contact errors. Appendix E.1 visualizes these results.

Figure 2 illustrates the recovered object contours and layouts under partial occlusion and overlap. Table 1 establishes complete-pipeline performance. Section 4.3 links semantic candidates and consensus admission to coverage and screening (Table 2), and tests measured-point retention and auxiliary coverage with fixed inputs and generator weights (Table 3). The controlled grasping comparison in Appendix F.3 evaluates the complete geometric representation with fixed perception, planning, and control.

4.3 Analysis of Key Mechanisms

4.3.1 Semantic Candidate Augmentation and Cross-View Consensus Admission

We compare geometry only (D_g), semantics only (D_s), dual stream (D_g+D_s), and dual stream with consensus on the same 205 observation pairs and 2050 queries. RGB-D inputs, queries, and presence/absence annotations are shared. Semantics only excludes geometric candidates; dual stream omits consensus gating. Acceptance requires a valid candidate satisfying the configuration's

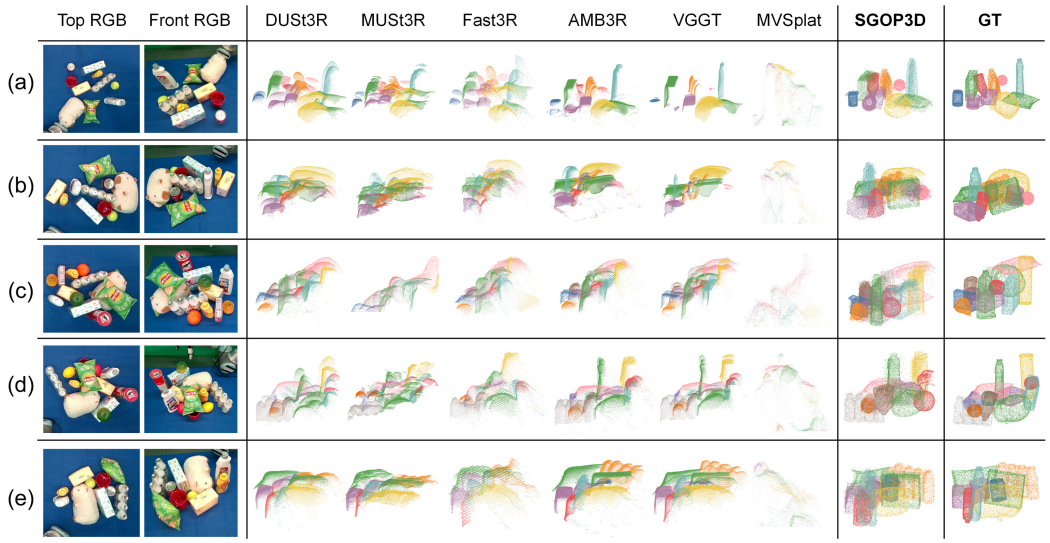

Figure 2: Object-level reconstruction in ten-object scenes with partial occlusion and overlap. The 20 objects are divided into two groups of 10, and five pairs of views are shown. Each row presents the RGB components of the top and front RGB-D observations, the six public methods, SGOP3D, and the reference point cloud (GT), in that order. Colors distinguish objects. Quantitative results use the per-object 8192-point evaluation protocol in Appendix D.2.

admission rules; no candidate, rejection, and UNKNOWN count as non-acceptance. Table 2(b) measures query admission; Table 1 measures geometry on 1808 CAD-based records. Appendix D.3 defines both statistical units and the scope of cross-metric interpretation.

Table 2: Class-query admission: example query states and results for four configurations.

(a) Ten example queries corresponding to Figure 3

Configuration	A+	B–	C+	D–	E–	F–	G–	H+	I+	J–
Geometry only	ACC	REJ*	ACC	REJ*	FA	REJ*	REJ*	ACC	ACC	REJ*
Semantics only	ACC	FA	ACC	REJ*	FA	REJ*	FA	ACC	ACC	REJ*
Dual stream	ACC	FA	ACC	REJ*	FA	REJ*	FA	ACC	ACC	REJ*
Dual + consensus	ACC	UNK	ACC	REJ*	UNK	REJ*	UNK	ACC	ACC	REJ*

