PVSR: SINGLE-IMAGE 3D TABLETOP RECONSTRUCTION WITH PHYSICAL STABILITY AND SEMANTIC PRESERVATION

Anonymous authors
Paper under double-blind review

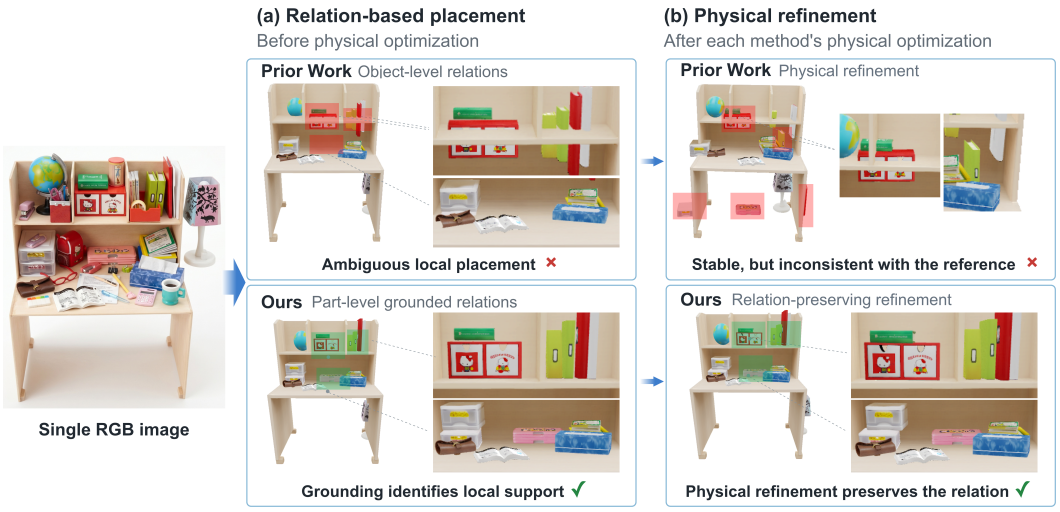


Figure 1: Given a single RGB image, PVSR reconstructs a tabletop scene while keeping each object on its intended local support. (a) Relation-based placement. Object-level relations leave the supporting surface unspecified, while grounding them in part-level geometry identifies the intended surface. (b) Physical refinement. Refinement without the grounded relations settles objects into stable but reference-inconsistent configurations, while our relation-preserving refinement retains the intended placement. Insets highlight the local structures involved.

ABSTRACT

Reconstructing a 3D tabletop scene from a single image remains challenging, as errors in object assets and pose estimates can cause misplaced supports, floating objects, and interpenetration. Object-level spatial relations capture how objects relate to each other, but they do not identify the local geometric regions on complex objects that realize these relations, leaving pose refinement without explicit geometric targets. We present PVSR (Part-level Visual-grounded Spatial Relations), a framework that grounds object-level spatial relations in local geometry and solves them as layout constraints for single-image tabletop reconstruction. PVSR first constructs an object-level SceneTree from the reference image and extracts addressable local geometric regions from the reconstructed assets. It then grounds each relation edge to the geometric regions on the child and parent objects that participate in the relation and formulates relation-conditioned layout constraints to refine the initial poses. We further introduce a relation-aware physical optimization module that leverages these grounded relations to resolve collisions and improve scene stability. We evaluate PVSR on the TabletopGen (TTG) benchmark and on PartRelBench, our proposed benchmark for part-level spatial relation modeling, covering both end-to-end reconstruction and controlled relation evaluation. Experiments show that PVSR produces more accurate relational layouts in structurally complex scenes, improves physical stability in simulation, and better preserves visual consistency with the reference image.

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.

1 INTRODUCTION

Reconstructing a complete 3D tabletop scene from a single RGB image is an important problem for embodied intelligence, robotic interaction, and 3D content creation. Beyond recovering the objects, the task requires estimating their spatial transformations, support structure, and mutual layout. Compositional systems that reconstruct or retrieve individual assets and assemble them into a shared 3D scene (Liu et al., 2025; Dai et al., 2025; Huang et al., 2025; Wang et al., 2025; Wu et al., 2025; Yao et al., 2025; Lee et al., 2026; Ma et al., 2026; Tian et al., 2026; Chen et al., 2026) dominate this task, yet the assembled scene is often neither structurally faithful nor physically valid. These systems start from instance masks, reconstruct or retrieve an asset per object, and estimate an initial pose (Ardelean et al., 2025; Dai et al., 2025; Wang et al., 2025; Wu et al., 2025; Yao et al., 2025; Ma et al., 2026; Tian et al., 2026). Occlusion, geometric ambiguity, and asset reconstruction errors can produce misplaced or interpenetrating objects (Wang et al., 2025; Wu et al., 2025; Yao et al., 2025; Lee et al., 2026; Ma et al., 2026). As shown in Figure 1(a), the drawer unit intersects the upper shelf that should support it. For furniture with multiple support surfaces, an object-level predicate such as $\mathrm{ON}(A, B)$ identifies the supporting object but leaves the intended local surface unspecified. Correcting such placement errors requires grounding the support relation to that specific surface and aligning the object with it. We call the local geometry that realizes a spatial relation a local geometric interface.

Recent work covers either side of this gap. Part-aware asset reconstruction (Liu et al., 2024) recovers part structure inside each asset but does not use it for assembly-time alignment. Relation-graph assembly and geometry-based support optimization (Bai et al., 2026; Wu et al., 2025) exploit explicit relations, but their graphs, annotations, or support surfaces are assumed or selected by pose proximity rather than recovered from the image. When assets lack part annotations, initial poses are inaccurate, and parents offer several candidate local structures, no existing mechanism grounds an image-observed relation to its intended local geometric interface.

These local correspondences must also be preserved during physical refinement. Existing methods improve physical validity through collision repair, layout optimization, and simulation (Dai et al., 2025; Wu et al., 2025; Yao et al., 2025; Lee et al., 2026; Li et al., 2026; Ma et al., 2026). However, physical stability alone does not ensure consistency with the image-indicated layout. Figure 1(b) illustrates this issue: after physical refinement, the pencil case ends up below the desktop, losing its intended support relation. Such a configuration may be stable while remaining inconsistent with the reference image. Single-image tabletop reconstruction therefore faces two coupled challenges: grounding image-observed relations to the intended local geometric interfaces, and preserving these correspondences during physical refinement.

To address both challenges, we introduce PVSR (Part-level Visual-grounded Spatial Relations), a framework that grounds object-level spatial relations in local geometry and solves them as layout constraints for single-image tabletop reconstruction. Without relying on given relation graphs or part-annotated assets, PVSR constructs an object-level SceneTree from the input image and parses each reconstructed asset into addressable local regions. A vision-language agent then grounds every tree edge in the geometric interface through which the child and parent interact, and these grounded relations yield layout constraints that correct the initial poses.

During physical refinement, each grounded relation defines an admissible pose space, which lets us distinguish immediate geometric infeasibility from dynamic instability. We repair the former within the relation constraint and resolve the latter through tree-factorized ancestor-chain searches that sample only the child's pose while evaluating dynamics across the full support chain. Part-level relations therefore act as persistent structural constraints throughout grounding, layout solving, and physical refinement.

We evaluate PVSR on the established TabletopGen benchmark (Wang et al., 2025) and introduce PartRelBench, a benchmark suite for studying tabletop scenes in which local object structure is critical to reconstruction. PartRelBench comprises a Free Track of 50 single-image scenes (OurBench-50) that spans both common layouts and recurring challenging configurations, and a Controlled Track (MiniBench) that provides ground-truth masks and assets to reduce interference from upstream reconstruction errors. Together, the two tracks let us compare methods on conventional and structurally demanding scenes and reveal where object-level layout reasoning falls short.

