GaussAsset: Evidence-Grounded Segmentation into Auditable Object Assets in Indoor 3D Gaussian Scenes

Anonymous authors
Paper under double-blind review

Abstract

3D Gaussian Splatting provides an explicit rendering representation for indoor scenes, but turning that representation into objects remains difficult: real captures contain floaters and missing coverage, and exhibit view-dependent effects such as transparency and reflections. Methods built for clean point clouds or surface annotations do not account for these Gaussian-specific artifacts and do not decide which Gaussians belong to which object, while fixed-view lifting methods do not establish persistent ownership. To address this problem, we present GaussAsset, an evidence-grounded met hod that turns a Gaussian carrier into auditable object assets by separating semantic proposal from ownership commitment. A frozen vision-language model proposes scene concepts and flags uncertain relations; a geometry-constrained executor lifts SAM3 masks through rendered depth, associates them across views, and commits mutually exclusive ownership over source Gaussians only when the evidence supports it, recording every decision in an auditable ledger. To evaluate this setting, we introduce GaussAsset-100, a benchmark of 100 furnished scenes with 3,590 human-verified object cuboids. The annotations target countable, exportable entities above a size gate and are provided as object-level cuboids, the unit downstream users export; membership diagnostics are box-derived and explicitly marked. Under a controlled shared-observation protocol that fixes the inputs for all methods, GaussAsset achieves the strongest object recall and all-GT bbox mIoU among the compared methods (77.5/67.2 at IoU 0.25/0.50; 55.4 bbox mIoU) and 56.1 scene-macro semantic mIoU without target-domain training. Together, these results support separating semantic proposal from ownership commitment as a practical route from a Gaussian carrier to auditable, exportable object assets.

1 Introduction

Robots, AR/VR systems and 3D editing pipelines increasingly start from a reconstructed indoor Gaussian scene and need *object-level assets*: stable, individually addressable sets of source Gaussians that share a semantic identity and a single owner. For robot perception, the immediately usable interface is an object's location, extent and orientation, but the underlying asset is the source-Gaussian subset from which that box is derived; for editing and asset reuse, the subset itself must remain stable, exportable and auditable. Re-scanning or re-optimizing the scene is not an option: categories are site-specific, real reconstructions contain floaters, missing coverage and view-dependent artifacts, and a wrong merge or split is not recoverable downstream. We therefore treat four requirements as fixed: reuse the existing carrier, survive real capture artifacts, work with an open vocabulary, and produce auditable object-level ownership. The central difficulty is *ownership commitment*: deciding which Gaussians belong to which object in a way that can be justified, audited and withheld when the evidence is insufficient. We address that problem from both sides, with a method for evidence-grounded decomposition and a dataset built for object-level Gaussian assets.

Method-side gap. 3DGS primitives carry no persistent semantic or instance ownership, so object editing and asset reuse require each object to own a stable subset of the source Gaussians. Existing methods either optimize scene-specific semantic features Zhou et al. (2024); Qin et al. (2024); Ye et al. (2024) or lift 2D predictions from a fixed image set Nguyen et al. (2024); Yin et al. (2024); Yan et al. (2024). Both are limited by what has already been observed, and independently predicted masks can fragment one object or merge neighboring instances. Because membership is committed

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

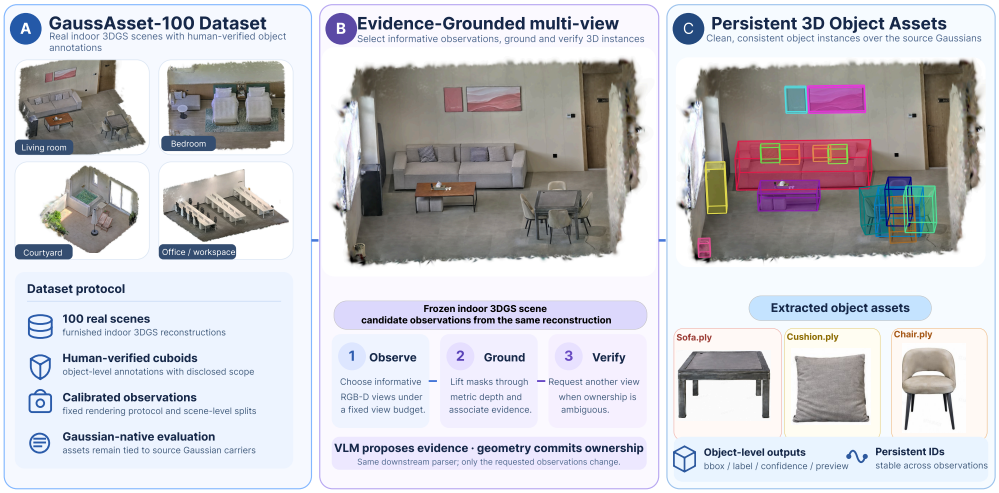

Figure 1: **Overview of GaussAsset and GaussAsset-100.** (a) A benchmark of real 3DGS scenes with human-verified object cuboids. (b) The pipeline proposes evidence with a frozen VLM and commits ownership with a geometric executor. (c) The result is a set of persistent objects over the source Gaussians.

implicitly, a confident but wrong prediction is hard to detect or reverse, and none of these methods separates a semantic proposal from a geometric ownership decision.

Method. We present GaussAsset, an evidence-grounded pipeline that separates semantic proposal from ownership commitment. A frozen VLM proposes scene concepts and flags uncertain relations; a geometry-constrained executor lifts SAM3 masks through rendered depth, associates them across views, and commits mutually exclusive ownership over source Gaussians only when the evidence supports it. Unsupported primitives remain unassigned rather than being forced into a merge or split, and every decision is recorded in an auditable ledger. The resulting assets are exported as per-object source Gaussian subsets, boxes and metadata; the carrier is never re-optimized and no target-domain training is required. The typed interface additionally specifies evidence goals and supplementary-view requests; we study this interface through an observation-policy pilot and report it separately from the main shared-observation results.

Data-side gap and dataset. Progress on ownership is also limited by the absence of suitable data. Real capture is the scarcer resource here: synthetic rooms offer control but omit the reconstruction artifacts that decomposition must survive. Existing 3D resources are synthetic Fu et al. (2021); McCormac et al. (2017), annotate meshes or point clouds Dai et al. (2017); Chang et al. (2017); Mo et al. (2019); Deitke et al. (2023), or label fixed 2D observations Kerr et al. (2023); the closest real-scene Gaussian datasets target per-point semantics or query segmentation rather than exportable objects Li et al. (2025); SpatialVerse Research Team (2025). We introduce GaussAsset-100: 100 audited, real furnished showrooms reconstructed as 3DGS with human-verified object cuboids, spanning 3,590 objects in 69 canonical categories across 203.1M source Gaussians, area-stratified into S/M/L/XL, with scene-level splits and a frozen protocol for object decomposition, semantic identity and observation-policy diagnostics. Cuboids provide the robot-facing annotation unit because downstream localization and interaction consume position, extent and orientation. For evaluation, we convert annotated cuboids into proxy Gaussian memberships by box inclusion to assess ownership quality in noisy 3DGS. Structural surfaces and objects too small to recover reliably remain context.

Evidence and scope. Our evaluation is organized in layers. The shared-observation protocol controls the carrier, views and mask front end, isolating the 2D-to-3D ownership decision; the semantic comparison uses the same carrier; and a separate observation-policy study examines view selection under a matched budget. Chorus is the primary same-carrier comparison, and additional comparisons follow the protocols described in the experiments. The main results, ablations and transfer experiments are reported in Section 5.

Our contributions are:

- **Evidence-grounded segmentation into auditable Gaussian object assets.** A VLM proposes semantics and flags uncertainty, depth-grounded projection and cross-view association establish support, and an auditable ledger commits persistent, mutually exclusive ownership over the source Gaussians; ambiguous relations are left unassigned or recorded for the reserved targeted-verification protocol.
- **GaussAsset-100: a real-scene Gaussian benchmark for object assets.** GaussAsset-100 contains human-verified oriented cuboids, a disclosed size-gated interactive-object scope, calibrated novel views, scene-level splits and a frozen evaluation protocol; its scale and full statistics are summarized in Section 4.
- **Protocols and diagnostics for asset-level evaluation.** All-GT axis-aligned bbox mIoU and object recall are primary; matched OBB and exported-box measures are secondary; membership diagnostics expose the welding and fragmentation that box overlap hides.

2 Related Work

Semantic and instance-aware Gaussians. 3D Gaussian Splatting Kerbl et al. (2023) provides explicit primitives that are convenient for rendering and editing. Feature 3DGS Zhou et al. (2024) and LangSplat Qin et al. (2024) attach semantic or language-aligned features, Gaussian Grouping Ye et al. (2024) adds identity features for instance grouping, and OpenMask3D Takmaz et al. (2023), Open3DIS Nguyen et al. (2024), SAI3D Yin et al. (2024), and MaskClustering Yan et al. (2024) lift multi-view masks or features into 3D instances. These methods optimize a scene-specific representation or a fixed set of observations; GaussAsset treats the input scene as a fixed carrier and asks how evidence can establish persistent *ownership* of its original primitives, committing them only through geometry-validated ownership.

Training-free Gaussian segmentation. Several methods transfer 2D masks to existing Gaussian scenes without learning a persistent semantic feature field. FlashSplat (Shen et al., 2024) directly solves Gaussian label assignments, while Lifting by Gaussians (Chacko et al., 2025) associates lifted fragments into scene-level instances. GaussianCut (Jain et al., 2024) supports interactive extraction through graph-cut inference. B^3-Seg (Kamata et al., 2026) accumulates mask evidence through Bayesian updates, and Seed2GS (Ding et al., 2026) propagates a reference mask across rendered views for target-object extraction. GaussAsset shares this training-free setting and emphasizes conservative, traceable ownership decisions for scene-level object-asset decomposition.

Datasets and benchmarks. Indoor 3D resources differ in what they reconstruct and what they supervise. ScanNet Dai et al. (2017), ScanNet++ Yeshwanth et al. (2023) and Matterport3D Chang et al. (2017) provide real RGB-D scans with point- or mesh-level labels, but that ground truth is attached to surfaces and does not carry over to Gaussian primitives. DL3DV-10K Ling et al. (2024) scales real capture for novel-view synthesis yet supplies no object-level annotation, and synthetic or object-centric resources such as 3D-FRONT Fu et al. (2021), SceneNet RGB-D McCormac et al. (2017), PartNet Mo et al. (2019) and Objaverse Deitke et al. (2023) annotate layouts, parts or individual shapes. Among Gaussian-native resources, LERF Kerr et al. (2023) evaluates open-vocabulary queries, InteriorGS SpatialVerse Research Team (2025) attaches instance semantics to 3D boxes and converts them to per-Gaussian labels through a separate assignment step Li et al. (2026), and SceneSplat Li et al. (2025) propagates surface annotations into a language-feature space. None of them defines mutually exclusive, exportable object ownership over the source primitives. Synthetic rooms and surface-annotated scans are abundant, whereas real Gaussian carriers that preserve capture artifacts and provide object-level asset annotations are rare. GaussAsset-100 addresses that gap with human-verified oriented cuboids for countable interactive objects; Appendix Table 7 positions GaussAsset-100 against these resources.

3 Method

Figure 2 summarizes the pipeline: geometry support cleans the carrier and proposes a bounded view pool; VLM-guided parsing consolidates the vocabulary and lifts SAM3 masks through depth; and the executor commits only compatible claims. Ambiguous relations remain read-only or reserved for next-evidence actions, and exports are derived from frozen membership.

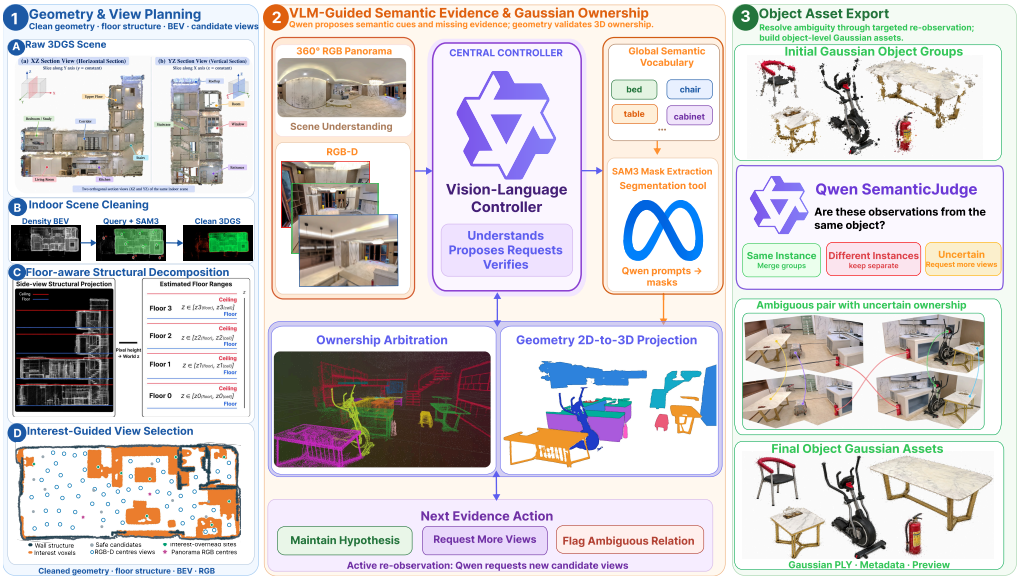

Figure 2: **Method overview.** GaussAsset extracts object assets from a fixed indoor 3D Gaussian scene. Scene cleaning and floor decomposition establish geometric support for interest-guided view selection. A frozen VLM proposes scene concepts, and SAM3 masks are lifted to source Gaussians through rendered depth. Cross-view association and geometric ownership gates determine mutually exclusive object memberships, which are exported with semantic labels and provenance metadata.