(b) Full evaluation: 205 observation pairs and 2050 queries

Configuration	**Absent-target non-acceptance↑**	**False acceptance↓**	**Present-target acceptance↑**	**Unknown rate**
Geometry only	95.74 (180/188)	4.26 (8/188)	91.41 (1702/1862)	0.00 (0/2050)
Semantics only	69.68 (131/188)	30.32 (57/188)	98.07 (1826/1862)	0.00 (0/2050)
Dual stream	67.02 (126/188)	32.98 (62/188)	**98.34** (1831/1862)	0.00 (0/2050)
Dual + consensus	**97.34** (183/188)	**2.66** (5/188)	89.80 (1672/1862)	10.54 (216/2050)

Notes. Geometry only, semantics only, and dual stream denote D_g, D_s, and D_g+D_s, respectively; Dual + consensus additionally enables cross-view consensus gating. In (a), +/– indicate target presence/absence; ACC, FA, and UNK denote acceptance, false acceptance, and unknown. REJ* retains the original REJECT label without distinguishing no-candidate cases from explicit rejection. In (b), percentages are followed by counts: absent-target rates use 188 queries, present-target acceptance uses 1862, and UNKNOWN uses all 2050. Acceptance denotes admission, not successful reconstruction. Bold marks the best acceptance or non-acceptance rate in each column; the unknown rate has no preferred direction.

Semantic candidate augmentation increases accepted present-target queries from 1702 to 1831 out of 1862: 129 additional acceptances and **6.93 percentage points** of pre-gating coverage. False acceptances also increase from 8 to 62 out of 188 absent-target queries. Consensus then reduces false acceptances to 5 (**2.66%**), while present-target acceptance falls to 1672/1862 (89.80%) and 216/2050 queries (10.54%) become UNKNOWN. Thus, semantics expands candidate coverage and consensus imposes conservative admission. Relative to geometry only, the complete configuration accepts three fewer absent targets and 30 fewer present targets. These results quantify the coverage–screening trade-off at the tested operating points; Appendix D.4 provides the corresponding plots.

Admission of cross-view class hypotheses. Figure 3 and Table 2(a) show how the current example changes from candidate generation to admission decisions. The incorrect class hypotheses B, E, and

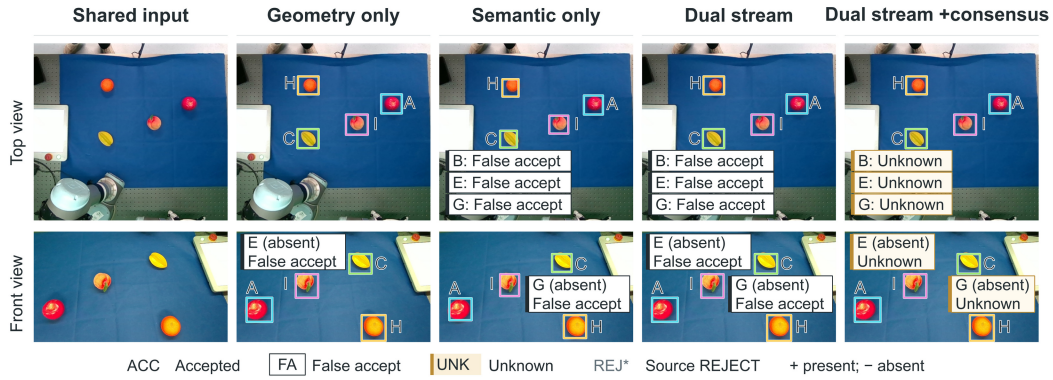

Figure 3: Candidates and admission outcomes from shared dual-view inputs. The leftmost column shows the shared inputs; the next four columns show geometry only, semantics only, dual-stream detection, and dual-stream detection with consensus gating. The top and bottom rows correspond to the top and front views. The 10 example queries and full statistics are provided in Tables 2(a) and 2(b), respectively. UNKNOWN denotes insufficient evidence.

G produced by dual-stream detection become UNKNOWN after consensus is added, while A, C, H, and I remain accepted. This is consistent with the reduction in false acceptance and the explicit uncertainty outputs reported in Table 2(b).

4.3.2 Measured-Surface Preservation and Missing-Region Coverage

The Q2 2×2 ablation (Table D1) switches measured-point retention and auxiliary coverage while fixing input clouds, reference poses, generator weights, and the output budget. The 8192-point capacity and evaluation budget do not imply equal numbers of distinct points: the network-only prediction contains 2048 native points, and padding duplicates add no geometric evidence.