Our main contributions are as follows.

- We identify the under-specification of object-level spatial relations as a key obstacle to pose correction in single-image 3D tabletop reconstruction and introduce an automatic image-to-interface grounding and solving framework.

- We introduce relation-preserving physical refinement, where grounded part-level relations define admissible pose spaces that stay active throughout optimization, separating immediate geometric violations from dynamic instability via relation-conditioned repair and tree-factorized ancestor-chain searches.
- We introduce PartRelBench, combining a 50-scene Free Track with a Controlled Track built on ground-truth masks and assets, for comparing reconstruction methods in scenes where local object structure matters.

2 Related Work

Image-Based 3D Scene Reconstruction Retrieval-based reconstruction matches and aligns existing CAD assets to image observations (Izadinia et al., 2017; Gümeli et al., 2022; Dai et al., 2025; Wu et al., 2025), with fidelity constrained by asset-library coverage. Generative approaches instead recover objects with pretrained priors, either in modular pipelines that combine per-object reconstruction with scene-level layout cues (Zhou et al., 2024; Ardelean et al., 2025; Han et al., 2025; Yao et al., 2025) or in joint models that capture inter-object interactions within the reconstruction process (Huang et al., 2025; Meng et al., 2026), and both produce compositional scenes with separate assets and spatial layouts. Spatial relations further guide layout solving. Diorama and CAST translate support and contact relations into geometric constraints (Wu et al., 2025; Yao et al., 2025), TabletopGen infers stacking relations from the reference image to determine object heights (Wang et al., 2025), and support trees organize physical optimization and assembly (Ma et al., 2026; Li et al., 2026). However, object-level relations alone do not specify local placement within complex objects. Placing an object on the top or a middle shelf of the same shelving unit satisfies the same support relation yet yields distinct positions and occlusion patterns, and contact, non-penetration, and stability alone cannot determine which layout matches the reference image. Identifying the parts and local geometric regions involved in a relation therefore requires part-level relation grounding, which we develop for single-image scene reconstruction.

Relation- and Physics-Guided Scene Modeling Relation-guided synthesis connects semantic intent to geometric layouts through explicit intermediate representations, from spatial constraints solved by search or differentiable optimization (Yang et al., 2024b; Sun et al., 2025) to semantic graph priors decoded into layouts (Lin & Mu, 2024). Scenethesis combines hierarchical scene graphs, visual alignment, and contact constraints for object placement (Ling et al., 2025), PARSE translates inter-part surface alignment and contact into feasible-pose constraints with a part-aware solver over part-annotated assets (Bai et al., 2026), and Diorama extracts local support surfaces and selects interaction regions by geometric proximity under the initial poses (Wu et al., 2025). Physical constraints further guide generation beyond relational objectives. PhyScene adds collision and reachability constraints to scene diffusion (Yang et al., 2024a), PAT3D improves physical validity and semantic consistency with differentiable simulation (Lin et al., 2025), physics-augmented predicate solvers refine configurations from simulation feedback (Wang et al., 2026), and support-aware optimization and diagnostic simulation improve stability and correct layouts in single-image reconstruction (Ma et al., 2026; Li et al., 2026; Lee et al., 2026). Despite these advances, a key challenge remains in single-image reconstruction: automatically grounding visual relations at the part level and preserving them during physical refinement. PVSR grounds image-observed spatial relations to part-level geometric interfaces on reconstructed assets, which guide layout solving and constrain pose adjustment during physical optimization.

3 Method

Given a single RGB image, our goal is to reconstruct a complete tabletop scene comprising 3D object assets and their spatial transformations. The reconstruction should preserve the visual content and spatial relations in the reference image and remain physically stable in simulation.

The central representation of PVSR is an image-grounded part-level relation, which binds a spatial relation to the local geometry that realizes it on the participating objects. This representation plays two roles in our framework. It provides explicit targets for layout correction, and it constrains pose changes during physical refinement so that stability does not come at the cost of the intended placement. As illustrated in Figure 2, we first initialize object geometry and poses (Section 3.1), then ground and solve part-level relations (Section 3.2), and finally perform relation-preserving physical refinement (Section 3.3) to improve geometric feasibility and dynamic stability.

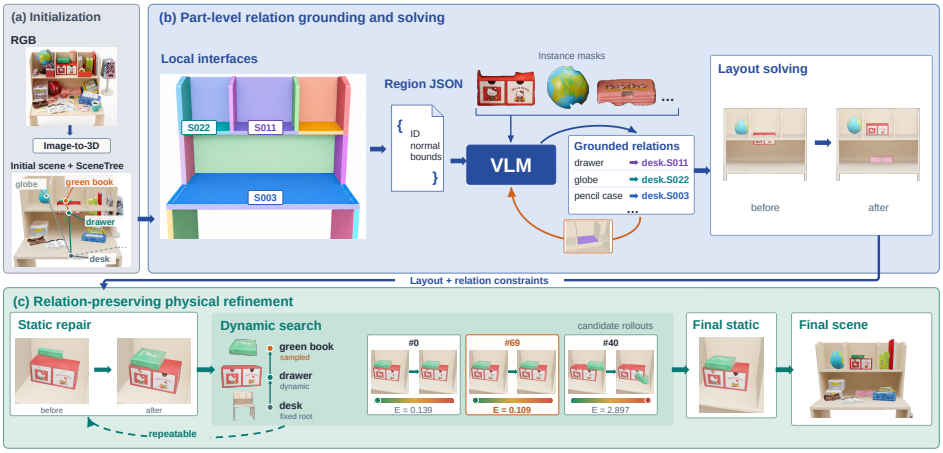


Figure 2: Overview of PVSR. **(a) Scene initialization.** We reconstruct object assets, estimate their initial poses, and organize object-level relations into a SceneTree. **(b) Part-level relation grounding and solving.** A vision-language agent grounds each tree relation to part-level interfaces, and the grounded interfaces define geometric constraints for layout correction. **(c) Relation-preserving physical refinement.** We alternate geometric repair and simulation-based dynamic search under the grounded relation constraints to resolve collisions and instability.

3.1 Scene Initialization

Given a reference image I, we follow the instance-level scene decomposition procedure of REST3D (Ma et al., 2026) to obtain object masks $\{M_i\}_{i=1}^N$, where N is the number of objects. SAM3D reconstructs an object mesh G_i from the image and the corresponding mask and estimates its initial pose T_i^0 and scale. After incorporating the estimated scale into each mesh, we represent the rigid pose of object i as $T_i = (R_i, \mathbf{t}_i) \in \mathrm{SE}(3)$, where R_i and $\mathbf{t}_i$ denote rotation and translation. In parallel, a vision-language model constructs an object-level SceneTree $\mathcal{T} = (V, E)$ from the reference image and the masks. Each edge $e = (c, p)$ identifies a child–parent pair that requires a spatial relation. The initial scene $S^0 = \{(G_i, T_i^0)\}_{i=1}^N$, with all poses expressed in a common scene coordinate system, serves as the input to relation grounding.

3.2 Part-Level Relation Grounding and Layout Solving

The SceneTree specifies which child–parent pairs require relations, but the meshes from Section 3.1 carry no part annotations, so the local structures that could realize these relations are not yet explicit. This subsection makes them explicit in three steps. We first parse each mesh into addressable local geometry (Section 3.2.1). A vision-language agent then grounds each image-observed relation in that geometry (Section 3.2.2). Finally, the grounded relations are solved as layout constraints that correct the initial poses (Section 3.2.3).