Problem and input contract. Given an unlabeled indoor Gaussian scene $\mathcal{G} = \{g_i\}_{i=1}^{N}$ with centers $\boldsymbol{\mu}_i$, covariances $\boldsymbol{\Sigma}_i$, opacities α_i and appearance coefficients $\mathbf{h}_i$, we recover semantic object instances directly over its *source* primitives without re-optimizing the scene. The output is $\mathcal{O} = \{(c_k, \mathcal{M}_k, q_k)\}_{k=1}^{K}$, where c_k is a semantic label, $\mathcal{M}_k \subseteq \{1, \ldots, N\}$ the source-Gaussian indices assigned to object k, and q_k summarizes its supporting evidence. Memberships are mutually exclusive, insufficient evidence leaves primitives in an explicit unassigned set $\mathcal{U}$, and scene preparation tracks rejected primitives in $\mathcal{I}_{\text{excluded}}$:

$$\{1, \ldots, N\} = \left(\biguplus_{k=1}^{K} \mathcal{M}_k \right) \uplus \mathcal{U} \uplus \mathcal{I}_{\text{excluded}}. \tag{1}$$

The only required input is a metrically scaled, renderable 3DGS in an approximately Z-up indoor frame; the pipeline neither re-optimizes nor edits the reconstruction, so membership stays an exhaustive, mutually exclusive and explicitly auditable index assignment over the original Gaussians.

An *object asset* is the exported tuple $(c_k, \mathcal{M}_k, q_k)$ together with its box, preview and metadata: the source-Gaussian index set, the label and quality summary, and the evidence that justified the commit. *Auditable* means the tuple is replayable. The ledger records the evidence, thresholds and version behind every commit, unassigned and excluded primitives stay visible instead of being silently merged or dropped, and the box and preview are derived from frozen membership and a validated scene basis rather than refit to improve a reported metric. Here, an object is a countable, exportable scene entity above the disclosed size gate. Its oriented cuboid is the robot-facing target; its source-Gaussian support set is the carrier evidence used to derive and audit that target. Structural surfaces, background context, parts, containers and objects below the recovery gate remain context or unassigned evidence rather than forced segments.

We evaluate object-asset quality on annotated cuboids and ownership diagnostics on box-derived proxies (purity, welding, fragmentation, coverage and $\text{mAP}^{G,\dagger}$), with bounded observation-policy and controller-role studies in Section 5.3.

Design rationale. The carrier is never re-optimized, and no commit rests on a single mask: real reconstructions contain floaters, missing coverage and view-dependent artifacts. Semantics are proposals from a frozen VLM and finite vocabulary; ownership must therefore pass an auditable gate.

Controller and authority. The central controller of Figure 2 is specified as a typed interface whose actions are clean, split_floor, discover_vocab, set_goal, verify, request_view and stop; every

action carries schema-validated target, evidence, payload and confidence fields. The VLM proposes concepts and identifies missing evidence; the geometric executor validates every request against floor, collision, clipping and visible-surface contracts; and only depth-grounded projection, cross-view association and the ownership gate can commit membership. Invalid, stale or conflicting actions are rejected and recorded as explicit abstentions. The evaluated mainline implements the scene-preparation and vocabulary roles, and the same interface specifies goal, verification and view-request actions (full schema in Appendix A.2). VLM outputs are therefore typed proposals; each must pass schema, geometry and camera-safety checks.

Evidence and persistent ownership. Scene preparation fixes the valid support and the camera pool (Appendix A.5). The VLM consolidates a finite scene vocabulary; SAM3 masks are lifted through rendered depth so that a mask supports only Gaussians inside it and depth-consistent, each keeping view, mask, ray and depth provenance. Projected masks are associated across views under geometric overlap, cross-view agreement and hard cannot-link constraints (Appendix A.7). Instance cores are frozen only when enough independent views support them and the best-to-second margin is sufficient; remaining carriers may attach to frozen cores but cannot rewrite them. A Gaussian is committed to an instance only when its accumulated support and its margin over the second-best candidate both pass the ownership gate,

$$A_{ik^{(1)}} \geq \tau_A, \qquad A_{ik^{(1)}} - A_{ik^{(2)}} \geq \tau_{\text{margin}}, \tag{2}$$

and otherwise remains in $\mathcal{U}$. Cores are frozen before attachment because association is order-sensitive: letting late views rewrite early cores would diffuse one ambiguous view through the scene. The margin term matters because high support can still be ambiguous between two nearby objects, such as a chair pushed under a table. Appendix A gives the full derivations.

Ambiguity, abstention and output. GaussAsset commits ownership when the accumulated geometric support and ownership margin satisfy the acceptance criteria. Primitives with insufficient support remain unassigned, while ambiguous relations and their supporting evidence are recorded in the ledger. After processing the selected observations, memberships are frozen and each object is exported as a subset of the original Gaussians, together with its semantic label, confidence, bounding box, and provenance metadata. Unassigned and excluded primitives are retained separately. The ledger records ownership decisions and supports audit and replay, preserving traceability from each exported asset to its source primitives.

4 GaussAsset-100: A Real-Scene Gaussian Benchmark

Scale, quality and provenance. GaussAsset-100 contains real furnished indoor Gaussian scenes; Table 1 gives its scale and Appendix B.1 gives construction and quality control. Every scene passes a usability audit, while floaters, incomplete coverage, transparency and reflection remain in the carrier.

Table 1: GaussAsset-100 at a glance. Frozen scale and annotation statistics for the audited benchmark.

Scale and split		*Annotation and distribution*	
Reconstructed indoor 3DGS scenes	100	Annotated objects	3,590
Source Gaussian primitives	203.1M	Canonical categories	69
Median Gaussians / scene	1.83M	Raw surface forms	86 → 69
Gaussian range / scene	0.97–5.31M	Top-10 category share	55.5%
Median objects / scene	34	Categories ≤ 5 objects	27 (10 singletons)
Objects p10/p90	25 / 48	Source Gaussians / object	56.6k (mean)
Median floor area	68.3 m^2	Carrier points / object	8.0k (median proxy)
Area strata (S/M/L/XL)	19/30/30/21	Evaluation split	scene-level (100 scenes)

Annotation protocol: why cuboids. Each object has a human-verified category and oriented cuboid (center, size, yaw) in scene coordinates; surface forms map to the canonical categories in Table 1. The cuboid is the robot-facing GT: it remains well defined under sparse or incomplete reconstruction, while a watertight mesh would be difficult to obtain and transfer onto noisy Gaussians SpatialVerse Research Team (2025); Li et al. (2026). We derive a carrier support set by box inclusion to fit and audit boxes; it sharpens ownership diagnostics but is not human point-level GT. Thus all-GT bbox mIoU and recall are direct annotation metrics, while box-derived membership quantities carry † (quality control is in Appendix B.1).

Scope and statistics. Targets are countable interactive/exportable objects above a disclosed size gate; structural surfaces and unrecoverably small objects are context. The category distribution is long-tailed (Table 1).

Tasks, splits and availability. GaussAsset-100 supports shared-observation decomposition, semantic identity and a bounded observation-policy pilot, and documents the full verification protocol. It provides the original carriers, calibrated novel views, scene-level split and frozen protocol; evaluation code, metadata and preview scene samples are released in the anonymous supplement, and the complete benchmark follows publicly. Cuboid-to-Gaussian diagnostics carry †, and unmatched objects count as zero in bbox mIoU.

The fixed carrier, splits and protocol keep welding-inflated recall and instance-level failures visible.

5 Experiments

The evaluation measures decomposition and semantic identity on the frozen benchmark, then studies controller/ownership ablations, the observation pilot, and transfer (protocol details are in Appendix B).

5.1 Setup, Metrics and Protocols

Setup, ground truth and baselines. GaussAsset-100 is the primary benchmark; transfer uses InteriorGS and LERF-OVS. GaussAsset is frozen and training-free, whereas Chorus uses its target-domain pseudo-label head. Compatible methods share scenes, cameras, RGB-D views and mask front ends; membership diagnostics are box-derived proxies. Where used, the VLM supplies proposals only: the vocabulary-only arm is a fixed-prompt control and no VLM output commits membership.

Protocols and metrics. The pipeline renders calibrated novel views and never reuses reconstruction-time RGB frames. Shared-observation decomposition is primary and gives compatible methods identical carriers, cameras and mask front ends; the observation pilot and controller ladder are separate layers, and closed-loop verification is specified in Appendix Table 8. Primary metrics are recall at IoU 0.25/0.50, all-GT bbox mIoU and matched oriented-box IoU; coverage, purity, welding and fragmentation are diagnostics.

5.2 Object Decomposition and Semantics on GaussAsset-100

Table 2 reports the primary comparison: GaussAsset leads recall and all-GT bbox mIoU with high purity and sub-2% welding. Full diagnostics and the separate 71-scene public bridge are in Appendix B.

Table 2: Object decomposition on GaussAsset-100 under the shared-observation protocol. Bbox mIoU is the all-GT axis-aligned metric, with unmatched objects counted as zero; membership diagnostics are box-derived proxies (†).

Method	R@0.25↑	R@0.50↑	bbox mIoU↑	purity† ↑	weld† ↓	frag† ↓
MaskClustering	74.7	43.5	47.6	92.1	**0.8**	14.8
SAI3D	72.6	61.7	51.4	94.9	6.9	**5.0**
GaussAsset (ours)	**77.5**	**67.2**	**55.4**	**95.6**	1.7	7.7

Under the exported-box protocol at IoU 0.50, the production rule scores precision/recall/matched IoU of 63.7/60.0/75.9.

Semantic identity. Chorus is the closest native-Gaussian comparison because it uses the same carrier. Its linear head is trained on target-domain pseudo-labels, whereas our VLM and SAM3 remain frozen; on the 53-class in-scope set (90.5% of foreground points), scene-macro mIoU rises 44.0 → 56.1 and pooled mIoU 37.7 → 50.4, with micro recall 51.2 → 62.8. Figure 4 shows the corresponding instance/class contrast.

5.3 Controller Roles, Ablations and Observation Pilot

Controller roles. Three controlled studies separate controller roles, view selection and ownership. The geometric executor, camera pool, SAM3 masks and rendering budget stay fixed; goal-conditioned

Table 3: Semantic segmentation on GaussAsset-100. The controlled comparison is the frozen 53-class in-scope set, covering 90.5% of annotated foreground points; scene-macro metrics average over scenes and pooled metrics aggregate foreground points. This panel is the semantic headline comparison.

Method	Scope	Scene macro		Pooled		Micro
	(classes / GT points)	fine-grained mIoU↑	fine-grained mAcc↑	fine-grained mIoU↑	fine-grained mAcc↑	recall↑
In-scope 53-class evaluation						
Chorus (trained)	53 / 90.5%	44.0	53.3	37.7	45.3	51.2
Ours (frozen)	53 / 90.5%	**56.1**	**62.6**	**50.4**	**58.4**	**62.8**

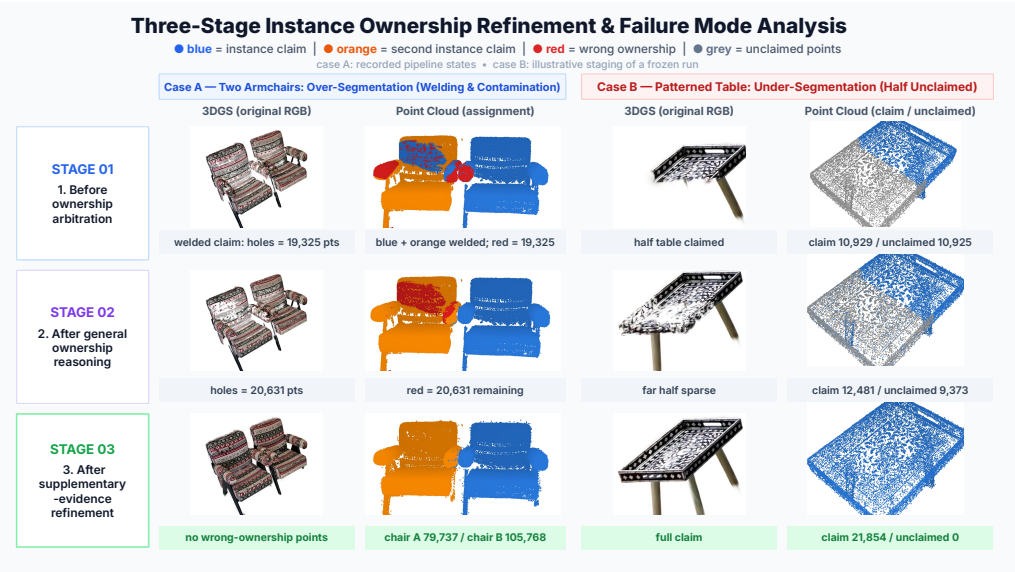

Figure 3: **Two ownership failure modes.** Case A shows over-clustering, where the unconstrained rule welds two adjacent armchairs into one instance; staged ownership reasoning with recorded supplementary evidence separates them. Case B shows under-clustering, where the same rule leaves part of a patterned table surface unclaimed and the refinement recovers it. Both cases are qualitative illustrations; benchmark-scale statistics are in Tables 5 and 12.

execution follows the same typed interface (Appendix A.2). On the frozen benchmark, $\text{mAP}^{G,\dagger}$ rises from 5.6 (geometry only) to 13.9 (vocabulary), 16.6 (cleaning/floor splitting) and 20.6 (production reference), while $\text{AP}^{\dagger}$@50 rises 10.7 → 38.5 and visible coverage 91.8% → 98.0%. The vocabulary tier uses a fixed prompt table; the benchmark membership and box diagnostics are in Appendix Table 15.