Table 3: Ablation of measured-point retention and auxiliary point addition under a unified output budget. Distance metrics are in mm, and F@10 mm is in %. Bold indicates the best result for each metric.

Configuration	Measured retention	Auxiliary points	CD-L1↓	Accuracy↓	Completeness↓	F@10 mm↑
A: Network only	×	×	16.959	7.556	9.403	73.13
B: No auxiliary	✓	×	13.498	**5.101**	8.397	80.03
C: No measured points	×	✓	13.870	7.811	6.058	81.08
D: SGOP3D	✓	✓	**10.747**	5.260	**5.486**	**88.07**

Note: ✓ denotes enabled and × disabled. All four configurations use the same fixed weights from a single training run; the results reflect differences in source composition under this frozen configuration. Configuration A is the offline network-only prediction and is reported separately from the hybrid network–template completion configuration in Appendix F.3.

Source-composition effects. Measured-point retention improves Accuracy and CD-L1; auxiliary coverage improves Completeness and F@10 mm at a slight Accuracy cost. The complete composition has the lowest CD-L1 (Table 3; Appendix D.4). These results establish complementary fitting and coverage benefits under frozen weights. Attribution concerns the tested source switches; surface-protection, free-space, and contact-use rules are evaluated jointly in the complete pipeline.

4.4 Real-World Robotic Manipulation

Controlled grasping. With detection, planning, control, calibration, and per-object grasp offsets fixed, we compare measured points (C1), measured points with template/rule-based completion (C2), hybrid network–template completion (C3), and complete SGOP3D. Each configuration has 100 trials across 20 objects. Success requires grasping, a 0.05 m lift, and a 5 s hold. Success rates are 76%, 69%, 79%, and 96%, respectively: SGOP3D improves over the strongest control by 17 percentage points (Table F2). This evaluates the complete geometric-input configuration, including its completion backend and geometry-availability/fallback behavior. Appendix F.1 provides the physical outcome criteria and failure breakdown.

4.4.1 Continuous Multi-Target Sorting with Dynamic Replenishment

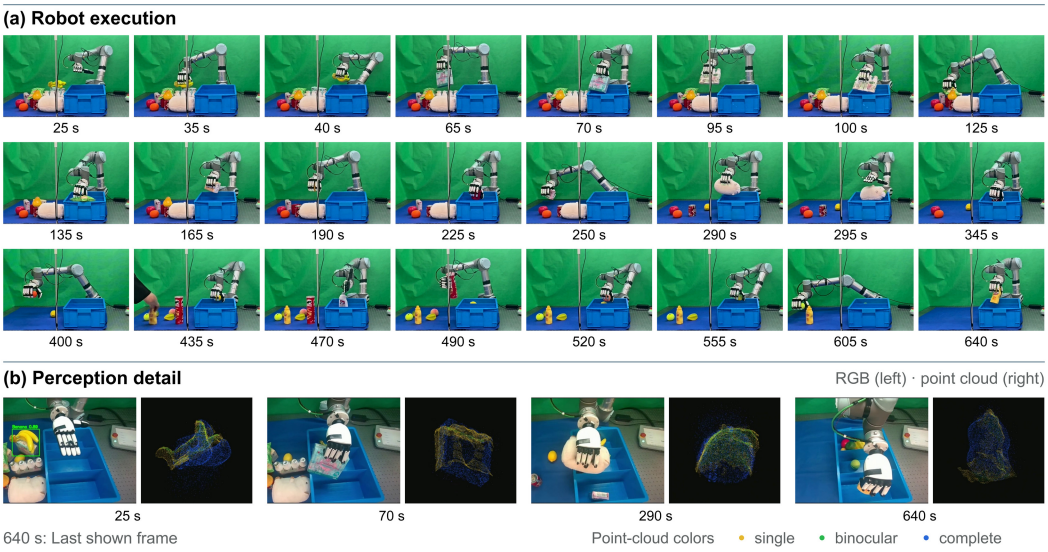

Figure 4: Continuous sorting with replenishment. (a) Twenty-four chronological keyframes; the asterisk marks replenishment at 435 s, and the last frame is at 640 s. (b) RGB observations and point clouds at 25, 70, 290, and 640 s. Yellow/green/blue indicate single-view/dual-view/completed clouds. Statistics: Table 4; events: Appendix F.1.