3.2.1 Addressable Local Geometry

The same object can contain several local regions that serve different geometric roles. Shelves at different heights, for example, belong to a single asset but offer distinct placement locations. Two common representations fail to capture this structure. Individual mesh triangles are too fine-grained to serve as useful references, while a whole-object bounding box erases these distinctions entirely. We therefore simplify each mesh, cluster triangles based on face adjacency, normal agreement, and coplanar distance, and merge compatible planar patches that are spatially adjacent. This parsing yields a set of local geometric candidates, each carrying an explicit identifier.

The identifier is what turns a geometric region into a parameter that both the agent and the solver can operate on. Each candidate is stored with its position, normal, boundary, local coordinate frame, and adjacency to other regions. These attributes allow the agent to compare the height and orientation of regions and their structural relationships, and they supply the computable geometry required for pose solving in Section 3.2.3. Since identifiers alone are hard to localize on a raw mesh, we additionally render a numbered visual index of the 9 largest visible regions from a default viewpoint, which allows the agent to refer to regions by number. The full candidate set remains available in the structured record, so the agent is not restricted to the indexed regions.

3.2.2 VISUAL GROUNDING AND RELATION CONSTRUCTION

For each child–parent pair in the SceneTree, the agent must decide which local structures realize the observed relation. We support this decision with two groups of evidence. Perceptual evidence includes the reference image, instance highlights and masks, rendered views of the object assets, and the numbered region index from Section 3.2.1, which allows the agent to refer to regions by identifier. Structural evidence consists of the region descriptions and the library of available relation functions, together with the geometric role that each parameter of a function requires. From the local arrangement in the image, the agent selects a relation function and binds candidate regions to its geometric parameters. The resulting part-level relation explicitly names the geometry involved rather than relying on object-level categories alone.

Different relation functions require local geometry with different roles. For support, the agent identifies the parent surface on which the child should rest. Containment requires the floor and inner boundaries of a specific compartment, while insertion requires a receiving opening and interior geometry on the parent, together with the insertion part of the child. Despite these differences, all relation functions share a common parameterization in terms of addressable local geometry, supplemented by mesh-derived features when needed.

Grounding errors would propagate into both layout solving and physical refinement, so we verify the binding before proceeding. We render multi-view images that highlight the selected regions and have the agent confirm their correspondence with the intended structures in the reference image. The verified grounding is stored as a relation record containing the child, parent, relation function, and bound geometric parameters, which is used in both layout solving and physical refinement.

3.2.3 RELATION-CONDITIONED LAYOUT SOLVING

A relation function maps its bound geometric parameters to pose conditions on the child relative to the parent. We illustrate this mapping with support. The selected parent region specifies the target placement area, and the child mesh provides its projected footprint. Let P denote the planar extent of the selected support region and $F_c(T_c)$ the footprint of child c projected onto the same plane under pose T_c. We score a candidate pose by the overlap ratio

$$\mathrm{overlap}(T_c, P) = \frac{\mathrm{area}\big(F_c(T_c) \cap P\big)}{\mathrm{area}\big(F_c(T_c)\big)}. \tag{1}$$

Intuitively, this ratio measures how much of the child rests on the intended region rather than on its surroundings. PVSR corrects the translation of the child relative to its initial pose so that its footprint overlaps the target region and its bottom reaches the height of the selected support surface, while retaining the initial rotation. Containment and insertion admit analogous conditions. The selected floor and inner walls jointly delimit a feasible placement region for containment, and the opening, interior boundaries, and insertion part of the child constrain admissible positions and orientations for insertion. Additional relations can be incorporated by specifying their required local geometric roles and corresponding pose conditions, without changing the representation of geometric candidates.

Corrections are composed along the SceneTree from parents to children. A pose update applied to a parent propagates to its descendants, and each child is corrected in the updated coordinate frame of its parent. The output of this stage is the relation-refined scene S^{rel} together with the part-level relation records, and the grounded local geometry remains available for the relation checks.

3.3 RELATION-PRESERVING PHYSICAL REFINEMENT

The relation-refined layout assigns objects to the local structures indicated by the image, but two failure modes remain. Objects may still interpenetrate, and the assembled scene may be unstable under gravity. The two failures differ in what reveals them. Penetration is a static property that can be detected and repaired from the current configuration alone, whereas stability can only be assessed from the simulated response of an object together with its supporting structure. A procedure that addresses only one of them inevitably reintroduces the other, so PVSR alternates between geometric repair and simulation-based search, with the grounded part-level relations as a shared constraint that keeps every pose change consistent with the intended placement.

The refinement proceeds as a geometric-dynamic-geometric cycle. Geometric repair first removes detectable penetrations across the scene, allowing dynamic search to concentrate on one object and its support chain. Dynamic search then corrects instabilities that static geometry cannot assess, and a final round of geometric repair removes the residual penetrations that the pose updates introduce.

3.3.1 RELATION-CONSTRAINED GEOMETRIC REPAIR

Geometric repair resolves inter-object penetrations with translation-only corrections that retain the current object rotations. Motion is coordinated through the SceneTree, which partitions the objects into fixed anchors and movable groups. A translation applied to a movable parent is inherited by its descendants as a rigid displacement, so a correction moves a whole supported group coherently. For a colliding pair (i, j), let $\mathbf{n}_{ij}$ denote a separating direction and d_{ij} the required separation. We solve for coordinated translations satisfying

$$\mathbf{n}_{ij}^{\top}(\Delta\mathbf{t}_i - \Delta\mathbf{t}_j) \geq d_{ij}, \tag{2}$$

while favoring small deviations from the current layout.

A candidate correction is accepted only after three checks. It must not create new collisions, it must respect the allowed displacement range, and it must preserve the part-level relations affected by the move. The grounded local geometry enters directly into the third check, so collision repair stays consistent with the target support region or the containment space.

3.3.2 PER-OBJECT DYNAMIC SEARCH ALONG THE SUPPORT CHAIN

Searching the poses of all objects jointly is impractical and largely unnecessary, because a pose change propagates only along the support chain. We decompose the search into per-object subproblems and traverse the SceneTree from parents to children. For each child we sample a batch of translation and rotation candidates around its current pose. The simulation always includes the child together with its full ancestor chain up to the root, where the root remains fixed and intermediate ancestors respond dynamically. Each subproblem optimizes a single object's pose while retaining its physical coupling to the supporting structure. This decomposition also isolates the scoring of each candidate. Under joint multi-object sampling, an improvement for a difficult object can mask degradation of another, whereas per-child search leaves children with satisfactory layouts untouched. We score the k-th candidate of child c by three terms that reflect the two objectives of this stage,

$$E_{\mathrm{c}}^{(k)} = w_{\mathrm{drift}} E_{\mathrm{drift}}^{(k)} + w_{\mathrm{vel}} E_{\mathrm{vel}}^{(k)} + w_{\mathrm{layout}} E_{\mathrm{layout}}^{(k)}. \tag{3}$$

The drift term aggregates the translational and rotational changes of dynamic objects in the support chain between the initial and final simulated states. The velocity term measures short-term motion during simulation. The layout term penalizes deviations from the relation-refined scene, which prevents stabilization from undoing the relations established in Section 3.2. Among the candidates that preserve the target local relation, we select the one with the lowest cost and update the child accordingly. When a parent changes pose, the update propagates to its descendants before their local searches begin.

After the per-object updates, we apply relation-constrained geometric repair once more to resolve the residual penetrations and inter-object conflicts introduced by dynamic search, which yields the final scene S^{final}.