Observation policy. The pilot isolates view selection: policies share initial observations, candidate pool, rendering budget and the same SAM3, projection, association and ownership modules. The three-stream selector gives the highest coverage recall, while quality remains close to uniform resampling (Table 4).

Ownership ablations and risk routing.

Ownership ablations. Table 5 evaluates ownership rules on the frozen benchmark under identical observations, masks, and thresholds. Removing identity consolidation increases welding from 1.69% to 7.35%, while naive same-label union reduces coverage to 28.6% and increases welding to 29.64%. Removing ownership closure yields a welding rate of 2.71%. These results support the role of identity consolidation and ownership gates in maintaining coverage while reducing incorrect merges. Detailed ownership-rule and threshold ablations are provided in Appendix B.4.

Figure 3 illustrates the same staged behaviour on two representative failures: two geometrically close armchairs are wrongly merged, and a patterned table surface is left partly unclaimed. The

Table 4: Observation policy on the three pilot scenes of the frozen benchmark, under an identical parsing backend. Cubemap rows are cost references at their actual view counts; all rows are bounded pilot numbers.

Observation policy	Visible G. ↑	$R^{\dagger}_{\text{cov}}$@0.5 ↑	mAP^{G,†} ↑	Views ↓
Uniform floor-level views	98.3	64.9	14.4	676
Random floor-level views	97.8	68.5	14.3	676
Geometry-only coverage views	98.1	68.0	13.9	676
Cubemap only	74.7	15.6	2.9	50
Cubemap + perspective RGB-D	98.3	52.7	7.7	676
Full three-stream policy (ours)	98.3	70.5	**14.7**	676

Table 5: Ownership-rule ablations on the frozen benchmark, with observations, masks and thresholds fixed. Membership quantities are box-derived proxies (†); box IoU is a Z-up refit diagnostic, and the exported-box point is reported in Section 5.2.

Rule	cov.† ↑	$R^{\dagger}_{50}$ ↑	purity† ↑	weld† ↓	frag† ↓	PQ$^{\dagger}_{50}$ ↑	box IoU↑	point F1† ↑
Production rule	**57.6**	**70.3**	**95.6**	**1.69**	**7.72**	51.6	**73.4**	64.2
No identity consolidation	57.4	69.7	94.0	7.35	10.06	54.7	67.2	66.7
Naive semantic union	28.6	34.1	85.2	29.64	2.59	25.3	45.0	42.2
No ownership closure	56.3	68.4	**95.0**	2.71	8.94	52.4	69.9	64.6

examples illustrate the ownership failure modes qualitatively; they are not evidence of a closed-loop verification result on the benchmark.

5.4 Transfer to Public Benchmarks

Transfer to public benchmarks. Appendix Table 10 reports the transfer numbers. On InteriorGS, GaussAsset reaches 11.82 foreground mIoU without target-specific 3D training, and its geometric correspondence on labeled Gaussians is much stronger than final membership coverage, suggesting that object discovery, which the semantic stage targets, dominates this transfer

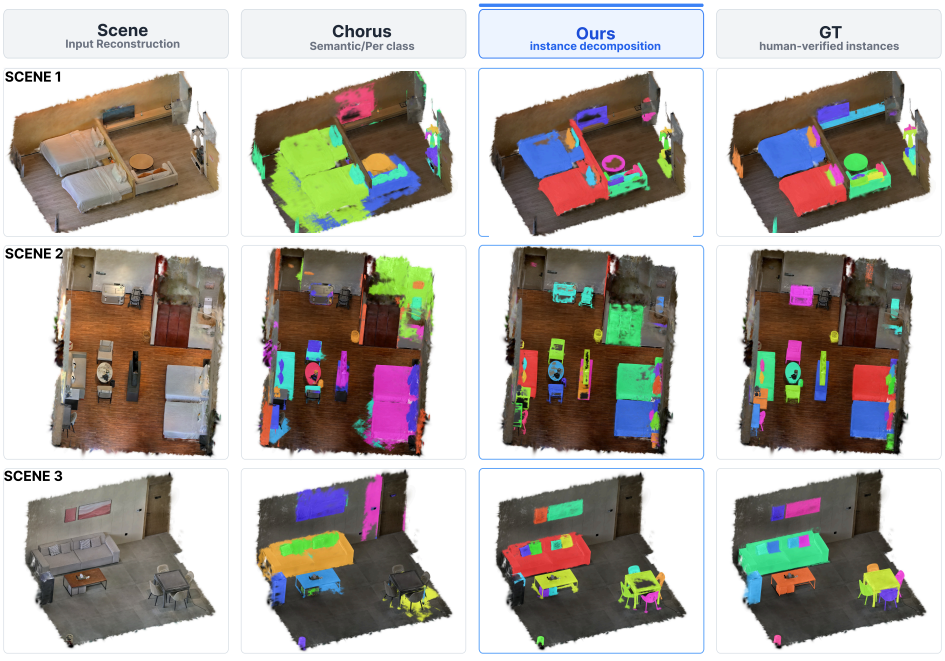

Figure 4: **Qualitative comparison with Chorus.** Columns show the same carrier and camera as the scene, Chorus, our decomposition and the ground-truth cuboids. Chorus uses its target-domain pseudo-label head, while our components are frozen; Chorus is shown per class, restricted to the classes annotated in that scene, while ground truth and our result are colored per instance. GT denotes human-verified cuboids, not point-level Gaussian membership.

probe. On LERF-OVS it reaches 38.46 query-mask mIoU and 52.80 scene-macro persistent-instance IoU. Published baselines use different protocols, so these are transfer results under each published protocol; additional cases are in Appendix B.13.

6 Limitations

Our claims are bounded by the carrier, annotation unit and evaluation protocols. The contract assumes metrically scaled indoor 3DGS with stable gravity; unknown scale, arbitrary gravity, outdoor scenes and strongly non-planar spaces remain out of scope. Floaters, transparency, specular or textureless surfaces, thin structures and heavy occlusion can violate depth or mask completeness. Vocabulary gaps and ambiguous container–content or object–part relations can therefore lead to abstention or incorrect proposals, especially near contact boundaries.

GaussAsset-100 provides human-verified cuboids rather than meshes or point-level Gaussian membership. Cuboids make robot-facing position, extent and orientation measurable, but do not supervise shape detail, parts, containers or heavily occluded extent; all membership quantities marked † are box-derived proxies. The carrier support set is reproducible evidence for box fitting and welding/fragmentation analysis, not an independent point-level annotation. Transfer rows follow each baseline's published protocol, and we report them as reference points under that protocol.

The controller ladder is measured on the frozen benchmark, whereas the observation-policy study is a bounded fixed-budget pilot. Goal-conditioned verification is specified as the next-stage protocol, building on the implemented controller roles. Supplementary observation adds rendering, SAM3 and VLM cost and cannot recover geometry absent from the input reconstruction. The semantic comparison uses a frozen GaussAsset configuration and a native baseline with a different training regime. The gain therefore measures protocol behavior, not a training-matched comparison. The category distribution is long-tailed and the size gate favors recoverable interactive objects, so headline averages are dominated by common furniture. Counts and area strata are reported, but uniform per-category performance remains open.

7 Conclusion

Reusing a reconstructed Gaussian scene requires object-level assets: stable sets of source Gaussians with a semantic identity and a single owner for downstream editing and simulation. Real captures contain floaters, missing coverage and view-dependent artifacts, while implicit ownership makes wrong merges and splits hard to audit. We presented GaussAsset, an evidence-grounded method that separates semantic proposal from ownership commitment. A frozen VLM proposes scene concepts; a geometry-constrained executor lifts SAM3 masks through rendered depth, associates them across views, and commits mutually exclusive ownership over source Gaussians through an auditable ledger when evidence suffices. Insufficient support is recorded as an abstention, and every asset remains tied to its source primitives.

To evaluate this setting, we introduced GaussAsset-100, a benchmark of 100 real furnished indoor scenes reconstructed as 3DGS, with 3,590 human-verified object cuboids in 69 categories over 203.1M source Gaussians, a size-gated interactive-object scope, calibrated novel views, scene-level splits and a frozen evaluation protocol. Box-derived membership diagnostics expose the welding and fragmentation that box overlap alone hides and are explicitly marked as such.

Under the controlled shared-observation protocol, GaussAsset achieves the strongest object recall and all-GT bbox mIoU among the compared methods (77.5/67.2 at IoU 0.25/0.50; 55.4 bbox mIoU) and 56.1 scene-macro semantic mIoU on the 53-class in-scope set, without target-domain training or scene re-optimization. Ownership ablations show that identity consolidation and the ownership gate help maintain high coverage while limiting welding: removing identity consolidation or replacing the ownership gate with a naive semantic union sharply increases welding. Persistent ownership, explicit abstention and auditable provenance preserve asset quality and failure visibility.

Together, GaussAsset and GaussAsset-100 provide a route from a Gaussian carrier to auditable object assets and a reproducible setting for measuring progress. Natural extensions include a full goal-conditioned verification policy, point-level audit sets for membership, long-tail vocabulary, and downstream tasks such as manipulation, simulation and scene editing.

References

Jaehun Bang, Jinhyeok Kim, Minji Kim, Seungheon Jeong, and Kyungdon Joo. Lightsplat: Fast and memory-efficient open-vocabulary 3d scene understanding in five seconds. *arXiv preprint arXiv:2603.24146*, 2026.

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

Rohan Chacko, Nicolai Häni, Eldar Khaliullin, Douglas Lee, and Lin Sun. Lifting by gaussians: A simple, fast and flexible method for 3d instance segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2025.

Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Nießner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. In *3DV*, 2017.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, 2017.

Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, et al. Objaverse: A universe of annotated 3d objects. In *CVPR*, 2023.

Zongjian Ding, Yudong Gao, Jiale Liu, Xinglin Yu, Junxing Ren, Dong Wei, Yajing Chen, Shan Huang, Mingjun Cheng, and Min Li. Seed2gs: Camera-free, training-free object extraction from 3d gaussian scenes via a single reference-view grounding. *arXiv preprint arXiv:2608.11928*, 2026.

Huan Fu, Bowen Cai, Lin Gao, Ling-Xiao Zhang, Jiaming Wang, Cao Li, Qixun Zeng, Chengyue Sun, Rongfei Jia, Binqiang Zhao, and Hao Zhang. 3d-front: 3d furnished rooms with layouts and semantics. In *ICCV*, 2021.

Umangi Jain, Ashkan Mirzaei, and Igor Gilitschenski. GaussianCut: Interactive segmentation via graph cut for 3d gaussian splatting. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Hiromichi Kamata, Samuel Arthur Munro, and Fuminori Homma. B^3-Seg: Camera-free, training-free 3DGS segmentation via analytic EIG and beta–bernoulli bayesian updates. *arXiv preprint arXiv:2602.17134*, 2026.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4):139:1–139:14, 2023. doi: 10.1145/3592433.

Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. LERF: Language embedded radiance fields. In *ICCV*, pp. 19729–19739, 2023.

Yue Li, Qi Ma, Runyi Yang, Huapeng Li, Mengjiao Ma, Bin Ren, Nikola Popovic, Nicu Sebe, Ender Konukoglu, Theo Gevers, Luc Van Gool, Martin R. Oswald, and Danda Pani Paudel. Scenesplat: Gaussian splatting-based scene understanding with vision-language pretraining. In *ICCV*, 2025. doi: 10.1109/ICCV51701.2025.00472.

Yue Li, Qi Ma, Runyi Yang, Mengjiao Ma, Bin Ren, Nikola Popovic, Nicu Sebe, Theo Gevers, Luc Van Gool, Danda Pani Paudel, and Martin R. Oswald. Chorus: Multi-teacher pretraining for holistic 3d gaussian scene encoding. In *CVPR*, pp. 21431–21442, 2026.

Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *CVPR*, 2024.

John McCormac, Ankur Handa, Stefan Leutenegger, and Andrew J. Davison. Scenenet rgb-d: Can 5m synthetic images beat generic imagenet pre-training on indoor segmentation? In *ICCV*, 2017.

Kaichun Mo, Shilin Zhu, Angel X. Chang, Li Yi, Subarna Tripathi, Leonidas Guibas, and Hao Su. Partnet: A large-scale benchmark for fine-grained and hierarchical part-level 3d object understanding. In *CVPR*, 2019.

Phuc Nguyen, Tuan Duc Ngo, Evangelos Kalogerakis, Chuang Gan, Anh Tran, Cuong Pham, and Khoi Nguyen. Open3dis: Open-vocabulary 3d instance segmentation with 2d mask guidance. In *CVPR*, pp. 4018–4028, 2024.

Minghan Qin, Wanhua Li, Jiawei Zhou, Haoqian Wang, and Hanspeter Pfister. Langsplat: 3d language gaussian splatting. In *CVPR*, pp. 20051–20060, 2024.

Qiuhong Shen, Xingyi Yang, and Xinchao Wang. FlashSplat: 2d to 3d gaussian splatting segmentation solved optimally. In *European Conference on Computer Vision*, 2024.

SpatialVerse Research Team. InteriorGS: A 3D gaussian splatting dataset of semantically labeled indoor scenes. `https://huggingface.co/datasets/spatialverse/InteriorGS`, 2025.

Ayça Takmaz, Elisabetta Fedele, Robert W. Sumner, Marc Pollefeys, Federico Tombari, and Francis Engelmann. Openmask3d: Open-vocabulary 3d instance segmentation. In *NeurIPS*, volume 36, 2023.

Mi Yan, Jiazhao Zhang, Yan Zhu, and He Wang. Maskclustering: View consensus based mask graph clustering for open-vocabulary 3d instance segmentation. In *CVPR*, pp. 28274–28284, 2024.