Using complete SGOP3D, we extend the grasp–lift–hold protocol in Appendix F.3 to continuous grasp–transfer–release (Figure 4). The experimenter replenishes and rearranges objects; the system autonomously performs perception, planning, and execution. Completion requires grasping, transfer to the collection bin, and release. Table 4 reports first-attempt completion per object and transfer-and-release completion per attempt.

Table 4: Completion of continuous sorting with dynamic replenishment. First-attempt completion is measured per object, and transfer-and-release completion per grasping attempt.

Batch	N	First attempt	Transfer / release
Initial batch	13	12/13(92.31%)	13/14(92.86%)
Replenishment batch	7	7/7(100.00%)	7/7(100.00%)
Total	**20**	**19/20(95.00%)**	**20/21(95.24%)**

Note: N denotes the number of objects. Both batches belong to the same sorting sequence. The seven objects in the replenishment batch are introduced during execution.

The milk can falls after its first grasp; following one manual pose-restoration intervention, the retry succeeds. Both attempts are retained. The controlled grasping comparison supports the manipulation value of the complete geometric representation, while continuous sorting demonstrates closed-loop operation with dynamic replenishment.

5 Conclusion

SGOP3D couples semantic–geometric anchors with measured-surface preservation, constrained residual completion, and provenance-based geometry use. Admission and source ablations support complementary coverage, screening, and surface-fitting mechanisms. The complete configuration improves strict separation and F@10 mm, achieves 96% grasp–lift–hold success under fixed perception, planning, and control, and supports continuous sorting with dynamic replenishment.

The current implementation is tailored to a predefined top–front observation configuration. Future work will explore adaptive viewpoint selection to acquire additional measured evidence while retaining object-level anchoring, measured-surface preservation, and provenance constraints. We will also investigate task-aware reconstruction updates to reduce redundant computation during continuous sorting.

AI USE STATEMENT

Generative AI tools were used to polish the language of the manuscript.

REPRODUCIBILITY STATEMENT

Appendix A describes the object-disjoint splits and instance records. Appendix B specifies the model architecture, training supervision, normalization, and training configuration. Appendix C gives the complete inference algorithms and robotic state transitions. Appendix D defines the input protocol, evaluation references, metrics, and statistical units; Appendices E and F provide additional reconstruction and robotic results. These sections specify frozen configurations, provenance handling, and the treatment of missing predictions used in the reported experiments.

REFERENCES

J.-B. Alayrac, J. Donahue, P. Luc, et al. Flamingo: A Visual Language Model for Few-Shot Learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

S. Bai, Y. Cai, R. Chen, et al. Qwen3-VL Technical Report. *arXiv preprint arXiv:2511.21631*, 2025.

P. J. Besl and N. D. McKay. A Method for Registration of 3-D Shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):239–256, 1992.

Y. Cabon, L. Stoffl, L. Antsfeld, et al. MUSt3R: Multi-view Network for Stereo 3D Reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1050–1060, 2025.

Y. Chen, H. Xu, C. Zheng, et al. MVSplat: Efficient 3D Gaussian Splatting from Sparse Multi-View Images. In *Proceedings of the European Conference on Computer Vision*, pages 370–386, 2024. doi: 10.1007/978-3-031-72664-4_21.

R. Cui, S. Qiu, S. Anwar, et al. P2C: Self-Supervised Point Cloud Completion from Single Partial Clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14351–14360, 2023.

B. Curless and M. Levoy. A Volumetric Method for Building Complex Models from Range Images. In *Proceedings of SIGGRAPH*, pages 303–312, 1996.

D. Driess, F. Xia, M. S. M. Sajjadi, et al. PaLM-E: An Embodied Multimodal Language Model. In *Proceedings of the 40th International Conference on Machine Learning*, PMLR 202:8469–8488, 2023.

H.-S. Fang, C. Wang, M. Gou, et al. GraspNet-1Billion: A Large-Scale Benchmark for General Object Grasping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11444–11453, 2020.

