# IMSCENE: AN AGENTIC FRAMEWORK FOR RECONSTRUCTING SIMULATION-READY 3D SCENES FROM A SINGLE IMAGE

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Physically reliable 3D scenes are essential for embodied navigation, robotic manipulation, and interactive environment simulation. Existing image-to-3D methods generate plausible objects and recover scene layouts, but inconsistencies between generated geometry, estimated poses, and inter-object relationships can still lead to interpenetration, floating objects, and unstable support. Resolving these inconsistencies from a single image requires reconciling incomplete visual evidence with geometric and physical constraints on the reconstructed scene. We present ImScene, an agentic framework for reconstructing editable 3D scenes from a single image. A detection agent corrects missing, duplicate, and mislabeled instances, while a vision-language model constructs a physical scene graph encoding support, containment, and attachment relations. Next, we reconstruct a scene point cloud from estimated depth and establish a gravity-aligned metric coordinate system. Guided by the scene graph, a differentiable optimizer then jointly refines the poses and scales of objects generated by an image-to-3D model, using image and depth alignment, surface contact constraints, and penalties for interpenetration and unsupported placement. Finally, a simulator-guided agent identifies unstable object placements from simulated motion and support contacts, then applies bounded pose adjustments or rigid-body settling to improve stability while maintaining consistency with the input image. Comprehensive comparisons demonstrate improved consistency with the input image and greater physical plausibility, providing a stronger foundation for subsequent scene editing and simulation.

## 1 INTRODUCTION

Reconstructing editable 3D indoor scenes from a single photograph offers a practical route to creating environments for embodied navigation, robotic manipulation, and interactive simulation. These applications require both consistency with the observed environment and physically plausible object placement. A reconstruction may resemble its input image yet contain intersecting furniture, floating objects, or objects placed outside valid support regions. Such errors become consequential when objects are viewed from new positions, edited, or introduced into a physics simulator. Single-image scene reconstruction therefore requires recovering a coherent arrangement of complete objects from incomplete visual evidence.

Recent advances in single-image reconstruction have enabled the recovery of detailed objects and compositional scene layouts. Systems such as Gen3DSR, MIDI, and SceneGen reconstruct object arrangements from images, while pretrained models such as SAM 3D recover individual object assets (Ardelean et al., 2025; Huang et al., 2025; Meng et al., 2026; SAM 3D Team et al., 2025). However, completed object geometry may be incompatible with the estimated scene layout, even when each object aligns with its visible image region. Occlusion limits the evidence available for resolving these conflicts, and instance detection errors can lead to omissions, duplicates, or incorrect labels in the object inventory. Improving individual object appearance or source-view alignment alone therefore does not ensure a coherent assembled scene.



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.