Mingqiao Ye, Martin Danelljan, Fisher Yu, and Lei Ke. Gaussian grouping: Segment and edit anything in 3d scenes. In *ECCV*, pp. 162–179, 2024. doi: 10.1007/978-3-031-73397-0_10.

Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *ICCV*, 2023.

Yingda Yin, Yuzheng Liu, Yang Xiao, Daniel Cohen-Or, Jingwei Huang, and Baoquan Chen. Sai3d: Segment any instance in 3d scenes. In *CVPR*, pp. 3292–3302, 2024.

Shijie Zhou, Haoran Chang, Sicheng Jiang, Zhiwen Fan, Zehao Zhu, Dejia Xu, Pradyumna Chari, Suya You, Zhangyang Wang, and Achuta Kadambi. Feature 3dgs: Supercharging 3d gaussian splatting to enable distilled feature fields. In *CVPR*, pp. 21676–21685, 2024.

A Detailed Method and Derivations

This appendix expands Section 3: the controller interface is in Appendix A.2, scene preparation in Appendices A.3 and A.4, observation planning in Appendix A.5, mask lifting and association in Appendices A.6 and A.7, ownership arbitration in Appendix A.8, the targeted supplementary-observation interface in Appendix A.9, and the end-to-end algorithm in Appendix A.10.

A.1 Overview

We keep the notation of Section 3 and do not restate it. $\mathcal{O} = \{(c_k, \mathcal{M}_k, q_k)\}_{k=1}^{K}$ is defined over the source indices, and Equation 1 partitions the primitives into disjoint instance sets, the unassigned set $\mathcal{U}$ and the scene-cleaning exclusion set $\mathcal{I}_{\text{excluded}}$. The remaining subsections follow the main-text order: scene preparation, observation planning, mask lifting, association, ownership arbitration, supplementary observation, and the end-to-end algorithm.

Unlike single-pass 3D segmentation, GaussAsset specifies preparation, observation, and verification as actions of one shared controller, as illustrated in Figure 2. The specification has three phases: **(1) scene initialization**, where clean and split_floor restrict the valid support; **(2) geometry-grounded segmentation**, where set_goal/vocab actions guide complementary observations and the executor forms depth-consistent instances; and **(3) verification and asset delivery**, where verify/request_view close residual ambiguities before export. The evaluated shared-observation mainline covers phase 1 and the segmentation backend of phase 2, the observation channel of phase 2 is reported as a bounded pilot, and phase 3 remains a reserved protocol (Section 5.3). The recurrent process is

$$(\mathcal{Z}_t, \mathcal{S}_t, \mathcal{H}_t) \xrightarrow{F_\theta} \mathbf{a}_t \xrightarrow{E} (\mathcal{Z}_{t+1}, \mathcal{S}_{t+1}, \mathcal{H}_{t+1}) \longrightarrow \mathcal{O}, \tag{3}$$

where F_θ is instantiated by a Qwen-family 27B vision-language model and E is the deterministic geometric executor. The final assets directly index the source representation with $\{\mathcal{M}_k\}$. No position, shape, or appearance parameter is refitted, so every output remains traceable to the original primitive, supporting view, and ownership operation.

A.2 Shared VLM Controller Interface

The VLM controller is a policy over visual evidence and a typed scene state, not an unrestricted text interface. At iteration t, the visual packet $\mathcal{Z}_t$ contains only views whose camera and source-scene hashes are recorded: the world-XY BEV, XZ/YZ side projections, selected RGB or RGB-D renders, and candidate-specific Evidence Atlases. The structured state $\mathcal{S}_t$ contains floor intervals, safe-camera identifiers, per-region coverage, candidate lineage, ownership margins, unresolved relations, and remaining budgets. The action history $\mathcal{H}_t$ prevents a controller from repeating an unsuccessful action without new evidence.

The seven legal actions are:

- clean(points, confidence): positive indoor and negative exterior BEV prompts;
- split_floor(segments, confidence): paired floor/ceiling segments in both side projections;
- vocab(concepts, aliases, evidence): a finite object vocabulary grounded to source views;
- set_goal: an object-centric region, view role, and priority;
- verify: a candidate pair, evidence IDs, and one of three verdicts (same, separate, or uncertain);
- request_view: a target, missing role, and reason for bounded re-observation request; and
- stop(reason): terminate when the task is resolved or no admissible action remains.

Every response is validated against a JSON schema, current evidence scope, and lineage hashes. A malformed response is retried a fixed number of times; stale or geometrically infeasible actions are rejected. Only the geometric executor can update camera sets or the OwnershipLedger. We use a Qwen-family 27B vision-language model for F_θ, with deterministic decoding, while keeping the action schema and executor independent of the backbone. The interface below is the full specified contract the executor validates against. The evaluated mainline implements the scene-preparation and vocabulary roles, and the same interface specifies goal, verification and view-request actions (Section 5.3).

A.3 World-Coordinate BEV-Based Scene Cleaning

Raw 3DGS reconstructions may contain exterior content visible through windows, isolated floaters, and poorly reconstructed regions. These regions can introduce irrelevant visual evidence and enlarge the search space for subsequent instance parsing. We therefore clean the scene before floor decomposition and observation planning.

Let the input scene be $\mathcal{G} = \{g_i\}_{i=1}^{N}$, where the center of Gaussian g_i is $\mathbf{p}_i = (x_i, y_i, z_i)$. We assume that the input reconstruction provides an approximately Z-up world coordinate system, so that the world XY plane approximates the physical horizontal plane. The cleaning module does not perform PCA rotation, gravity estimation, or axis reorientation.

World-coordinate BEV construction. We construct a bird's-eye view (BEV) directly from Gaussian centers in the original world coordinate system. This BEV is a density map, not a virtual-camera RGB rendering. For all finite Gaussian centers, we determine the horizontal scene extent and add a fixed boundary margin δ:

$$\begin{aligned} x_{\min} &= \min_i(x_i) - \delta, \quad x_{\max} = \max_i(x_i) + \delta, \\ y_{\min} &= \min_i(y_i) - \delta, \quad y_{\max} = \max_i(y_i) + \delta. \end{aligned} \tag{4}$$

Given a BEV grid resolution r, Gaussian g_i is mapped to pixel (u_i, v_i) by

$$u_i = \left\lfloor \frac{x_i - x_{\min}}{r} \right\rfloor, \qquad v_i = \left\lfloor \frac{y_{\max} - y_i}{r} \right\rfloor. \tag{5}$$

Image direction $+u$ corresponds to world direction $+X$, whereas $+v$ corresponds to world direction $-Y$; the center of BEV pixel (u, v) maps back to $x_c = x_{\min} + (u + 0.5)r$ and $y_c = y_{\max} - (v + 0.5)r$.

The height coordinate z does not affect the BEV pixel location. For a BEV pixel p, we define its Gaussian count and log-density as

$$n(p) = \sum_{i=1}^{N} \mathbf{1}\left[(u_i, v_i) = p\right], \qquad B(p) = \log\left(1 + n(p)\right). \tag{6}$$

Because the Gaussian-to-pixel mapping is retained, the BEV provides a deterministic spatial interface for scene cleaning without requiring camera intrinsics or extrinsics.

Prompt-guided indoor-support extraction. We provide the BEV density map and its world-coordinate extent to the shared controller, whose clean action proposes positive and negative point prompts for SAM3. Positive points indicate continuous, high-density indoor support, whereas negative points indicate distant window-exterior structures or isolated Gaussian regions. SAM3 predicts an indoor-support mask $M_{\text{in}}(u, v)$, and each Gaussian g_i takes its mask value by direct pixel lookup

$$\begin{aligned} m_i &= M_{\text{in}}(u_i, v_i), \\ \mathcal{G}_{\text{clean}} &= \{g_i \in \mathcal{G} \mid m_i = 1\}. \end{aligned} \tag{7}$$

This operation removes Gaussians outside the estimated indoor support, but does not modify the position, color, opacity, scale, or other attributes of the retained primitives. The excluded set is $\mathcal{I}_{\text{excluded}} = \mathcal{G} \setminus \mathcal{G}_{\text{clean}}$. It is not assigned to indoor object instances and is tracked separately for stage-level quality analysis.

The BEV resolution r and margin δ inherit the spatial units of the input PLY. Their physical interpretation therefore requires a metric-scale reconstruction. The Z-up assumption and the sensitivity of the cleaning mask to r, δ, and VLM/SAM3 errors define the applicability boundary of this front end.

A.4 Multi-Floor Indoor Scene Decomposition

An XY-plane support mask alone cannot distinguish floors that overlap in horizontal projection. We therefore decompose the cleaned scene into floor-level sub-scenes before generating observations. For this purpose, we project Gaussian centers to the world-coordinate XZ and YZ planes:

$$\begin{aligned} (x_i, z_i) \to (u_i^{xz}, v_i^{xz}), \qquad (y_i, z_i) \to (u_i^{yz}, v_i^{yz}), \\ \ell_i^x = x_i^{\max} - x_i^{\min}, \qquad \ell_i^y = y_i^{\max} - y_i^{\min}. \end{aligned} \tag{8}$$

The two side projections retain the mapping between image pixels and world coordinates: for each indoor region i, the XZ projection gives the side-view extent along X, whereas the YZ projection gives the corresponding extent along Y, with mapped horizontal bounds $[x_i^{\min}, x_i^{\max}]$ and $[y_i^{\min}, y_i^{\max}]$. These lengths are measured in the original world units and retained as geometric metadata for subsequent floor-level scene analysis. The same side projections also provide the vertical evidence needed for floor-boundary estimation.

We provide both side projections to the same controller. Its split_floor action returns floor and ceiling segments with structured confidence values. After converting segment pixels back to height candidates, we sort them and establish one-to-one XZ–YZ matches within tolerance ϵ_z. Consistent predictions are combined by a confidence-weighted median. A one-view prediction is accepted only when it agrees with a local peak or valley of the vertical Gaussian-density profile. If the projections conflict or yield crossing intervals, processing stops for that scene and does not commit a cross-floor cut. For an accepted floor l, let its bounds be z_l^{floor} and z_l^{ceil}; its Gaussian subset is

$$\mathcal{G}^{(l)} = \left\{ g_i \in \mathcal{G}_{\text{clean}} \mid z_l^{\text{floor}} \le z_i \le z_l^{\text{ceil}} \right\}. \tag{9}$$

The cleaned scene is consequently decomposed into floor sub-scenes $\{\mathcal{G}^{(1)}, \ldots, \mathcal{G}^{(L)}\}$, where L is the estimated number of floors. Subsequent geometry, camera generation, and view selection operate independently per floor. Adjacent floors use half-open intervals $[z_l^{\text{floor}}, z_l^{\text{ceil}})$, with a closed upper bound for the highest floor, so that a Gaussian belongs to at most one sub-scene. Primitives within ϵ_z of a boundary are marked as boundary elements and assigned only when later depth-visible evidence is decisive.

A.5 VLM-Guided, Geometry-Constrained Observation Planning

The observation planner below is specified in full; the main comparison uses shared observations, and the supplementary-view study is evaluated as a bounded pilot (Section 5.3).

The quality of 3D instance decomposition is constrained by *what the system observes*. After scene initialization, GaussAsset alternates between a semantic observation goal proposed by the shared controller and a safe camera chosen by the geometric executor. The controller never receives ground-truth instance labels; its goals are grounded to the current BEV, RGB evidence, or candidate lineage. The executor uses the geometry-only object-interest representation below to enforce feasibility and coverage.

Floor-level geometric representation. For each floor $\mathcal{G}^{(l)}$, we analyze the spatial distribution of Gaussian centers in the horizontal XY plane, together with local density, structural extent, and valid free space. The scene is voxelized into $\mathcal{X}$ and partitioned into ordinary structure $\mathcal{X}_{\text{ord}}$ (walls and other near-planar structural regions) and interest regions $\mathcal{X}_{\text{int}}$ with high geometric value for object observation. The latter is derived from geometry alone and does not encode semantic categories or final instance labels.

For reproducibility, we first estimate navigable space as the largest connected component of low-height free-space voxels. For each occupied voxel x, we compute height $h(x)$, local density $\rho(x)$, and planarity $\pi(x)$. Wall-connected, near-planar components that span most of the floor height are assigned to $\mathcal{X}_{\text{ord}}$. The remaining occupied components enter $\mathcal{X}_{\text{int}}$ when they exceed a minimum density, lie between the floor and ceiling, and do not belong to navigable space. Voxels that satisfy neither test remain eligible for surface coverage but do not contribute to the interest or structure terms. All scale-normalized thresholds are selected once on the validation set and then frozen.

Complementary candidate views. We generate three complementary candidate-view sets $\mathcal{V}_c = \mathcal{V}_{\text{cube}} \cup \mathcal{V}_{\text{top}} \cup \mathcal{V}_{\text{persp}}$. Here, $\mathcal{V}_{\text{cube}}$ contains Cubemap candidates from open regions near room centers. Each candidate produces six cube-face RGB observations for broad room-level context. $\mathcal{V}_{\text{top}}$ contains vertical top-down candidates placed above spatially extended object-interest regions. These candidates observe large horizontal structures, such as beds and sofas, that may be fragmented under lateral views. Their viewing direction follows the world $-Z$ direction.

$\mathcal{V}_{\text{persp}}$ contains ordinary perspective cameras sampled from valid floor-level free space. Their viewing directions prioritize nearby $\mathcal{X}_{\text{int}}$ after excluding wall regions, and each candidate produces a calibrated RGB image with its corresponding metric depth map.