H.-S. Fang, C. Wang, H. Fang, et al. AnyGrasp: Robust and Efficient Grasp Perception in Spatial and Temporal Domains. *IEEE Transactions on Robotics*, 39(5):3929–3945, 2023. doi: 10.1109/TRO.2023.3281153.

Z. Fu, L. Wang, L. Xu, et al. VAPCNet: Viewpoint-Aware 3D Point Cloud Completion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12108–12118, 2023.

H. Guo, H. Zhu, S. Peng, et al. Multi-view Reconstruction via SfM-guided Monocular Depth Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5272–5282, 2025.

S. Hong, M. Yavartanoo, R. Neshatavar, et al. ACL-SPC: Adaptive Closed-Loop System for Self-Supervised Point Cloud Completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9435–9444, 2023.

E. J. Hu, Y. Shen, P. Wallis, et al. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*, 2022.

S. Jin, I. Armeni, M. Pollefeys, et al. Multiway Point Cloud Mosaicking with Diffusion and Global Optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20838–20849, 2024.

A. Laurentini. The Visual Hull Concept for Silhouette-Based Image Understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(2):150–162, 1994.

J. Li, D. Li, S. Savarese, et al. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *Proceedings of the 40th International Conference on Machine Learning*, PMLR 202:19730–19742, 2023.

I. Loshchilov and F. Hutter. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*, 2019.

S. S. Mohammadi, N. F. Duarte, D. Dimou, et al. 3DSGrasp: 3D Shape-Completion for Robotic Grasp. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 3815–3822, 2023. doi: 10.1109/ICRA48891.2023.10160350.

R. A. Newcombe, S. Izadi, O. Hilliges, et al. KinectFusion: Real-Time Dense Surface Mapping and Tracking. In *Proceedings of the IEEE International Symposium on Mixed and Augmented Reality*, pages 127–136, 2011.

J. Varley, C. DeChant, A. Richardson, et al. Shape Completion Enabled Robotic Grasping. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2442–2447, 2017.

H. Wang and L. Agapito. AMB3R: Accurate Feed-forward Metric-scale 3D Reconstruction with Backend. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14612–14625, 2026.

J. Wang, M. Chen, N. Karaev, et al. VGGT: Visual Geometry Grounded Transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5294–5306, 2025.

J. Wang, M. Chen, S. Zhang, et al. VGGT-Ω. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21486–21499, 2026.

S. Wang, V. Leroy, Y. Cabon, et al. DUSt3R: Geometric 3D Vision Made Easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20709, 2024.

G. Wei, Y. Feng, L. Ma, et al. PCDreamer: Point Cloud Completion Through Multi-view Diffusion Priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27243–27253, 2025.

E. B. Wilson. Probable Inference, the Law of Succession, and Statistical Inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927. doi: 10.1080/01621459.1927.10502953.

L. Wu, X. Cheng, Y. Xu, et al. Unsupervised 3D Point Cloud Completion via Multi-View Adversarial Learning. *IEEE Transactions on Visualization and Computer Graphics*, 31(10):7890–7905, 2025. doi: 10.1109/TVCG.2025.3559340.

P. Xiang, X. Wen, Y.-S. Liu, et al. SnowflakeNet: Point Cloud Completion by Snowflake Point Deconvolution with Skip-Transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5499–5509, 2021.

J. Yang, A. Sax, K. J. Liang, et al. Fast3R: Towards 3D Reconstruction of 1000+ Images in One Forward Pass. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21924–21935, 2025.

X. Yu, Y. Rao, Z. Wang, et al. PoinTr: Diverse Point Cloud Completion with Geometry-Aware Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12498–12507, 2021.

Z. Yu, Z. Qin, L. Zheng, et al. Learning Instance-Aware Correspondences for Robust Multi-Instance Point Cloud Registration in Cluttered Scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19605–19614, 2024.

W. Yuan, T. Khot, D. Held, et al. PCN: Point Completion Network. In *Proceedings of the International Conference on 3D Vision*, pages 728–737, 2018.

H. Zhou, Y. Cao, W. Chu, et al. SeedFormer: Patch Seeds Based Point Cloud Completion with Upsample Transformer. In *Proceedings of the European Conference on Computer Vision*, pages 416–432, 2022.

B. Zitkovich, T. Yu, S. Xu, et al. RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control. In *Proceedings of the Conference on Robot Learning*, PMLR 229:2165–2183, 2023.