4 EXPERIMENTS

We evaluate PVSR along three axes, overall scene reconstruction quality, part-level relation modeling, and relation-preserving physical refinement. First, we compare against existing single-image scene reconstruction methods on the established TabletopGen benchmark (TTG-78) (Wang et al., 2025) and the Free Track of PartRelBench, measuring both visual-semantic consistency and physical validity. We then use the Controlled Track to reduce errors from upstream segmentation and asset reconstruction and more directly evaluate part-level relation grounding and layout solving. Finally, we isolate the contributions of part-level relation modeling, relation-preserving physical refinement, and the initial geometric repair through controlled ablations.

4.1 EXPERIMENTAL SETUP

Benchmarks. We evaluate on the established TTG-78 benchmark and on PartRelBench, our benchmark suite with two tracks. TTG-78 covers common tabletop objects and arrangements. The Free Track contains 50 single-image scenes organized around recurring failure patterns of existing reconstruction systems, including multi-level support, multiple candidate local surfaces on the same parent object, structurally complex supporting objects, and missing critical local structures caused by imperfect asset reconstruction. The Controlled Track provides ground-truth instance masks and assets and covers support, containment, and insertion, reducing interference from upstream reconstruction errors. Within each setting, all methods receive the same reference images and instance masks.

Table 1: Perceptual and physical evaluation on TTG-78 and PartRelBench (Free Track). VF, IA, and PP are scored from 1 to 7; Avg. is their arithmetic mean. Stability and net pose drift are evaluated after 5 s of physical simulation. Collision Obj. uses fixed 10 mm verified separation.

	Perceptual quality				Physical quality			
Method	VF ↑	IA ↑	PP ↑	Avg. ↑	Stable (%) ↑	ΔTrans. (cm) ↓	ΔRot. (rad) ↓	Collision Obj. (%) ↓
TTG-78 (78 scenes)								
MIDI	3.245	3.200	2.373	2.940	25.65	42.14	1.1964	69.46
TabletopGen	3.960	4.221	4.982	4.388	85.86	5.19	0.2799	0.40
REST3D	4.056	4.213	5.170	4.480	83.99	4.68	0.1905	**0.21**
REST3D-CD	**4.095**	4.239	5.215	4.516	85.67	5.01	0.2184	0.36
PVSR	4.066	**4.253**	**5.250**	**4.523**	**95.61**	**1.20**	**0.1097**	0.43
PartRelBench Free Track (50 scenes)								
MIDI	3.895	4.185	2.982	3.688	29.00	83.46	1.0683	65.38
TabletopGen	3.272	3.510	3.595	3.459	43.23	50.58	0.7049	14.60
REST3D	3.322	3.730	3.405	3.486	73.23	16.24	0.5092	18.84
REST3D-CD	3.458	3.805	3.797	3.687	88.83	6.70	0.2066	31.40
PVSR	**3.915**	**4.338**	**5.085**	**4.446**	**91.77**	**6.26**	**0.1859**	**12.82**

Compared Methods. We compare PVSR with four representative methods. **MIDI** is a multi-instance diffusion method for single-image 3D scene generation. **TabletopGen** assembles scenes through per-instance asset generation followed by pose and scale alignment; we use SAM3D as its asset backbone and leave the remainder of its pipeline unchanged. **REST3D**, the closest method to ours, represents support relations with an object-level SceneTree and improves physical plausibility through simulation-based CEM optimization. Because CEM depends on its collision representation, we additionally report **REST3D-CD**, a variant that replaces its optimization-time collision geometry with the convex-decomposition bodies of our evaluation protocol, so that the collision constraints seen by CEM match those used in the final physical evaluation.

Evaluation Metrics. Following TabletopGen, a vision-language model scores each method pair on Visual Fidelity (VF), Image Alignment (IA), and Physical Plausibility (PP) from 1 to 7, and the aggregate score is their arithmetic mean. Each pair is evaluated in both presentation orders to reduce order bias, covering all scenes and all pairs of the five methods (protocol details in Appendix A). We additionally run 5 s of physical simulation for each scene and evaluate dynamic stability, net pose drift, and residual collisions, with the primary stability criterion requiring translation drift below 10 cm and rotation drift below 1 rad. Collision rates use a fixed 10 mm verified-separation check.

4.2 Main Results on Single-Image Reconstruction

Table 1 reports perceptual and physical results on TTG-78 and the PartRelBench Free Track. We first use TTG-78 to test whether PVSR remains effective on conventional tabletop scenes, and then use the PartRelBench Free Track to evaluate its ability to handle complex local structures and part-level layouts. In the perceptual evaluation, PVSR obtains the highest IA (4.253), PP (5.250), and aggregate score (4.523), while remaining competitive in visual fidelity (4.066). In the physics-based evaluation, it achieves the highest dynamic stability rate of 95.61%, exceeding the strongest baseline, TabletopGen, by 9.75 percentage points, with translation and rotation drift of 1.20 cm and 0.1097 rad that are 74.4% and 42.4% lower than the second-best results. Its collision-object rate of 0.43% is slightly above the best baseline result of 0.21%, the one metric where relation-constrained corrections trade a small collision margin for a much larger reduction in drift. The aggregate scores of REST3D and REST3D-CD differ by only 0.036, which indicates that collision representation is not the bottleneck on conventional scenes and that the remaining gains come from relation grounding itself. Although part-level relations primarily target complex local structures, these results show that they do not compromise performance on conventional tabletop scenes.

The perceptual ranking on the Free Track separates generative methods from relation-aware ones. MIDI attains a reasonable VF (3.895) but the lowest PP (2.982), which shows that plausible geometry alone does not yield physically credible layouts, while TabletopGen drops on every perceptual axis. PVSR attains the highest VF (3.915), IA (4.338), PP (5.085), and aggregate score (4.446), 0.758 points above the strongest baseline. The physical results expose a complementary failure mode of object-level optimization. REST3D-CD reaches 88.83% stability, only 2.94 percentage points below PVSR, yet its collision-object rate of 31.40% is more than double that of PVSR (12.82%), and its translation and rotation drift are both higher. It therefore reaches most of its stability by settling objects into configurations that remain interpenetrating, whereas relation-constrained



Figure 3: Qualitative results on the PartRelBench Free Track. For each scene, the upper and lower rows show reconstructions before and after physical simulation. The magnified views highlight PVSR's placements on (a) the lower shelf and (b) the keyboard tray. Each method uses the same crop and scale at both time points.

Table 2: Controlled evaluation on PartRelBench. Best results are highlighted in bold.

	Perceptual quality				Physical quality			
Method	VF ↑	IA ↑	PP ↑	Avg. ↑	Stable (%) ↑	ΔTrans. (cm) ↓	ΔRot. (rad) ↓	Collision Obj. (%) ↓
REST3D-CD	3.889	3.222	3.556	3.556	78.19	7.34	**0.3957**	79.28
PVSR	**4.611**	**4.361**	**4.194**	**4.389**	**80.75**	**1.93**	0.5687	**74.75**

refinement removes 59.2% of colliding objects while simultaneously improving stability. The margin persists under a stricter rotation tolerance of 0.1 rad (Appendix B.1). These gains are consistent with the failure patterns around which the Free Track is organized, and Figure 3 shows representative cases where object-level methods miss multi-level support or pick an unintended candidate surface while PVSR retains the intended placement.

Figure 3 compares reconstructed scenes before and after 5 s of physical simulation. In (a), TabletopGen places the books on an upper surface, and they fall off during simulation, while PVSR places them on the lower shelf indicated by the reference image and retains that placement. In (b), REST3D-CD places the keyboard below the desk, whereas PVSR places it on the keyboard tray and preserves this relation after simulation. In both cases the perceptual gap in Table 1 is visible as a difference in which local structure supports the object, and the before-and-after pairing shows that the gap widens after simulation because unstable placements are corrected by falling rather than by restoring the intended relation.