The controller first observes a low-cost global view and emits $q_t = (R_t, \rho_t, \kappa_t)$, containing target region R_t, requested role ρ_t, and priority κ_t. All three candidate types then compete under a geometry-constrained criterion. Let $\nu_v(x) \in \{0, 1\}$ indicate that voxel x lies inside the frustum of v, faces the camera, and passes z-buffer visibility. For a voxel set A, define its marginal coverage as

$$\Delta C_A(v \mid \mathcal{V}_s) = \frac{\sum_{x \in A} \nu_v(x) \prod_{v' \in \mathcal{V}_s} (1 - \nu_{v'}(x))}{\max(1, |A|)}. \tag{10}$$

Using $A = \mathcal{X}_{\text{int}}$ or $\mathcal{X}_{\text{ord}}$ yields the interest and structural terms; ΔC_{surf} applies the same definition to visible surfels derived from the source Gaussians. The candidate score is

$$\begin{aligned} S_t(v) = \lambda_{\text{int}} \Delta C_{\text{int}}(v \mid \mathcal{V}_s) + \lambda_{\text{ord}} \Delta C_{\text{ord}}(v \mid \mathcal{V}_s) \\ + \lambda_{\text{surf}} \Delta C_{\text{surf}}(v \mid \mathcal{V}_s) + \lambda_g \Delta C_{\text{goal}}(v \mid q_t) + \lambda_u \Delta U_t(v). \end{aligned} \tag{11}$$

Here, ΔC_{goal} is the visible fraction of R_t satisfying the requested role, while ΔU_t is the visible fraction of uncertain members or candidate pairs implicated by the current state. The five nonnegative weights sum to one, with $\lambda_{\text{int}} > \lambda_{\text{ord}}$. To prevent a view family from vanishing during greedy selection, initialization first takes the highest-scoring view from each nonempty family; subsequent iterations compete over all remaining candidates.

At each iteration we select and append

$$\begin{aligned} v^* &= \arg\max_{v \in \mathcal{V}_c \setminus \mathcal{V}_s} S_t(v), \\ \mathcal{V}_s &\leftarrow \mathcal{V}_s \cup \{v^*\}. \end{aligned} \tag{12}$$

Here, total coverage C is the geometric union coverage. Selection stops when the controller emits stop, no safe candidate can satisfy the current goal, or

$$C(\mathcal{V}_s) \geq \tau \qquad \text{or} \qquad |\mathcal{V}_s| \geq B, \tag{13}$$

where B is the selected-view budget. This greedy policy favors views that add object evidence or new surfaces while reducing redundant observations. The final observation set is $\mathcal{V}_s = \{v_1, v_2, \ldots, v_K\}$.

Each selected view is represented as $O_v = (I_v, D_v, \Pi_v)$, where I_v is the RGB observation, D_v is the corresponding depth information when available, and Π_v contains the camera parameters. Cubemap and top-down RGB views may have no depth, whereas perspective RGB-D views provide calibrated depth for subsequent projection. The final observation package is $O = \{O_v \mid v \in \mathcal{V}_s\}$. The selected RGB observations return to the shared controller, whose vocab action constructs a scene-level vocabulary reused as SAM3 prompts. The perspective RGB-D observations additionally provide the calibrated geometric evidence used for Gaussian ownership.

Figure 2 summarizes the complete pipeline; the equations above specify its observation stage.

A.6 Semantic Vocabulary, Masks, and Geometric Evidence

Finite scene vocabulary. We arrange the selected RGB observations by floor and invoke the controller's vocab action to return structured countable-object nouns, necessary appearance modifiers, and alias groups. Lowercasing, lemmatization, and alias consolidation produce a finite vocabulary $\mathcal{L}$. Walls, floors, and ceilings are structural classes and do not share instance IDs with countable objects. To prevent free text from rewriting geometry, VLM outputs may only (i) generate vocabulary or point prompts for SAM3 Carion et al. (2025), (ii) label an existing candidate, or (iii) flag a relation for additional evidence. They cannot create, remove, or transfer Gaussian members.

For each perspective RGB-D view v, SAM3 produces a mask set $\mathcal{S}_v = \{(m_{vj}, \ell_{vj}, p_{vj})\}_{j=1}^{n_v}$, where m_{vj} is a binary mask, ℓ_{vj} its normalized label, and p_{vj} its confidence. Within a view, highly redundant masks are suppressed by confidence; overlapping pixels remain unresolved until single-owner arbitration.

Metric-consistent Gaussian projection. Let K_v be the camera intrinsics and $(R_v, \mathbf{t}_v)$ the world-to-camera transform. Gaussian i projects as

$$\mathbf{x}_{iv} = R_v \boldsymbol{\mu}_i + \mathbf{t}_v, \quad z_{iv} = [\mathbf{x}_{iv}]_z, \quad \mathbf{u}_{iv} = \pi(K_v \mathbf{x}_{iv}). \tag{14}$$

The renderer returns RGB-aligned metric depth D_v, accumulated opacity, and the pixel contribution ω_{iv} of each Gaussian. Depth is camera-axis z depth; ray distance must be converted before use. Define

$$d_{iv} = \frac{|z_{iv} - D_v(\mathbf{u}_{iv})|}{\max(D_v(\mathbf{u}_{iv}), \epsilon)}. \tag{15}$$

Gaussian i is visible only if its projection lies inside the image, $z_{iv} > 0$, $\omega_{iv} \geq \tau_{\text{vis}}$, and $d_{iv} \leq \tau_d$. Its support for mask j is

$$e_{ivj} = p_{vj}\omega_{iv} \exp\left(-\frac{d_{iv}^2}{2\sigma_d^2}\right) \mathbf{1}[\mathbf{u}_{iv} \in m_{vj}]\mathbf{1}[d_{iv} \leq \tau_d]. \tag{16}$$

The candidate carrier $H_{vj} = \{i : e_{ivj} > 0\}$ retains its view, mask, label, and depth provenance. Thus a 2D mask cannot expand through an unbounded frustum, and an occluded surface gains no support merely from pixel overlap.

A.7 Staged Cross-View Association

Each carrier initializes a graph node with immutable provenance. For nodes a, b, define symmetric membership overlap and directional coverage:

$$J_{ab} = \frac{|H_a \cap H_b|}{|H_a \cup H_b|}, \qquad Q_{ab} = \frac{1}{2}\left(\frac{|H_a \cap H_b|}{|H_a|} + \frac{|H_a \cap H_b|}{|H_b|}\right). \tag{17}$$

Let $P_{ab} = \exp(-\|\mathbf{c}_a - \mathbf{c}_b\|_2 / s_{ab})$ be scale-normalized proximity, and let V_{ab} be the cosine similarity between cross-view support vectors of shared members. The geometric association score is

$$G_{ab} = \beta_J J_{ab} + \beta_Q Q_{ab} + \beta_P P_{ab} + \beta_V V_{ab} - \beta_B R_{ab}^{\text{bridge}}, \tag{18}$$

where the nonnegative weights sum to one and R_{ab}^{bridge} is the fraction of low-density voxels along the shortest path after a hypothetical merge. Different floors, mutually exclusive structural classes, and nodes fully separated with a clear depth gap in the same high-confidence view induce hard cannot-link edges.

Association proceeds in two stages. First, we select anchors by mask confidence, depth-inlier ratio, visible member count, and camera-pose diversity. One-to-one Hungarian matching over candidates satisfying $G_{ab} \geq \tau_{\text{core}}$, label compatibility, and all cannot-link constraints forms high-confidence cores. A core freezes after support from at least n_{core} independent perspective views and a best-to-second match margin greater than δ_{assoc}. The second stage can only attach remaining nodes to frozen cores. A node remains unmatched if it is compatible with several cores without sufficient margin or would violate a transitive cannot-link constraint. This core-then-complete schedule prevents a weak early mask from initiating a chain of false merges.

A.8 STAR-3D: Selective Topology Arbitration and Single Ownership

The staged graph may still contain same-class fragments, local bridges, and overlapping members. **STAR-3D** (**S**elective **T**opology **A**rbitration and **R**obust Evidence control) handles only these residual ambiguities.

Five-factor relation evidence. For a candidate relation $r = (a, b)$ between instance groups, construct

$$\mathbf{f}^r = (f_{\text{sem}}^r, f_{\text{mask}}^r, f_{\text{mv}}^r, f_{\text{topo}}^r, f_{\text{VLM}}^r) \in [0, 1]^5. \tag{19}$$

Here, f_{sem} measures normalized label and part–whole compatibility; f_{mask} aggregates confidence and depth-inlier statistics from Equation 16; f_{mv} measures independent-view support and pose diversity; $f_{\text{topo}} = 1 - R_{ab}^{\text{bridge}}$ measures local connectivity; and $f_{\text{VLM}} \in \{0, 0.5, 1\}$ encodes keep_separate, uncertain, or same_instance. Continuous factors are clipped to $[0, 1]$ using bounds fixed on the validation set. No human-corrected reference or test annotation is used at inference.

These factors are not averaged unconditionally. A merge must have no cannot-link edge and must satisfy $f_{\text{mask}} \geq \tau_{\text{mask}}$, $f_{\text{mv}} \geq \tau_{\text{mv}}$, and $f_{\text{topo}} \geq \tau_{\text{topo}}$. Semantics or the VLM may veto a candidate or request a new observation, but cannot rescue a merge that fails a geometric gate. This asymmetry makes one design principle enforceable: semantics proposes, geometry commits.

Single-owner arbitration. Let $\mathcal{A}_k$ contain the mask nodes assigned to instance k. The ownership score of cleaned Gaussian i is

$$A_{ik} = \frac{\sum_{(v,j)\in\mathcal{A}_k} e_{ivj}}{\sum_v \omega_{iv}\mathbf{1}[i \text{ is visible in } v] + \epsilon}. \tag{20}$$

Let k_1, k_2 be the highest- and second-highest-scoring instances. We assign i uniquely to k_1 only if $A_{ik_1} \geq \tau_{\text{own}}$ and $A_{ik_1} - A_{ik_2} \geq \delta_{\text{own}}$; otherwise, i enters $\mathcal{U}$. A primitive never visible in a perspective RGB-D observation is not completed by nearest-center heuristics. Instance confidence q_k is the median ownership score of its members, multiplicatively calibrated by the number of independent supporting views and the minimum association margin.

Every merge, split, transfer, and abstention is written by a single writer to an **OwnershipLedger**. Each event stores before/after instance IDs, Gaussian indices, membership hashes, supporting views and masks, thresholds, and a reason code, enabling deterministic rollback and replay.

Sparse contamination and bridge guards. For each membership set, project centers onto the floor plane and compute

$$r_i^2 = (\mathbf{y}_i - \widetilde{\mathbf{y}})^\top (\widetilde{\Sigma} + \epsilon I)^{-1} (\mathbf{y}_i - \widetilde{\mathbf{y}}). \tag{21}$$

A primitive moves to $\mathcal{U}$ only when r_i^2 exceeds a validation-set quantile, its k-NN density is below τ_ρ, and at least two views provide negative ownership evidence. One audit can remove at most η_{tail} of an instance and cannot break its main connected component. Separately, adjacent instances without co-visible support cannot merge solely because their oriented bounding boxes (OBBs) intersect. OBBs are diagnostic and export summaries, never membership writers.

A.9 Targeted Supplementary Observation

This section specifies the verification interface; the results reported in this paper cover the measured shared-observation and observation-policy studies.

We expose three ambiguity types: (i) a relation score G_{ab} within δ_{rel} of its acceptance threshold; (ii) an instance with more than η_{amb} low-margin members; or (iii) nearby same-label fragments never jointly visible in a perspective view. For relation ambiguity, the controller sees only an Evidence Atlas containing candidate-specific RGB crops, masks, depth inliers, and visibility flags. It first emits verify; a verdict of uncertain, or disagreement with the geometric gates, permits a lineage-bound request_view action. The request specifies the target, missing view role and reason, and not an unconstrained camera pose.

For ambiguity a, let $\mathcal{U}_a$ denote its disputed members or instance pair and $\mathcal{V}_a^{\text{safe}}$ the cameras passing free-space, near/far-plane, and collision checks. Score a candidate by

$$S_a^{\text{sup}}(v) = \gamma_u C_{\mathcal{U}_a}(v) + \gamma_p C_{\text{pair}}(v) + \gamma_\theta A_{\text{div}}(v) - \gamma_o R_{\text{occ}}(v) - \gamma_c R_{\text{cost}}(v), \tag{22}$$

where the terms reward visibility of disputed members, simultaneous visibility of the candidate pair, and baseline/azimuth diversity, while penalizing occlusion and rendering cost. The selected camera renders RGB-D, and its new mask re-enters Equation 16 and the same association and ownership procedure. Each ambiguity receives at most B_a views and each scene at most B_{sup}. Once the budget is exhausted, unresolved relations remain separate and disputed Gaussians remain in $\mathcal{U}$.

A.10 End-to-End Algorithm, Output, and Complexity

The specified procedure is: (1) invoke clean on the BEV and validate its support; (2) invoke split_floor and cross-check XZ/YZ intervals; (3) alternate set_goal with geometry-constrained view selection via Equation 11; (4) invoke vocab and obtain perspective SAM3 masks; (5) project source Gaussians with Equation 16; (6) create and complete cores with Equation 18; (7) apply STAR-3D gates and contamination guards; (8) commit unique ownership with Equation 20; (9) the reserved branch may invoke verify and, if needed, request_view, select a supplementary camera with Equation 22, and return to step 4; and (10) accept stop only after budget and unresolved-state checks, then export native Gaussian assets and audit metadata. The evaluated mainline covers scene preparation, vocabulary formation and the geometry-grounded segmentation backend; the request–verify branch is specified through the same interface (Section 5.3).

The exporter copies only $\{\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \alpha_i, \mathbf{h}_i\}_{i \in \mathcal{M}_k}$ and does not update their values; the per-asset artifact contract and the failure handling that accompany the export are defined in Appendix D.1. Given $M = \sum_v n_v$ mask nodes and a sparse candidate-edge set E, projection cost is linear in visible Gaussian–view pairs, graph construction costs $O(|E|)$, and each view-pair Hungarian match is cubic in the number of candidates in those two views. Floor, spatial-neighborhood, and label-compatibility pruning avoid a fully connected graph.