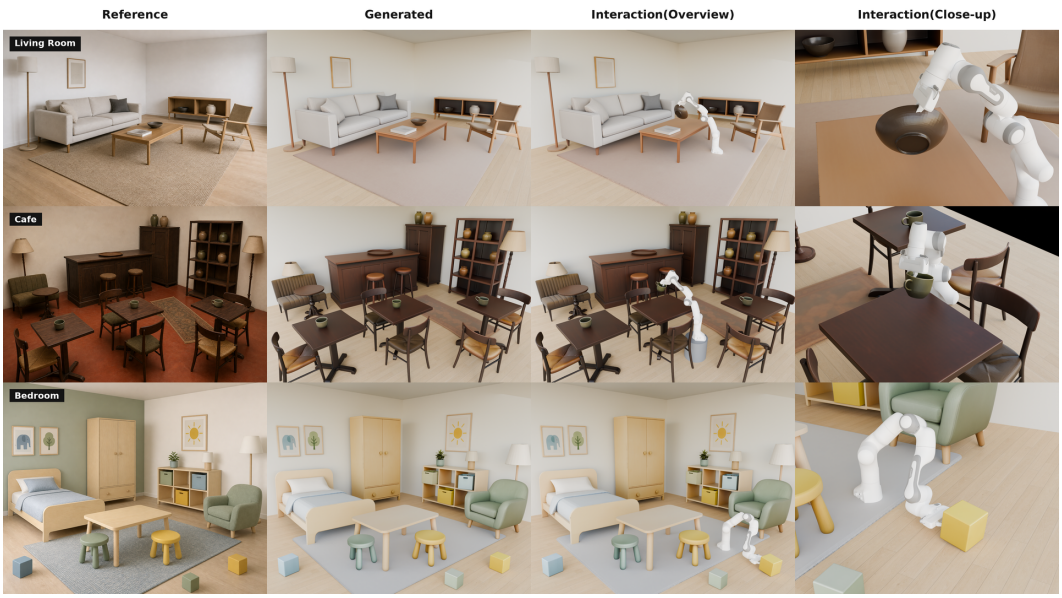


Figure 1: **Single-image scene reconstruction and robotic interaction in MuJoCo.** ImScene reconstructs diverse indoor environments, including a living room, a café, and a bedroom, which are imported into MuJoCo for simulation (Todorov et al., 2012). Each row presents the input image and reconstructed scene, followed by overview and close-up views of a robotic arm interacting with reconstructed objects.

To address these inconsistencies, we present ImScene, an agentic framework that integrates scene understanding, geometric reconstruction, and simulation feedback. A detection agent inspects the source image and instance overlays annotated with instance IDs to correct missing, duplicate, or mislabeled instances. A vision-language model (VLM) then infers support, containment, and attachment relations, which undergo structural validation to form a physical scene graph. Room geometry and objects reconstructed by SAM 3D are initialized in a common, gravity-aligned metric frame using estimated depth and camera geometry.

With many objects, jointly recovering their positions, rotations, and scales entails a large search space for single-image 3D scene reconstruction. ImScene constrains this space by integrating visual evidence with inferred physical relations grounded in mesh surfaces and finite support regions. Supported objects are projected onto their assigned surfaces to establish contact. A differentiable optimizer jointly refines rigid-object poses and scales using source-view masks, predicted depth, and penalties for interpenetration and placement outside valid support regions or the room boundaries. A simulator-guided agent then identifies residual instability through physics simulation and applies bounded pose adjustments or rigid-body settling. This physical verification complements geometric refinement, since source-image alignment and low geometric errors alone do not ensure stable behavior. Experiments show that ImScene better reproduces the observed layout, reduces interpenetration and unsupported placements, and improves object stability in simulation.

Our contributions are twofold:

- We introduce ImScene, an agentic framework that combines visual evidence, inferred physical relations, and simulation feedback to reconstruct editable 3D scenes from a single image, improving source-image consistency and physical plausibility.
- We develop a scene refinement method that grounds physical relations in generated geometry for joint differentiable optimization of object poses and scales. A simulator-guided agent further corrects residual instability under geometric and visual constraints.

## 2 Related Work

**Compositional and physically grounded reconstruction.** IM2CAD retrieves and aligns CAD models, while Total3D jointly predicts room layout, object poses, and geometry (Izadinia et al., 2017; Nie et al., 2020). Recent approaches adopt object-wise or context-aware generation, including Gen3DSR, MIDI, and SceneGen (Ardelean et al., 2025; Huang et al., 2025; Meng et al., 2026), while SAM 3D provides object geometry, appearance, and initial layout (SAM 3D Team et al., 2025). For geometric alignment, 3D-RE-GEN refines object placement using silhouettes and estimated scene geometry (Sautter et al., 2025). Physical priors improve reconstruction and pose refinement (Bauer et al., 2022; Ni et al., 2024). Physical inconsistencies are addressed through relation-based contact and support constraints in CAST (Yao et al., 2025), agentic scene-tree construction and simulator-guided pose optimization in REST3D (Ma et al., 2026), and simulation-driven geometry correction in SimuScene (Lee et al., 2026). ImScene grounds inferred relations in finite mesh regions and jointly refines object poses and scales using direct mask and depth alignment together with geometric placement constraints.

**Agentic scene construction and refinement.** Language and vision-language agents combine reasoning, tool use, and iterative feedback (Yao et al., 2023; Nasiriany et al., 2024). For scene construction, Lucida uses a VLM-operated 3D interface to refine placements from multiple captured views (Qin et al., 2026), while SceneSmith, SAGE, and Scenethesis combine language-guided generation with visual or physical critics (Pfaff et al., 2026; Xia et al., 2026; Ling et al., 2026). These systems primarily address multi-view reconstruction or scene generation. ImScene targets the particular arrangement in one RGB image, treating that image as a fixed visual reference and using agents to improve reconstruction fidelity and physical plausibility.

## 3 Method

Given one RGB image, ImScene reconstructs an editable 3D scene with physically plausible object placement. It first audits object instances and infers a physical scene graph, then initializes generated meshes and finite relation surfaces in a gravity-aligned metric frame. A differentiable optimizer jointly refines object poses and scales, after which a simulator-guided agent repairs residual dynamic failures under visual and geometric acceptance constraints.