4.3 PartRelBench Controlled Track

The Controlled Track provides ground-truth instance masks and object assets. We compare PVSR with REST3D-CD, providing both methods with the same reference images, ground-truth masks and assets, and initial poses estimated by GigaPose.

For REST3D-CD, we replace its SAM3D-based initialization with the provided assets and GigaPose poses, while retaining its object-level SceneTree, relation-driven layout adjustment, and CEM-based physical optimizer. Its optimization-time collision geometry remains aligned with the convex-decomposition representation used for physical evaluation. This shared initialization enables a controlled comparison of downstream relation modeling, layout solving, and physical refinement, while controlling for differences in segmentation and asset reconstruction. We report the same perceptual and physical metrics as in Table 1, with bidirectional perceptual comparisons between the two methods.

Table 2 shows that PVSR improves all three perceptual scores, raising the aggregate score from 3.556 to 4.389. It also increases the stability rate from 78.19% to 80.75%, reduces translation drift from 7.34 cm to 1.93 cm, and lowers the collision-object rate from 79.28% to 74.75%. Mean

Table 3: Component analysis and refinement stages on the PartRelBench Free Track. Layout denotes part-level relation grounding and solving, and Refine. denotes our relation-preserving physical refinement. Perceptual scores are produced within a separate three-method comparison and are not directly comparable to Table 1. Physical metrics use the same metric as Table 1.

	Components		Perceptual quality				Physical quality			
Method	Layout	Refine.	VF ↑	IA ↑	PP ↑	Avg. ↑	Stable (%) ↑	ΔTrans. (cm) ↓	ΔRot. (rad) ↓	Collision Obj. (%) ↓
(a) Component analysis										
REST3D-CD PVSR Layout	×	×	3.340	3.625	3.695	3.553	88.83	6.70	0.2066	31.40
+ REST3D-CEM	✓	×	3.595	4.010	4.350	3.985	89.57	9.11	0.2247	22.81
PVSR Full	✓	✓	**3.770**	**4.195**	**5.030**	**4.332**	**91.77**	**6.26**	**0.1859**	**12.82**
(b) Refinement stages from the same part-level layout										
Initial geometric repair only							79.30	11.35	0.4299	
w/o initial geometric repair							88.18	7.27	0.2695	
PVSR Full							**91.77**	**6.26**	**0.1859**	**12.82**

rotation drift, however, increases from 0.3957 rad to 0.5687 rad. These results demonstrate improved perceptual quality and complementary gains in stability and collision reduction under shared ground-truth assets and pose initialization.

4.4 Ablation and Component Analysis

To analyze the contributions of part-level relation modeling and relation-preserving physical refinement, we construct three incremental settings on the PartRelBench Free Track. **REST3D-CD** uses object-level relations and REST3D's original CEM-based physical optimizer. **PVSR Layout + REST3D-CEM** adds part-level relation grounding and layout solving while retaining REST3D's CEM optimizer. **PVSR Full** further replaces it with our complete relation-preserving physical refinement. Table 3(a) summarizes the perceptual and physical results, and Table 3(b) examines the stages of physical refinement from the same part-level layout.

Adding part-level relation grounding and layout solving to REST3D-CD raises the aggregate perceptual score from 3.553 to 3.985, with PP rising from 3.695 to 4.350, and reduces the collision-object rate from 31.40% to 22.81%. Stability is nearly unchanged at 89.57%, yet translation and rotation drift increase from 6.70 cm to 9.11 cm and from 0.2066 rad to 0.2247 rad. Because CEM moves objects freely once a relation is expressed only at the object level, the optimizer trades the corrected layout for short-term stability. Part-level grounding therefore fixes the layout, but an object-level optimizer can still undo the fix. Replacing CEM with our refinement further raises the aggregate perceptual score to 4.332 and PP to 5.030, while translation and rotation drift fall to 6.26 cm and 0.1859 rad and the collision-object rate drops from 22.81% to 12.82%. The informative comparison is between the second and third rows of Table 3(a), which share the same part-level layout and differ only in the optimizer. CEM converts the layout improvements into additional drift, whereas our refinement converts the same layout into additional stability, raising the stability rate from 89.57% to 91.77%. Stability gains no longer come at the cost of the image-grounded layout.

Geometric repair alone reaches 79.30% stability, compared with 91.77% for the full refinement (Table 3(b)). Omitting the initial repair while retaining dynamic search and the final repair reduces stability to 88.18% and increases both translation and rotation drift. The initial repair matters because a deeply interpenetrating object has no nearby stable configuration for dynamic search to sample, so the search either spends its candidate budget on collisions or settles for a partially corrected pose. Removing static violations first lets dynamic search concentrate on genuine instability, which is why the geometric-dynamic-geometric cycle outperforms either extreme.

5 Conclusion

We presented PVSR, a framework that grounds image-observed spatial relations in part-level geometric interfaces and keeps them active as constraints throughout layout solving and physical refinement, enabling single-image tabletop reconstruction that is both structurally faithful to the reference image and physically stable. Its two tracks, the Free Track and the Controlled Track, further provide a testbed for studying reconstruction in scenes where local object structure is decisive, and we hope it supports future work on finer-grained relation modeling.

AI USE STATEMENT

We used ChatGPT to assist with writing portions of the research code, generating a subset of the dataset used in our experiments, and drafting and improving the readability of limited portions of the manuscript. The authors reviewed the AI-assisted code, data, and text before use and verified the technical claims and experimental results in the final manuscript. We take full responsibility for the content of this work, including materials produced with the aid of generative AI.

ETHICS STATEMENT

This work does not involve human subjects, personal data, or applications with foreseeable harm, and we are not aware of ethical concerns raised by the released benchmark or reconstruction pipeline.

REPRODUCIBILITY STATEMENT

We will release the PartRelBench benchmark, evaluation protocols, and implementation code upon acceptance, and the appendix documents the perceptual evaluation prompts, physical simulation settings, and aggregation procedures used in all experiments.

REFERENCES

Andreea Ardelean, Mert Özer, and Bernhard Egger. Gen3DSR: Generalizable 3d scene reconstruction via divide and conquer from a single view. In *International Conference on 3D Vision (3DV)*, pp. 616–626, 2025.

Yinuo Bai, Peijun Xu, Kuixiang Shao, Yuyang Jiao, Jingxuan Zhang, Kaixin Yao, Jiayuan Gu, and Jingyi Yu. PARSE: Part-aware relational spatial modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Weixing Chen, Zhuoqian Feng, Yang Liu, Yexin Zhang, Yifan Wen, Yinghong Liao, Weichao Qiu, Guanbin Li, and Liang Lin. Physcene3d: Physically consistent interactive 3d tabletop scene generation. *arXiv preprint arXiv:2606.01649*, 2026.

Tianyuan Dai, Josiah Wong, Yunfan Jiang, Chen Wang, Cem Gokmen, Ruohan Zhang, Jiajun Wu, and Li Fei-Fei. Automated creation of digital cousins for robust policy learning. In *Proceedings of the 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 4912–4943. PMLR, 2025.

Can Gümeli, Angela Dai, and Matthias Nießner. ROCA: Robust CAD model retrieval and alignment from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. URL https://arxiv.org/abs/2112.01988.

Haonan Han, Rui Yang, Huan Liao, Jiankai Xing, Zunnan Xu, Xiaoming Yu, Junwei Zha, Xiu Li, and Wanhua Li. REPARO: Compositional 3d assets generation with differentiable 3d layout alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. URL https://arxiv.org/abs/2405.18525.

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

Hamid Izadinia, Qi Shan, and Steven M. Seitz. IM2CAD. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. URL https://arxiv.org/abs/1608.05137.