B Extended Experimental Protocol and Results

This appendix follows the main-text order and adds what the main text only summarizes. Appendix B.1 documents how the benchmark is built (selection, annotation, quality control, statistics); Appendices B.2 and B.3 define the metrics and the protocol layers of Section 5.1, including the measured / protocol-only / reserved status of every study; Appendices B.4–B.11 report the full tables and ablations behind Sections 5.2–5.3, and Appendix B.12 specifies timing and reproducibility records, and gives the additional qualitative case study. Reserved studies are marked as reserved and are never filled with pilot numbers.

B.1 GaussAsset-100 Benchmark

Section 4 motivates the benchmark and reports its headline statistics (Table 1). This appendix documents how the benchmark is constructed: which scenes enter the pool, how cuboids are produced and audited, how box labels become membership proxies, and what the benchmark does and does not contain. For each annotated object O_k the benchmark provides a semantic category c_k and a human-verified oriented 3D cuboid, and predictions are evaluated on the *original* Gaussian primitives without converting them to another 3D representation. There is no point-level Gaussian membership ground truth because the annotation platform does not expose such a tool, so every membership metric in this paper is computed from boxes converted to proxy point labels and carries a †.

Construction and provenance. The reconstructed scenes come from a furnished-indoor capture pool. Scenes enter the benchmark only after a reconstruction and annotation usability audit. The frozen benchmark retains 100 scenes from the annotated pool, area-stratified by occupied floor area toward S/M/L/XL = 2/3/3/2, yielding S 19 / M 30 / L 30 / XL 21. All splits are at the physical-scene level.

Annotation and quality control. Annotations are oriented 3D cuboids (category, center3D, size3D, yaw) in the reconstructed scene coordinate frame, produced by human annotation, verified against the reconstruction, and mapped onto the canonical category set of Section 4. Boxes are converted to proxy point labels by testing carrier points for inclusion; nested boxes are resolved smallest-volume first, and points outside all boxes are treated as unannotated. Quality control includes coordinate-frame verification against the recorded per-box point count (for example, a bed box with pointN= 69,224) and a comparison of box-inclusion counts between the source LiDAR PCD and the Gaussian PLY consumed by the baselines. The benchmark carries no per-object occlusion label and does not annotate parts, containers or context regions; the derived masks used by the 2D evaluation layer are explicitly derived and are never described as human annotation.

Benchmark tasks. The three evaluation settings of Section 4 are implemented as the four protocols of Table 8. Shared-observation decomposition provides all methods with the same cleaned floor-level scenes, identical predefined observations and matched 2D masks, so it evaluates only the 3D instance-parsing backend and produces the primary results. The observation-policy pilot compares view-selection policies under a controlled budget without claiming a complete loop, and targeted verification, which would resolve merge, fragmentation and ownership errors from ambiguous hypotheses, remains reserved.

Reconstruction-quality audit. Before a scene enters the annotated pool it must satisfy four checks: metric scale consistent with the annotated cuboids, renderable RGB-D from the benchmark camera table, no gross unclosed reconstruction hole within the annotated floor area, and stable gravity and floor support. Scenes that fail any check are not annotated and are not included in the frozen benchmark. The audit is a usability gate rather than a cross-dataset quality comparison; the benchmark benchmark tooling supports protocol-matched PSNR/SSIM comparisons for follow-up studies.

Scene-cleaning boundary. The full workflow may exclude exterior or isolated Gaussian primitives before instance parsing. Such primitives are retained in a separate excluded set and are not treated as indoor object predictions. To prevent erroneous preprocessing from being hidden by downstream metrics, we report scene-cleaning retention and exterior-residual statistics separately whenever corresponding audit annotations are available.



Figure 5: **Overview of the GaussAsset-100 benchmark.** Representative scenes from the four pre-registered floor-area strata (S/M/L/XL) of GaussAsset-100, shown as rendered cut-away views of the reconstructed source Gaussians.

Benchmark statistics. Table 6 expands the headline statistics of Section 4 with the area-band boundaries (near 40/65/150 m^2), the density quartiles formed on annotated objects per square metre, and the full long-tailed category distribution. Scenes are split at the physical-scene level; because GaussAsset and the open-vocabulary baselines are training-free, the split is used for evaluation and not for training, whereas Chorus trains a linear head on target-domain pseudo-ground-truth and is marked as trained in the semantic comparison. Table 1 lists 69 canonical categories, while its long-tail statistics (top-10 share, categories with at most five objects, singletons) are computed over the 86 raw surface forms before canonical mapping; over the 69 canonical groups the same statistics are 58.8%, 21 and 8.

B.2 Evaluation Metrics

Coverage alone is insufficient for this task. A method can obtain high coverage by welding neighboring objects, while a pure but fragmented prediction can omit much of an object. We therefore measure both membership completeness and instance correctness.

Gaussian instance overlap. For predicted membership $\mathcal{M}_p$ and ground-truth membership $\mathcal{M}_g$, we define Gaussian IoU as

$$\mathrm{IoU}^G(\mathcal{M}_p, \mathcal{M}_g) = \frac{|\mathcal{M}_p \cap \mathcal{M}_g|}{|\mathcal{M}_p \cup \mathcal{M}_g|}. \tag{23}$$

Predicted and ground-truth instances are matched using Gaussian IoU. We report Gaussian instance AP at IoU 0.25 and 0.50 and their mean. Here G denotes overlap measured over source Gaussian primitives, while † marks the box-derived proxy ground truth used for these scores.

Asset completeness and purity. For each matched instance we report Gaussian precision and recall,

$$P_G = \frac{|\mathcal{M}_p \cap \mathcal{M}_g|}{|\mathcal{M}_p|}, \qquad R_G = \frac{|\mathcal{M}_p \cap \mathcal{M}_g|}{|\mathcal{M}_g|}, \tag{24}$$

Table 6: Statistics of the frozen audited 100-scene benchmark. Object and category counts come from the benchmark annotations; Gaussian counts are source PLY vertex counts. All listed fields come from the frozen benchmark; the split is scene-level; GaussAsset and the open-vocabulary baselines are training-free, whereas the Chorus semantic row uses its documented target-domain head.

Statistic	GaussAsset
Reconstructed Gaussian scenes	100
Object instances	3,590
Semantic categories	69 canonical
Total Gaussian primitives	203.1M
Median Gaussians / scene	1.83M (range 0.97M–5.31M)
Median objects / scene	34 (p10/p90 = 25/48)
Scene-size bands (S/M/L/XL)	19 / 30 / 30 / 21
Median occupied floor area	68.3 m^2 (range 22.0–303.2)
Top-10 category share	55.5% of instances
Categories with ≤ 5 objects	27 (10 singletons)
Raw platform surface forms	86 (mapped to 69 canonical)
Source Gaussians / annotated object	56.6k (mean)
Carrier points / annotated object	8.0k (median proxy)
Split	Scene-level; method training-free
Evaluation set	100 scenes (scene-level split)

which measure asset purity and completeness, respectively. We additionally report semantic-matched Gaussian IoU and object recall. A prediction is counted as correct only when its matched semantic identity is also correct.

Metric hierarchy and box-derived proxy labels. The benchmark mixes three levels of evidence, and we keep them separate everywhere in the paper. *Direct annotation metrics* are computed from the human-verified cuboids: all-GT axis-aligned bbox mIoU, matched oriented-box IoU, and object recall. *Box-derived proxy metrics* convert each ground-truth cuboid $\mathcal{B}_g$ to a proxy Gaussian membership

$$\mathcal{M}_g^\dagger = \{i : \boldsymbol{\mu}_i \in \mathcal{B}_g\}, \qquad \mathrm{IoU}^\dagger(p, g) = \frac{|\mathcal{M}_p \cap \mathcal{M}_g^\dagger|}{|\mathcal{M}_p \cup \mathcal{M}_g^\dagger|}, \tag{25}$$

and carry a † in every table and in the main text; they are not human point-level ground truth. When cuboids overlap, ownership is assigned smallest-volume first, and Gaussian primitives outside all annotated cuboids are treated as unannotated and not as foreground negatives. *Bridge metrics* re-evaluate published protocols on a shared RGB-D reconstruction stack and are reported in a separate table, never pooled with the two levels above.

Structural errors. We separately report **welding rate** (↓), which measures erroneous merging of multiple physical objects into one prediction, and **fragmentation rate** (↓), which measures cases in which one physical object is split into multiple substantial predictions.

Front-end quality. We separately audit the world-coordinate cleaning and floor decomposition stages. For scene cleaning, we report the retained indoor-Gaussian ratio and the residual exterior or isolated-Gaussian ratio when corresponding audit annotations are available. For floor decomposition, we report floor-boundary accuracy and cross-floor Gaussian leakage. These metrics distinguish errors introduced before observation planning from errors in the downstream parsing backend.

Observation efficiency. For observation-policy experiments, the release records the number of Cubemap, top-down, and perspective RGB-D observations, visible-Gaussian ratio, discovered-object recall, supplementary-view count, and associated rendering, VLM and SAM3 cost; this paper reports the view and quality metrics.

B.3 Experimental Protocol and Baselines

We distinguish two run types throughout the paper. A full_workflow_run starts from a fixed reconstructed Gaussian scene and regenerates BEV cleaning, floor decomposition, candidate-view planning, RGB observations, RGB-D observations, semantic evidence, instance memberships, targeted verification, and final assets. A downstream_replay reuses an existing cleaned scene, floor partition, view set,

Table 7: 3D resources referenced in this paper. Annotation units differ across resources: surface labels on meshes or points, object-level boxes, or object-centric assets. The ScanNet bridge is an appendix comparability check and not a primary claim; its scenes are a subset of the benchmark with shared-stack predictions for all compared methods. Protocols are never mixed numerically.

Benchmark	Representation	Scale	Supervision provided	Protocol and role
GaussAsset (ours)	Reconstructed furnishing-showroom 3DGS scenes	100 scenes, 3,590 objects, 203.1M Gaussians	Human-verified oriented 3D cuboids; no point-level Gaussian membership	Membership protocol; primary controlled benchmark; shared-observation results, ablations and observation study
ScanNet	Real RGB-D scans of indoor scenes	1,513 scans	Point- and mesh-level semantic and instance labels	Supervised 3D segmentation and instance benchmarks; surface ground truth, not Gaussian ownership
ScanNet++	High-fidelity real RGB-D scans	460 scenes, 280K registered DSLR images	Instance and semantic labels with high-fidelity geometry	Surface ground truth for segmentation and reconstruction; reference resource
Matterport3D	Real RGB-D panoramas with reconstructed meshes	90 buildings, 10,800 panoramas	Region and object labels on the reconstructed mesh	Supervised semantic and region understanding; reference resource
DL3DV-10K	Real capture for novel-view synthesis	10,510 videos, 51.2M frames, 65 scene types	Videos and camera poses; no object-level annotation	Novel-view synthesis benchmark; motivates the absence of object-level GT on real capture
3D-FRONT	Synthetic furnished rooms with layout	Thousands of furnished rooms	Layout and furniture models	Synthetic layout and object composition; no real capture artifacts
SceneNet RGB-D	Synthetic indoor RGB-D sequences	5M rendered frames	Semantic and instance labels from rendering	Synthetic pre-training resource; no reconstruction artifacts
PartNet	Object-centric 3D models	26,671 models, 573,585 part instances	Hierarchical part-level labels over 24 object categories	Part-level object understanding; no rooms or scenes
Objaverse	Object-centric 3D assets	800K+ objects	Object-level 3D assets with metadata	Asset generation and understanding; no indoor scenes
InteriorGS	3DGS scenes with object-level 3D box annotations; per-Gaussian labels are obtained by a separate assignment step	1,000 Gaussian scenes; official test split 001–100	Foreground semantic categories attached to 3D boxes	Foreground mIoU; external semantic-transfer check, no target training
SceneSplat	3DGS scenes with language features	SceneSplat-7K scale	Scene-level language features for open-vocabulary 3D understanding	Public bridge; representation alignment differs from our carrier protocol
LERF / LERF-OVS	3DGS scenes with language queries	Official test views	Query masks at official cameras	Query-mask mIoU and Acc@0.25/0.50; separate persistent-3D-instance analysis
ScanNet bridge	Shared RGB-D reconstruction stack prepared by the baseline authors	71 scenes with predictions from all compared methods	Point sets derived from cuboids	Greedy point-set IoU, $AP/AP_{25}/AP_{50}$, bbox mIoU; appendix comparability bridge only

or rendered observations. Unless otherwise stated, headline results use full_workflow_run; replays are used only for controlled component analysis.

Protocol layers. Table 8 records, for every protocol used in the paper, what the method receives, what it is allowed to do, which scene set is evaluated, which metrics are reported, and whether the result is measured, protocol-only, or reserved. The layers are not interchangeable: a bridge row answers whether a published protocol can be reproduced on a shared stack, whereas a native row answers what the method achieves on the benchmark Gaussian carriers. Training regimes and evaluators are also distinct: GaussAsset uses frozen components without target-domain training, the Chorus semantic row uses its documented target-domain pseudo-label head, and bridge rows retain the evaluator defined by their published protocol.

VLM usage in the reported rows. The shared-observation results use the frozen proposal head of the production stack, and the observation-policy pilot adds no VLM call because only the rendered view set changes. In the controller-role ladder only the production-reference tier calls the VLM (20.6 calls per scene on average); the lower tiers read fixed prompt tables or their frozen geometric substitutes, and the closed-loop tier is reserved.