Inhee Lee, Sangwon Baik, Sungjoo Kim, Hyeonwoo Kim, Hyunsoo Cha, and Hanbyul Joo. SimuScene: Simulation-ready compositional 3d scene reconstruction from a single image. *arXiv preprint arXiv:2606.03994*, 2026.

Haodong Li, Lulu Shao, Haolin Lu, Yu Fu, Yen-Ru Chen, Seemandhar Jain, and Manmohan Chandraker. ϕ-Scene: Physically grounded image-to-3d scene reconstruction. *arXiv preprint arXiv:2606.21596*, 2026.

Chenguo Lin and Yadong Mu. InstructScene: Instruction-driven 3d indoor scene synthesis with semantic graph prior. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2402.04717.

Guying Lin, Kemeng Huang, Michael Liu, Ruihan Gao, Hanke Chen, Lyuhao Chen, Beijia Lu, Taku Komura, Yuan Liu, Jun-Yan Zhu, and Minchen Li. PAT3D: Physics-augmented text-to-3d scene generation. *arXiv preprint arXiv:2511.21978*, 2025. URL https://arxiv.org/abs/2511.21978.

Lu Ling, Chen-Hsuan Lin, Tsung-Yi Lin, Yifan Ding, Yu Zeng, Yichen Sheng, Yunhao Ge, Ming-Yu Liu, Aniket Bera, and Zhaoshuo Li. Scenethesis: A language and vision agentic framework for 3d scene generation. *arXiv preprint arXiv:2505.02836*, 2025. URL https://arxiv.org/abs/2505.02836.

Anran Liu, Cheng Lin, Yuan Liu, Xiaoxiao Long, Zhiyang Dou, Hao-Xiang Guo, Ping Luo, and Wenping Wang. Part123: Part-aware 3d reconstruction from a single-view image. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–12, 2024.

Yang Liu, Weixing Chen, Yongjie Bai, Xiaodan Liang, Guanbin Li, Wen Gao, and Liang Lin. Aligning cyber space with physical world: A comprehensive survey on embodied ai. *IEEE/ASME Transactions on Mechatronics*, 2025.

Xiaoxuan Ma, Jiashun Wang, Nicolas Ugrinovic, Yehonathan Litman, and Kris Kitani. REST3D: Reconstructing physically stable 3d scenes from a single image. *arXiv preprint arXiv:2605.30338*, 2026.

Yanxu Meng, Haoning Wu, Ya Zhang, and Weidi Xie. SceneGen: Single-image 3d scene generation in one feedforward pass. In *International Conference on 3D Vision (3DV)*, 2026. URL https://arxiv.org/abs/2508.15769.

Fan-Yun Sun, Weiyu Liu, Siyi Gu, Dylan Lim, Goutam Bhat, Federico Tombari, Manling Li, Nick Haber, and Jiajun Wu. LayoutVLM: Differentiable optimization of 3d layout via vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. URL https://arxiv.org/abs/2412.02193.

Zefan Tian, Yuteng Ye, Yiheng Zhang, Yuhang Yang, Xueqiang Lv, Shizhou Zhang, Le Liu, and Di Xu. SceneReGen: Generative reconstruction of 3d scenes from a single image. *arXiv preprint arXiv:2608.23930*, 2026.

Yian Wang, Han Yang, Minghao Guo, Xiaowen Qiu, Tsun-Hsuan Wang, Wojciech Matusik, Joshua B. Tenenbaum, and Chuang Gan. PhyScensis: Physics-Augmented LLM Agents for Complex Physical Scene Arrangement. *arXiv preprint arXiv:2602.14968*, 2026. URL https://arxiv.org/abs/2602.14968.

Ziqian Wang, Yonghao He, Licheng Yang, Wei Zou, Hongxuan Ma, Liu Liu, Wei Sui, Yuxin Guo, and Hu Su. TabletopGen: Instance-level interactive 3d tabletop scene generation from text or single image. *arXiv preprint arXiv:2512.01204*, 2025.

Qirui Wu, Denys Iliash, Daniel Ritchie, Manolis Savva, and Angel X. Chang. Diorama: Unleashing zero-shot single-view 3d indoor scene modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8896–8907, 2025.

Yandan Yang, Baoxiong Jia, Peiyuan Zhi, and Siyuan Huang. PhyScene: Physically interactable 3d scene synthesis for embodied AI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024a. URL https://arxiv.org/abs/2404.09465.

Yue Yang, Fan-Yun Sun, Luca Weihs, Eli VanderBilt, Alvaro Herrasti, Winson Han, Jiajun Wu, Nick Haber, Ranjay Krishna, Lingjie Liu, Chris Callison-Burch, Mark Yatskar, Aniruddha Kembhavi, and Christopher Clark. Holodeck: Language guided generation of 3d embodied AI environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024b. URL https://arxiv.org/abs/2312.09067.

Kaixin Yao, Longwen Zhang, Xinhao Yan, Yan Zeng, Qixuan Zhang, Lan Xu, Wei Yang, Jiayuan Gu, and Jingyi Yu. CAST: Component-aligned 3d scene reconstruction from an RGB image. *ACM Transactions on Graphics*, 44(4):86:1–86:19, 2025. doi: 10.1145/3730841.

Junsheng Zhou, Yu-Shen Liu, and Zhizhong Han. Zero-shot scene reconstruction from single images with deep prior assembly. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2410.15971.

A Evaluation Details

A.1 Perceptual Evaluation

We follow the three perceptual dimensions described in the main text: Visual Fidelity (VF), Image Alignment (IA), and Physical Plausibility (PP). The evaluator receives the reference image and the rendered montages of two reconstructions, and assigns a score from 1 to 7 to each reconstruction along each dimension. We use `gpt-5.6-luna` and the same evaluation prompt throughout the finalized comparisons. Each method pair is presented in both orders, and scores are aggregated across presentation orders, scenes, and opponents. The reported average is the arithmetic mean of VF, IA, and PP.

For five-method comparisons, each scene contributes ten unordered method pairs. This gives 780 scene-pair evaluation jobs on TTG-78 and 500 on the PartRelBench Free Track; each job includes both presentation orders. The three-method component analysis contains 150 scene-pair jobs on the Free Track. Because a method's scores are conditioned on the alternatives presented to the evaluator, we compare methods within the same evaluation pool. The main comparison and the component analysis therefore retain their own perceptual aggregates.

A.2 Physics-Based Evaluation

Each final reconstruction is evaluated in PhysX for 5 s at 60 Hz. Let $\mathcal{D}_s$ denote the set of dynamic objects in scene s, $\Delta x_{s,i}$ the net displacement of object i's geometric centroid, and $\Delta\theta_{s,i}$ its net rotation between the initial and final simulation states. The stability rate at rotation tolerance τ is

$$\mathrm{Stable}(\tau) = \frac{1}{|\mathcal{S}|}\sum_{s\in\mathcal{S}}\frac{1}{|\mathcal{D}_s|}\sum_{i\in\mathcal{D}_s}\mathbf{1}[\Delta x_{s,i} < 0.10\,\mathrm{m} \ \wedge\ \Delta\theta_{s,i} < \tau]\,. \tag{4}$$

The primary criterion uses $\tau = 1$ rad; the stricter analysis below uses $\tau = 0.1$ rad. Mean translation and rotation drift follow the same object-within-scene and then scene-equal aggregation. Consequently, a scene with many objects does not receive greater weight than a scene with fewer objects. Five seconds specifies the evaluation horizon; stability is determined by the displacement and rotation criteria rather than assumed from elapsed time.