Table 8: Protocol layers. *Native* rows share the benchmark Gaussian carrier and the frozen evaluator; *bridge* rows reproduce a published protocol on a shared RGB-D stack and are reported separately. The last column states which results are measured in this paper, protocol-only, or reserved.

Protocol	Observation source	Action set	Scene set	Metrics	Status
Shared-observation decomposition	Frozen cleaned carrier, identical rendered RGB-D for every method	None; parsing only	Frozen 100-scene benchmark	Object recall, all-GT bbox mIoU, matched OBB, membership diagnostics	Measured (primary)
Observation-policy pilot	Same initial observations plus a shared supplementary-view pool	View selection under a matched budget	Three pilot scenes	Visibility, coverage recall, $mAP^{G,\dagger}$	Measured (pilot)
Controller-role ladder	Fixed geometric executor, roles opened one at a time	clean, split_floor, discover_vocab	Frozen 100-scene benchmark	$mAP^{G,\dagger}$, $AP^{\dagger}$@50/25, visible-Gaussian ratio, VLM calls	Measured (ablation)
Closed-loop execution and verification	New observations requested from unresolved hypotheses	Adds set_goal, verify, request_view	Reserved	Coverage recall, harmful-action and abstention ledgers	Reserved
Public bridge	Shared RGB-D reconstruction stack prepared by baseline authors	Published front ends	71 shared-stack scenes	Greedy point-set AP, all-GT bbox mIoU	Measured (secondary)

Table 9: Secondary public-protocol bridge on 71 shared-stack scenes. All rows use the same greedy point-set evaluator and axis-aligned all-GT box mIoU; this table is separate from the native Gaussian-asset comparison.

Method	Pred. inst.	AP↑	AP_{25} ↑	AP_{50} ↑	bbox mIoU↑
MaskClustering	3264	10.7	60.2	31.3	58.9
SAI3D	2414	30.2	66.2	54.0	60.1
SceneSplat*	14371	9.7	35.9	19.9	48.0
GaussAsset (ours)	2476	**31.5**	**67.5**	**56.5**	**64.4**

Novelty boundary. We position GaussAsset as a pipeline for a specific combination of conditions: it consumes an existing Gaussian PLY without re-optimizing the scene, it decomposes that carrier into persistent and mutually exclusive object ownership, and it assigns semantic proposals without target-domain training under the stated class scope. Each condition is checked against the resources and baselines above: methods that optimize scene-specific features fail the first, lifting pipelines that keep only per-frame masks fail the second, and trained heads fail the third. The claim is scoped to this combination; cross-dataset results are reported under their published protocols.

Public ScanNet-protocol bridge. The following table is a secondary comparability bridge on the 71 shared-stack scenes. It uses greedy point-set matching and constant-confidence exports; its AP and box values are not compared numerically with the shared-observation asset metrics in the main text.

For parsing-backend comparisons, every method receives the *same* cleaned floor-level Gaussian scenes, the same observations, and, where appropriate, the same 2D masks. This prevents scene-cleaning and observation quality from confounding the comparison. For observation-policy comparisons, the downstream GaussAsset parsing backend is fixed, all policies use the same cleaned candidate pool, and view or rendering-cost budgets are matched.

Baselines. MaskClustering and SAI3D are evaluated as aligned protocol bridges under the shared-observation membership evaluator. Chorus is the native-Gaussian semantic comparison, while SceneSplat appears only in the public bridge because its language-feature instance interface is not a source-membership head.

Reference configuration. Per-scene records document the renderer, backend, device, camera and PLY provenance, view-selection decisions, and membership-change provenance; the frozen reference configuration is summarized in Appendix D.3.

B.4 Ownership-Rule and Single-Gate Ablations

Table 12 varies only the rule that assigns carrier Gaussians to instances; the rendered observations, SAM3 masks, and projected evidence are fixed. All rows are scored on the frozen 100-

Table 10: Public-benchmark reference results. InteriorGS baselines follow Chorus Li et al. (2026); LERF-OVS reference values follow LightSplat Bang et al. (2026). The persistent-instance protocol is a separate 3D analysis and is not directly comparable to query-mask mIoU.

Benchmark	Method / protocol	Metric	Value
InteriorGS foreground semantics			
InteriorGS	Mosaic3D	fine-grained mIoU	9.4
	SceneSplat Li et al. (2025)	fine-grained mIoU	10.0
	Chorus Li et al. (2026)	fine-grained mIoU	**15.7**
	GaussAsset (novel-view RGB-D; no target training)	fine-grained mIoU	11.82
LERF-OVS official query scoring			
LERF-OVS	LightSplat (official protocol)	mIoU / Acc@0.25	**47.58 / 68.32**
	Dr.Splat (official protocol)	mIoU / Acc@0.25	43.58 / 63.87
	OpenGaussian (official protocol)	mIoU / Acc@0.25	42.15 / 56.22
	GaussAsset (novel-view 3D prediction)	mIoU / Acc@0.25	38.46 / 57.69
LERF-OVS persistent 3D instances			
LERF-OVS	GaussAsset (fully automatic)	scene-macro instance IoU	52.80

Table 11: Detailed membership diagnostics on the frozen 100-scene benchmark. The class-agnostic bridge rows use the same evaluator as the main comparison; mask IoU is box-derived and marked with †. Coverage recall R_{cov}@0.5 is a membership quantity and is distinct from the box-layer object recall of Table 2.

Method	cov† ↑	$R^{\dagger}_{\text{cov}}$@0.5↑	purity† ↑	weld† ↓	frag† ↓	mask IoU† ↑	bbox mIoU↑
Frozen 100-scene benchmark							
MaskClustering	44.1	49.3	92.1	0.8	14.8	41.5	47.6
SAI3D	53.1	64.4	94.9	6.9	5.0	52.3	51.4
GaussAsset (ours, production rule)	**57.6**	**70.3**	**95.6**	1.7	7.7	**55.4**	**55.4**

scene benchmark by the same evaluator at the production gates (assignment = 0.50, span = 0.10, low purity = 0.70). The *A1* rows replace the production membership: fine stops before identity consolidation, label_merge unions instances that share a canonical label, and geometry_only_d groups carrier points by d-metre voxel connectivity with no semantics and no cross-view evidence. The *A2* rows disable exactly one ownership gate and keep everything else unchanged, so each row is the counterfactual of accepting that gate's removals. The last two *A2* rows are distinct recorded states that coincide on every membership-level metric and differ only in their fitted boxes. Box columns refit one Z-up box per membership with an identical fitter in every row, which measures the geometry implied by each rule; the exported-box operating point of Section 5.2 (63.7/60.0/75.9 for the production rule) is reported in the main text. Box IoU follows the matched oriented-box convention rather than the axis-aligned all-GT bbox mIoU, and purity is averaged over scenes with at least one assigned prediction.

Among the evaluated rows, the production rule pairs the highest coverage with sub-2% welding. Disabling a single gate raises welding from 1.69% to 2.7–7.6% for most components; stray quarantine and cabinet-neighborhood merging instead reduce coverage by 2.4 points without a welding benefit. The rule-level alternatives are more damaging: stopping before identity consolidation raises welding to 7.35%, naive same-label union reduces coverage to 28.6% and raises welding to 29.64%, and class-agnostic voxel states are rejected by the assignment gate. Across the ablations the production rule retains the strongest coverage–exclusivity balance.

Assignment-gate sensitivity. Table 13 re-scores three rules at assignment gates 0.50 (production), 0.35, 0.25 and 0.00 with everything else frozen. For the production rule a looser gate trades purity for coverage (95.6% → 94.6% purity and 57.6% → 58.5% coverage from 0.50 to 0.25), which is

Table 12: Ownership-rule (*A1*) and single-gate (*A2*) ablations on the frozen 100-scene benchmark; only the membership rule changes. Rows marked "–" are class-agnostic negative controls rejected by the assignment gate, and paired bootstrap and Wilcoxon statistics are in Table 14.

Membership rule	Cov.† ↑	Purity† ↑	Weld.† ↓	Frag.† ↓	$PQ^{\dagger}_{50}$ ↑	mask IoU† ↑	box IoU↑
A1: membership rules							
final (production)	**57.6**	**95.6**	**1.69**	**7.72**	51.6	55.4	73.4
fine (no identity consolidation)	57.4	94.0	7.35	10.06	54.7	54.0	67.2
label_merge (naive semantic union)	28.6	85.2	29.64	**2.59**	25.3	34.3	45.0
geometry_only_0.10 m	0.0	97.2	0.84	4.12	0.0	–	–
geometry_only_0.15 m	0.0	94.6	2.64	0.33	0.0	–	–
geometry_only_0.25 m	0.0	80.8	9.07	0.00	0.0	–	–
A2: single-gate removals (recorded snapshots)							
identity consolidation	57.5	93.9	7.44	9.86	54.7	54.1	67.1
general ownership arbitration	57.5	93.9	7.56	9.16	54.8	54.1	67.3
ownership closure	56.3	95.0	2.71	8.94	52.4	54.0	69.9
container ownership	57.5	93.9	7.50	9.22	54.7	54.1	67.2
attached-part ownership	56.5	94.8	5.06	8.97	52.8	54.0	69.9
seating consolidation	57.5	93.9	7.46	9.28	54.6	54.1	67.2
stray quarantine	55.2	95.5	1.65	6.55	51.0	53.2	72.7
cabinet neighborhood merge	55.2	95.5	1.65	6.55	51.0	53.2	72.3

Table 13: Assignment-gate sensitivity on the frozen 100-scene benchmark. Only the gate changes; states, observations, GT and evaluator are identical. Gate 0.00 is a degenerate control (see text).

Gate	Rule	Cov.† ↑	Purity† ↑	Weld.† ↓	Frag.† ↓
0.50	final (production)	57.6	95.6	1.69	7.72
0.35	final	58.4	95.0	1.69	7.72
0.25	final	58.5	94.6	1.69	7.72
0.00	final (degenerate)	2.2	89.0	1.69	7.72
0.50	label_merge	28.6	85.2	29.64	2.59
0.25	label_merge	34.7	77.9	29.64	2.59
0.50	geometry_only_0.15 m	0.0	–	2.64	0.33
0.25	geometry_only_0.15 m	0.0	–	2.64	0.33

the conservative operating point we report. The class-agnostic voxel states stay at zero coverage even at 0.25: their rejection is not an artefact of the gate threshold but of cross-object clusters failing the purity requirement at every admissible setting, so they remain negative controls. Gate 0.00 is a labeled degenerate control: with purity ≥ 0 every prediction becomes assignable, the tightest-match rule hands predictions to the smallest ground-truth box, and coverage collapses for *every* rule, so the row must not be read as a model result.

The sweep also shows that the ablation ordering in Table 12 is stable under a ± 0.25 change of the assignment gate for the rules where a sweep is defined: the production rule keeps its coverage advantage and its sub-2% welding, and the naive same-label union keeps its 29.6% welding.

B.5 Paired Statistics and Benchmark Robustness

The means in Table 12 are per-scene paired differences against the production rule on the same 100 scenes. We resample the 100 scenes with replacement 10,000 times (bootstrap 95% confidence intervals) and apply a two-sided Wilcoxon signed-rank test; no per-scene tuning or scene removal is involved. Table 14 reports the coverage and welding contrasts under the same pooled definition as Table 12. The two ordering claims survive the bootstrap analysis: stopping before identity consolidation leaves coverage statistically unchanged (-0.2 points, $p = 0.44$) while welding rises by 5.7 points ($p < 2 \times 10^{-4}$), and the naive same-label union loses 29.0 coverage points and gains 28.0 welding points ($p < 2 \times 10^{-4}$). The stray_quarantine and cabinet_neighborhood rows lose 2.4 coverage points ($p < 2 \times 10^{-4}$) while their welding change is statistically indistinguishable from zero ($p = 0.76$), which is why they are reported as the coverage-cost counterfactuals of the production rule.

The completed ownership and gate ablations are reported in the ownership-rule, gate-sensitivity and paired-statistics tables above. Association-backend, full observation-budget, non-chair five-factor and risk-routed verification studies are documented as protocols with their measurement status in Table 8.

Table 14: Paired statistics for the ownership ablation on the frozen benchmark. Δ columns are pooled percentage-point differences against the production rule (coverage averaged over scenes, welding pooled over predictions); intervals and p are from 10,000 bootstrap resamples of the 100 scenes. Geometry-only rows are degenerate negative controls.

Rule (vs. production)	$\Delta\text{cov}^{\dagger}$	95% CI	p	$\Delta\text{weld}^{\dagger}$	95% CI	p
fine (no identity consolidation)	−0.2	[−0.8, +0.4]	0.44	+5.7	[+4.9, +6.4]	$< 2 \times 10^{-4}$
label_merge (naive union)	−29.0	[−31.2, −26.8]	$< 2 \times 10^{-4}$	+28.0	[+26.0, +29.9]	$< 2 \times 10^{-4}$
geometry_only_0.10 m	−57.6	[−58.8, −56.4]	$< 2 \times 10^{-4}$	−0.8	[−1.3, −0.4]	$< 2 \times 10^{-4}$
geometry_only_0.15 m	−57.6	[−58.8, −56.4]	$< 2 \times 10^{-4}$	+1.0	[+0.4, +1.6]	0.003
geometry_only_0.25 m	−57.6	[−58.8, −56.4]	$< 2 \times 10^{-4}$	+7.4	[+5.4, +9.9]	$< 2 \times 10^{-4}$
identity consolidation	−0.1	[−0.6, +0.4]	0.78	+5.8	[+5.0, +6.5]	$< 2 \times 10^{-4}$
general ownership arbitration	−0.1	[−0.7, +0.4]	0.64	+5.9	[+5.1, +6.6]	$< 2 \times 10^{-4}$
ownership closure	−1.3	[−1.8, −0.8]	$< 2 \times 10^{-4}$	+1.0	[+0.6, +1.4]	$< 2 \times 10^{-4}$
container ownership	−0.1	[−0.7, +0.4]	0.74	+5.8	[+5.1, +6.6]	$< 2 \times 10^{-4}$
attached-part ownership	−1.1	[−1.6, −0.6]	$< 2 \times 10^{-4}$	+3.4	[+2.8, +4.0]	$< 2 \times 10^{-4}$
seating consolidation	−0.1	[−0.7, +0.4]	0.73	+5.8	[+5.0, +6.5]	$< 2 \times 10^{-4}$
stray quarantine	−2.4	[−2.9, −1.8]	$< 2 \times 10^{-4}$	0.0	[−0.3, +0.2]	0.76
cabinet neighborhood merge	−2.4	[−2.9, −1.8]	$< 2 \times 10^{-4}$	0.0	[−0.3, +0.2]	0.77

Table 15: Controller-role diagnostics for the main-text ladder, on the same benchmark runs. The table reports the benchmark evaluation of the four configurations under the frozen protocol. All object-level quantities are computed against box-derived proxy labels (†); box IoU is the direct annotation metric. Coverage-recall and harmful-action/abstention verification columns remain reserved and are described in the surrounding text.