Residual collisions are evaluated separately using simplified triangle meshes and FCL. For each initially intersecting pair, the evaluation tests whether a relative translation of 10 mm yields a verified separation along one of six coordinate-axis directions or 1,334 sampled sphere directions. Pairs for which no separation is verified are retained. We report the fraction of objects incident to at least one retained pair, including fixed supporting objects, and average this fraction equally over scenes. This finite-direction separation test is a collision diagnostic rather than an exact penetration-depth measurement.

B Additional Quantitative Analyses

B.1 Sensitivity to the Stability Criterion

Table 4 tightens the rotation tolerance from 1 rad to 0.1 rad while keeping the translation tolerance and evaluated reconstructions unchanged. On the Free Track, PVSR remains the highest-scoring method under both criteria. Its strict stability rate is 83.38%, compared with 80.36% for REST3D-CD. The improvement therefore persists when substantially less rotational drift is allowed.

Table 4: Stability under two rotation tolerances on the PartRelBench Free Track (50 scenes). The translation tolerance is 10 cm in both columns. Higher is better. Both rates are computed from the same finalized per-object evaluation records.

Method	Stable, 1 rad (%) ↑	Stable, 0.1 rad (%) ↑
MIDI	29.00	13.83
TabletopGen	43.23	38.79
REST3D	73.23	57.63
REST3D-CD	88.83	80.36
PVSR	**91.77**	**83.38**

The component analysis further illustrates why stability should be considered together with continuous drift and collision measurements. PVSR Layout + REST3D-CEM and PVSR Full obtain

nearly identical strict stability rates, 83.34% and 83.38%, respectively. Nevertheless, the complete refinement reduces mean translation drift from 9.11 cm to 6.26 cm, rotation drift from 0.2247 rad to 0.1859 rad, and the collision-object rate from 22.81% to 12.82%. A binary stability decision alone does not capture the magnitude of the remaining errors.

B.2 Repeatability of Perceptual Scoring

We compare two completed evaluations of the same five-method Free Track inputs. Their input locks contain identical hashes for the 50 reference images, 250 reconstruction montages, and the evaluation prompt. The second evaluation re-scores all 500 scene-pair jobs without reusing previous judgments; it supplies the main-paper scores. Table 5 shows that PVSR obtains the highest aggregate score in both evaluations. Its margin over the strongest baseline is 0.753 points in the first evaluation and 0.758 in the second. Individual scores and some baseline rankings vary, but the leading aggregate result is consistent across these two evaluations.

Table 5: Repeated perceptual evaluation on identical Free Track inputs. Each evaluation covers 500 scene-pair jobs with both presentation orders. The second evaluation is used in the main table. Higher is better.

	IA		PP		Avg.	
Method	Eval. 1	Eval. 2	Eval. 1	Eval. 2	Eval. 1	Eval. 2
MIDI	4.173	4.185	3.018	2.982	3.705	3.688
TabletopGen	3.467	3.510	3.583	3.595	3.449	3.459
REST3D	3.640	3.730	3.362	3.405	3.428	3.486
REST3D-CD	3.842	3.805	3.803	3.797	3.706	3.687
PVSR	**4.378**	**4.338**	**5.082**	**5.085**	**4.459**	**4.446**

C Additional Qualitative Results

C.1 Additional Free Track Examples

We supplement the main-paper examples with five Free Track scenes: 015, 037, 075, 136, and 188. The examples cover compartmentalized furniture, stacked support, multiple candidate support levels, thin objects, and residual instability. They are selected for diagnosis of distinct reconstruction behaviors rather than for estimating their frequency in the benchmark.

Visualization protocol. Each of these five comparisons includes the reference photograph and all five methods: MIDI, TabletopGen, REST3D, REST3D-CD, and PVSR. We render frames 0 and 300 from the saved 60 Hz evaluation trajectories, corresponding to the final reconstruction before simulation and its state after 5 s. These are not the inputs and outputs of our physical refinement. Each method uses the same camera, crop, and scale for both time points. Crops include the union of visible content across the two frames. The supplementary figures use the original material renderings without additional color correction, including for MIDI. Viewpoints can differ between methods, so the figures compare structural placement and temporal changes rather than pixel-level alignment.

Compartment-specific placement. In case 015 (Figure 4), the reference image distributes objects between upper compartments and the main work surface. PVSR places the globe and the red drawer box in the upper structure while keeping the pink case on the lower work surface, and these placements persist after simulation. In the REST3D-CD reconstruction, the red drawer box is instead placed on the lower level and the pink case is below the work surface. This example illustrates why identifying the parent furniture alone is insufficient: the relevant compartment or support level must also be determined.

Stacked support and temporal consistency. Case 037 (Figure 4) combines stacks of instruments on an elevated shelf with objects on the lower work surface. PVSR retains both the upper instrument arrangement and the lower red container after simulation. In the REST3D-CD sequence, the red container and a covered object end up below the work surface. The paired frames distinguish recovery of an initial arrangement from its persistence under physical simulation.

Multiple candidate support levels. The workstation in case 075 (Figure 5) contains several vertically separated surfaces, including a printer shelf, a monitor shelf, and a low platform for the computer tower. TabletopGen places the monitor on the upper shelf and initializes the tower beside the

upper structure; the tower subsequently drops. PVSR instead places the printer and monitor on their respective upper and middle shelves and retains the tower on the low platform. These differences concern selection among local support structures within the same furniture asset.

Thin objects and remaining pose errors. Case 136 (Figure 5) separates objects on the hutch top from those on the main desktop. PVSR preserves the placement of the box on the upper structure and the keyboard on the desktop, but a thin, paper-like object changes from an upright pose to a flat pose during simulation. Correct support-level selection therefore does not eliminate every residual orientation error. The reference also contains fine-grained stationery that is absent or simplified in the reconstruction, illustrating the remaining dependence on upstream instance decomposition and asset reconstruction.

Residual instability in a populated scene. Case 188 (Figure 6) is a deliberately included limitation example. Much of the reconstructed layout remains recognizable, but the plant moves off its initial placement and some objects on top of the cabinet tip over in the PVSR sequence. The example shows that improved average stability does not imply that every object in every scene is stable. Preserving a useful hierarchy of support relations and achieving complete dynamic stability remain distinct requirements.

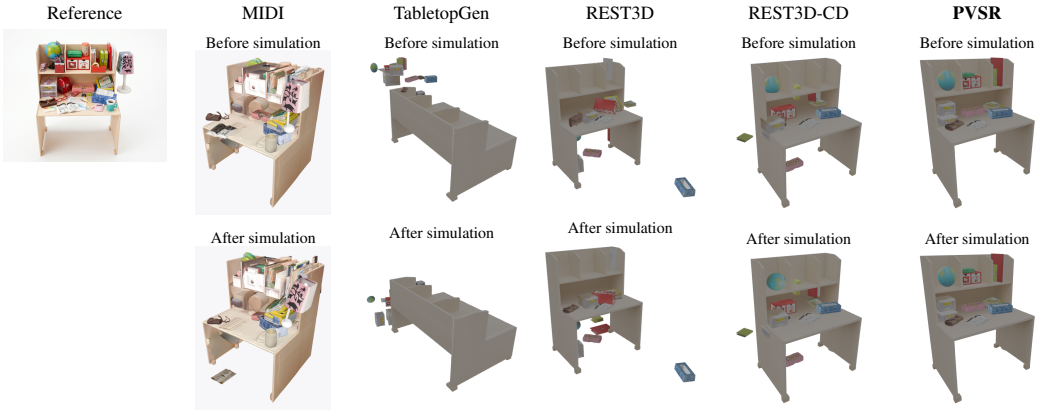
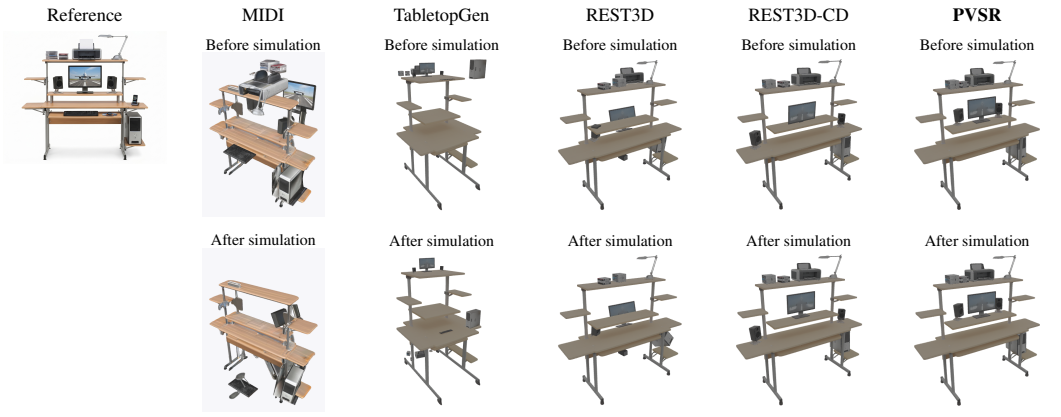


Figure 4: Additional Free Track comparisons: compartment-specific placement (case 015) and stacked support (case 037). For each method, the upper image is before simulation and the lower image is after simulation (5 s). Before/after crops and scales are shared. See the visualization protocol in Appendix C.

Tiered workstation (case 075). Multiple support heights distinguish the monitor, printer, and computer tower.



Desk with a hutch (case 136). Main support-level placement coexists with a residual orientation error of a thin object.

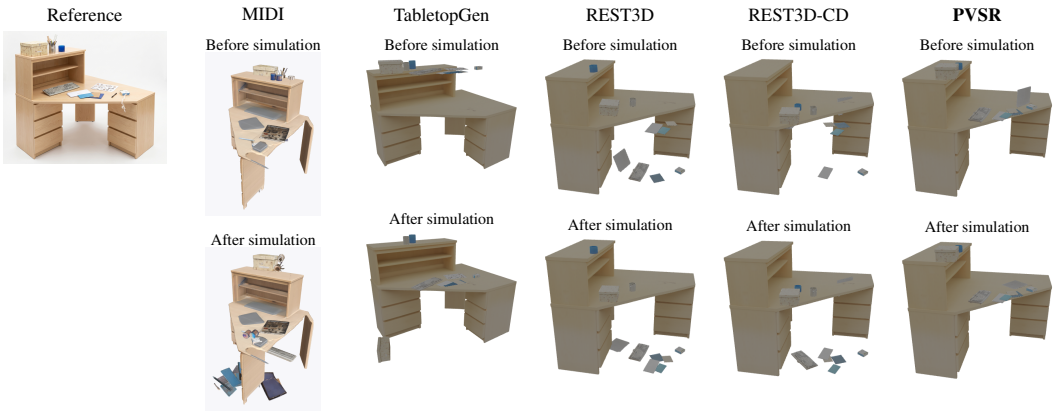


Figure 5: Additional Free Track comparisons: multiple support levels (case 075) and remaining thin-object pose errors (case 136). Upper and lower images show the states before and after 5 s of simulation. All five methods are included with fixed before/after framing.

Residual instability (case 188). The plant and some upper-cabinet objects move despite preservation of much of the scene layout.

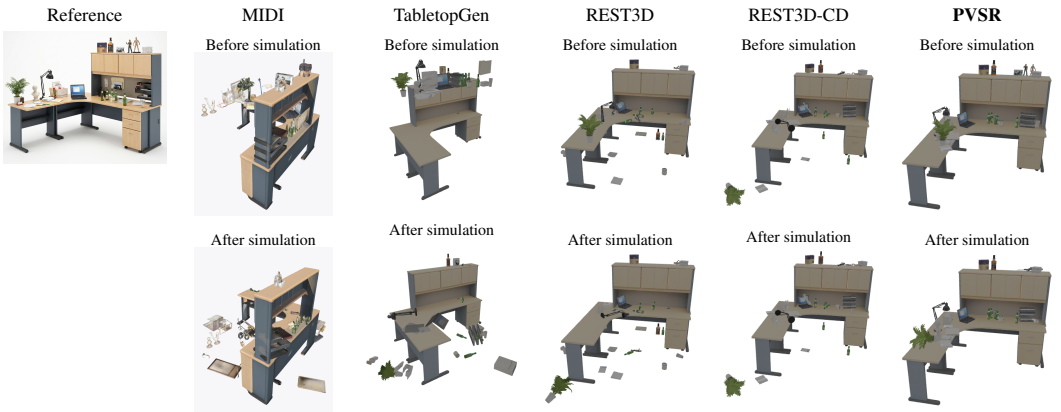


Figure 6: Remaining instability in case 188. The paired images show that PVSR retains much of the scene layout but does not prevent all local motion: the plant and some objects on the cabinet change pose during simulation. Upper: before simulation; lower: after simulation (5 s).

C.2 Case Study: A Slatted Shelf

We additionally examine a single-image scene containing a green three-tier shelf and six movable objects. The vertically separated tiers and the gaps between their slats make the local support geometry important to placement. Both methods receive the same instance masks, SAM3D-reconstructed meshes, and VLM-inferred SceneTree, and use the same precomputed collision bodies. The shelf is fixed during the 5 s PhysX evaluation at 60 Hz. PVSR applies part-level layout solving followed by geometric repair and support-chain dynamic refinement, whereas REST3D-CD uses its object-level layout and CEM-based optimization. This is a comparison of complete pipelines, not an ablation isolating their samplers.

Figure 7 shows the final reconstructed layouts from a shared rendering camera. In the reference image, the plush toy and the two noodle cups occupy the middle tier. PVSR places these objects on that tier, whereas the REST3D-CD output places the plush toy outside the shelf. This difference is present despite the shared object meshes and collision geometry.

The subsequent physical evaluation further distinguishes the two layouts (Table 6). Under the strict criterion of translation drift below 10 cm and rotation drift below 0.1 rad, PVSR keeps all six movable objects stable, compared with four for REST3D-CD. Mean geometric-centroid drift is 5.77 mm versus 31.28 mm, and mean rotation drift is 0.0273 rad versus 0.2753 rad. The largest difference occurs for the plush toy: its centroid drift is 5.45 mm with PVSR and 147.30 mm with REST3D-CD. The middle cup also fails REST3D-CD's strict stability criterion, with 0.1337 rad rotation drift, compared with 0.0123 rad for PVSR. Together, the rendered layouts and simulation results illustrate the complete pipeline's benefit on this slatted structure.

Table 6: Single-scene physical evaluation of the slatted shelf. Strict stability uses 10 cm/0.1 rad thresholds after 5 s. Drift is averaged over the six movable objects, with translation measured at each object's geometric centroid. This case is reported separately from the benchmark aggregates.

Method	Strict stable objects ↑	ΔCentroid (mm) ↓	ΔRot. (rad) ↓
REST3D-CD	4/6	31.28	0.2753
PVSR	**6/6**	**5.77**	**0.0273**



Figure 7: A single-image slatted-shelf comparison using shared object meshes and collision bodies. Both reconstruction panels use the same camera and framing. PVSR places the plush toy on the middle tier indicated by the reference image; REST3D-CD places it outside the shelf. The panels are static renders of the final method outputs, not frames after the 5 s evaluation rollout.