Benchmark membership and box metrics (frozen protocol)

Configuration	$\text{cov}^{\dagger} \uparrow$	$R^{\dagger}_{50} \uparrow$	$\text{purity}^{\dagger} \uparrow$	$\text{weld}^{\dagger} \downarrow$	$\text{frag}^{\dagger} \downarrow$	box IoU↑	point $\text{F1}^{\dagger} \uparrow$
Geometry only	13.8	16.4	95.1	1.9	2.2	59.4	28.5
Vocabulary only	52.9	63.5	94.6	1.3	14.5	62.8	60.0
+ clean/split_floor	58.7	71.4	94.7	1.4	15.1	66.6	64.2
Production reference	54.8	66.6	95.3	1.8	9.5	68.5	64.0

B.6 Shared VLM Controller: Full-Benchmark Diagnostics

We hold the geometric executor, camera candidate pool, SAM3 masks, and budgets fixed and progressively enable the controller roles. This comparison separates the benefit of using a strong VLM from the benefit of organizing the full workflow as one stateful control loop. The ladder is reported once, in the main text, and is measured on the benchmark. Table 15 reports the membership and box diagnostics for the same four configurations and the same runs; reserved verification columns are described in prose.

The frozen executor implements the scene-preparation and vocabulary roles; the goal, verification and view-request actions are specified through the same interface and are listed with their measurement status in Table 8. Table 15 reports the benchmark coverage, welding and fragmentation, box and point metrics; coverage-recall and harmful-action/abstention verification remain reserved. The main-text ladder and this table come from the same runs, so the two are read as one ranking.

B.7 Risk-Routed Verification: Reserved Protocol

Risk-routed verification is reserved for a rerun on the 100-scene benchmark. The pilot scan and capture study used scenes outside the frozen benchmark, so its decision counts are not reported here. The protocol remains defined by the same offline ledger and evaluator, with no membership writer; it is future work rather than an empty numerical result table.

B.8 Detailed Fixed-View Diagnostics

The main-text table is the single headline comparison. This appendix expands it with coverage, $R_{\text{cov}}@0.5$, purity, welding, fragmentation, proxy mask IoU, and fitted OBB quality in Table 11, followed by area and density stratification. Observation-policy and verification studies use the same frozen parsing backend and are reported separately below.

Table 16: Stratified results on the frozen benchmark. Bands are the pre-registered floor-area strata (S/M/L/XL); density quartiles are formed on annotated objects per square metre. Occlusion strata are not reported because the annotations carry no occlusion label.

Stratum	cov† ↑	$R^{\dagger}_{\text{cov}}$@0.5↑	purity† ↑	weld† ↓	frag† ↓	box IoU↑
S	56.4	70.4	96.4	1.97	6.83	72.1
M	57.5	71.2	95.0	2.08	5.76	75.3
L	56.1	67.7	95.1	1.90	8.59	73.7
XL	**61.0**	**72.8**	**96.6**	**1.06**	9.22	**76.8**
Density Q1 (low)	**61.1**	**73.3**	96.2	**0.91**	8.49	**76.1**
Density Q2	54.9	66.9	95.9	2.23	8.60	75.8
Density Q3	57.6	71.2	95.6	1.83	6.62	73.3
Density Q4 (high)	56.7	69.9	94.9	2.20	**6.45**	71.7

Table 16 shows no scene-complexity failure mode: the largest scenes reach the best coverage, purity, and box IoU, and the high-density quartile trades a small amount of purity and box IoU for the lowest fragmentation. Scene composition is therefore recorded alongside the headline operating point, not collapsed into a single average.

B.9 Observation-Policy Pilot

The six-policy comparison is reported as main-text Table 4; this subsection records the protocol behind it. We isolate observation while holding mask generation, projection, association and ownership fixed: every policy starts from the same cleaned floor scenes and candidate pool, uses the same GaussAsset parsing backend and receives a matched view budget. Uniform and random policies resample the perspective camera subset without querying the controller and keep the Cubemap and blindspot streams; the Cubemap rows change the stream composition and are cost references at their actual view counts (50 views for Cubemap only vs. 676 otherwise); geometry-only coverage uses the first three terms of Equation 11 without the interest terms; and the full three-stream policy is the production selector. This comparison is a view-selection pilot, not a completed goal-conditioned active-execution evaluation.

Read against each other, the rows separate two objectives. Detection quality is close between uniform resampling and the production selector, whereas coverage recall separates them, and the Cubemap-only row shows that cheap views still leave much of the object set unseen. Geometry-only coverage reproduces most of the full three-stream operating point, so on this pilot the 4:1 interest prior is not the dominant lever. With three pilot scenes and an unmatched Cubemap-only view count, this is a cost-aware observation study rather than a significance claim, and a quality–observation curve that varies the rendering budget remains future work. Targeted verification is evaluated separately in Section B.11.

B.10 Parsing Ablations under a Fixed View Set

A controlled parsing-backend ablation under identical observations and masks is specified as a follow-up protocol (Table 8); it requires per-point frame-count and morphology artifacts beyond the frozen benchmark. The parsing comparisons reported in this paper are the membership-protocol comparison (Table 2), the ownership-rule and single-gate ablations (Tables 12 and 13), and the controller-role ladder (Table 15).

B.11 Targeted Supplementary Observation: Reserved Protocol

The interface for this study is specified in Appendix A.9; what follows is its protocol and reserved result table. We evaluate re-observation on an ambiguity subset frozen *before* supplementary rendering. Under a matched additional-view budget, we compare no supplementation, random safe views, geometry-only coverage views, and controller-requested views whose target and missing evidence role are recorded before rendering. The targeted study is reserved for a rerun on the benchmark; the pilot used scenes outside the frozen benchmark and its decision-level counts are therefore not quoted. The full-benchmark decision table will be produced only when that protocol is run under the same frozen unary schema.

B.12 Efficiency, Reproducibility, and Qualitative Analysis

For each full_workflow_run, the release records end-to-end wall-clock time and its decomposition into BEV construction and scene cleaning, floor decomposition, candidate-view planning and Gaussian rendering, Cubemap and top-down observations, perspective SAM3 inference, depth projection and association, semantic verification and ownership refinement, supplementary observation, and final asset export. It also records observation counts, VLM call/token statistics and the fraction of total cost attributable to targeted verification. SAM3 image embeddings are computed once per perspective image and reused across scene-level prompts.

The workflow records camera and PLY hashes, metric-depth definitions, view-selection decisions, and the provenance of membership changes.

No paper-level wall-clock table is reported: the available timing pilot used a different resource profile and included scenes outside the frozen benchmark. Stage-level timing, resource details and reproducibility records remain part of the release artifacts rather than submission claims.

Qualitative results. The qualitative ownership case is provided in Appendix B.13; it complements the shared-observation quantitative comparison without introducing a separate metric.

B.13 Additional Qualitative Results

The instance-state case study of the main-text Figure 3 shows the two adjacent armchairs in one representative furnished room at three membership stages. The top row renders the original RGB Gaussians and marks, as holes, points whose ownership is wrong at that stage; the bottom row colors the two instance claims (blue: first instance; orange: second instance) and marks the same wrong-ownership points in red. Left to right: the initial fine membership (19,325 wrong-ownership points, of which 16,465 are in core clusters and 2,860 in low-support clusters), after general ownership arbitration (20,631), and the supplement-refined completion (0). The third panel uses the recorded supplementary-evidence refinement for this qualitative case study; the quantitative claims in the main text use the frozen shared-observation evaluation.

C Reproducibility and Responsible Research Statements

AI use statement

We used generative-AI tools to assist Chinese–English translation, manuscript organization and readability editing, methodology framing and completeness checks, formulation of the typed controller interface, mathematical notation, LaTeX formatting, and experimental-design and reproducibility checklists. We did not use generative AI to create, alter, or interpret experimental data or quantitative results. The authors reviewed all AI-assisted material and remain responsible for the algorithmic definitions, citations, experimental setup, and final claims. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

Ethics and societal impact statement

Indoor reconstructions may inadvertently contain identifiable people, private objects, or location information. Collection and release must be authorized, personally identifying content must be removed before reconstruction, and all source-data licenses must be respected. Open-vocabulary models may systematically miss long-tail, culturally specific, or atypical objects. Accordingly, this system should not be used for security, personnel monitoring, or other high-stakes decisions without human review. The reversible ledger and explicit abstention state support auditing but do not eliminate these risks.

Dataset availability and release

We will release the GaussAsset implementation and GaussAsset-100, including the original Gaussian scenes, calibrated novel views, human-verified object cuboids, scene-level splits, the evaluation protocol, and scripts for reproducing the reported results. The accompanying documentation specifies

the data processing procedures and evaluation configurations. Data release will follow the privacy provisions described in the ethics statement.

Reproducibility statement

We organize every experiment as a versioned, auditable workflow. The planned release will include scene-level splits; exact renderer, SAM3, and VLM versions; coordinate and depth conventions; thresholds, weights, seeds, and camera configurations; controller interfaces; action histories, evidence IDs, rejected actions, and token counts; cleaning masks, floor intervals, view traces, mask records, and OwnershipLedgers; baseline adapters; and a single-command path from source scenes to tables for the released components. The VLM prompt suite (controller, discriminator and risk-evaluation prompts, and the JSON structured-output instructions) and the prompting vocabulary are not part of the release, and neither are the model reasoning traces; a short usage note describes how to connect a compatible VLM. A full-workflow run recomputes cleaning, views, and masks from the source scene. A downstream replay must expose the hashes of every reused artifact. The anonymous supplementary package contains evaluation code, frozen configurations, benchmark metadata, annotation examples and preview scene samples; the complete 100-scene benchmark will be released publicly.

D Implementation, Output, and Pipeline Details

D.1 Output Definition and Failure Handling

Each scene asset contains one PLY per instance together with scene-level metadata. The asset contract is: semantic label, source-PLY hash, ordered Gaussian indices, confidence, supporting views, and OBB, with the exporter copying only the original attributes $\{\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \alpha_i, \mathbf{h}_i\}_{i \in \mathcal{M}_k}$. The metadata records class, confidence, source-index hash, supporting views, OBB, ownership version, and unresolved warnings. Structural classes may be exported as a separate background layer but are not countable objects. A Gaussian may not occur in two instance PLY files; the unassigned and scene-cleaning exclusion sets are exported separately. A geometrically stable instance with uncertain semantics receives unknown-object. Invalid coordinates, intersecting floor intervals, inconsistent depth conventions, or changed source hashes are scene-level hard failures and never produce a silently complete-looking output.

D.2 Pipeline Stage Mapping

Table 17 maps the scientific components to persistent artifacts. Every stage consumes only the declared output of its predecessor; all arrays and indices carry source-scene hashes and coordinate conventions.

Table 17: Mapping from scientific components to auditable artifacts.

Component	Primary input	Persistent output	Invariant
VLM controller	Visual packet, scene state, history	Typed action, evidence IDs, action trace	No direct membership writes
Scene cleaning	Source PLY, BEV, clean	Prompts, support mask, exclusion set	Retained attributes unchanged
Floor decomposition	XZ/YZ maps, split_floor	Floor intervals, boundary set, consistency status	Non-overlapping floors
Observation planning	Safe candidates, set_goal, budget	Selected cameras, coverage/goal trace, RGB(-D)	Executor validates every pose
Projection	RGB-D, cameras, vocabulary, SAM3	Masks, residuals, visibility, carrier provenance	Metric inliers only
Arbitration	Carriers and provenance	Frozen cores, cannot-link edges, relation margins	Weak edges cannot rewrite cores
Ownership/export	Arbitrated graph, support tensor	Disjoint memberships, unassigned set, ledger, assets	At most one owner per Gaussian

D.3 Reference Configuration and Annotation-Free Evaluation

Reference configuration. The reference configuration fixes the metric-depth convention, projection gate $|z - D| \leq 0.06 + 0.02D$, association gates and ownership support recorded in Appendix A.7 and Appendix A.8. Model and runtime versions, decoding settings and resource details are recorded in the release for exact reproduction; per-scene records remain authoritative for scene-specific counts.

Annotation-free evaluation. The pipeline starts from a pretrained 3DGS representation and renders its own novel views; reconstruction-time RGB frames are not consumed after reconstruction. No target-dataset semantic or instance labels are used for training, prompt construction, view selection, association or verification. Prompts, decoding parameters, geometric thresholds and stopping criteria are frozen before evaluation. For semantic evaluation, we report both scene-macro metrics, which average per-scene scores, and pooled metrics, which aggregate foreground-point statistics across the evaluation split, as summarized in Table 3.