### 3.1 Agentic Scene Understanding

**Instance detection and correction.** A VLM proposes open-vocabulary categories, Grounding DINO localizes candidate instances (Liu et al., 2024), and SAM 3 predicts instance masks conditioned on the resulting boxes (Carion et al., 2025). A detection agent then compares the image with a labeled instance overlay and iteratively adds missing objects, removes false positives, relabels instances, or merges separately detected parts of the same object. The loop stops when no correction is proposed or its budget is exhausted. The final inventory contains $N$ instances, each with a persistent identifier, category label, mask $M_i$, and image box $b_i$; see Appendix A.1.

**Physical relation inference.** From the image and labeled instances, the VLM proposes a graph $\mathcal{G}$ by assigning each object a primary parent $p(i)$ and relation $r_i \in \{\texttt{on}, \texttt{inside}, \texttt{mounted_on}, \texttt{hang}\}$. Parents may be objects or room surfaces, allowing an internal cabinet shelf, for example, to be represented without a separate detected instance. We check identities, relation-parent compatibility, coverage, and acyclicity, returning violations to the VLM for correction. After bounded retries, any remaining structural violations are resolved using the fallback rules. The resulting hierarchy orders placement and surface association; see Appendix A.2.

### 3.2 Metric Scene Initialization

Depth Pro estimates metric depth and focal length (Bochkovskii et al., 2025). Using the resulting camera intrinsics, we reconstruct a metric point map and provide it to SAM 3D together with the RGB image and instance masks. SAM 3D generates object meshes with initial poses and scales (SAM 3D Team et al., 2025), forming the independent assembly $\mathcal{S}_{\text{init}}$. Planes fitted to floor

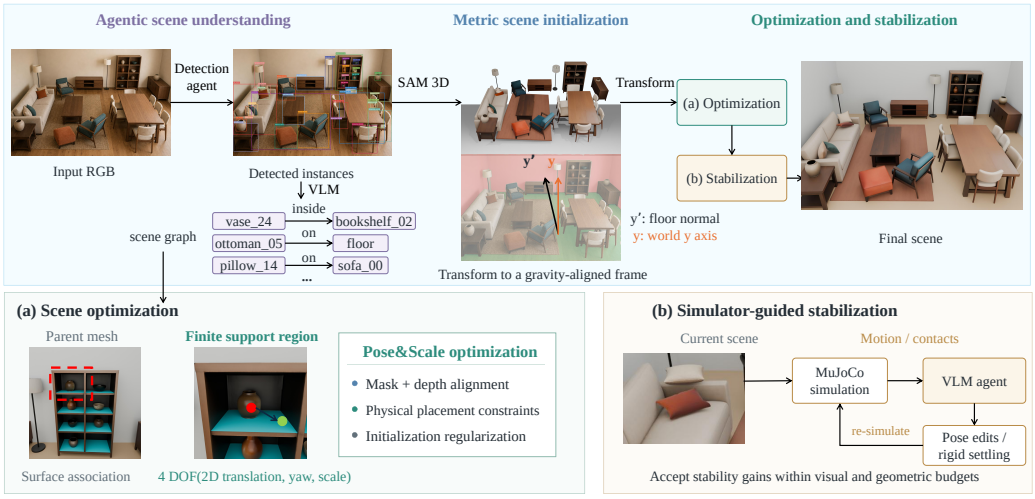


Figure 2: **ImScene overview. Agentic scene understanding** detects and corrects instances and constructs a physical scene graph. **Metric scene initialization** aligns generated assets in a gravity-aligned metric frame. **(a) Scene optimization** refines object poses and scales using image/depth alignment and physical placement constraints. **(b) Simulator-guided stabilization** corrects residual instability through simulation feedback, accepting improvements within visual and geometric preservation budgets.

and, when available, ceiling points establish gravity. Lines fitted to horizontally projected wall points determine room yaw under a rectangular-room assumption; camera-aligned yaw is retained when wall evidence is insufficient. We estimate finite room bounds from visible boundaries and size priors. The assets and source camera are transformed consistently into a common $Y$-up metric layout frame, preserving their source-view projection.

Following HSM and SceneSmith (Pun et al., 2026; Pfaff et al., 2026), we extract finite support regions from parent meshes and associate inferred relations with compatible regions or room planes. Projecting each movable object onto its assigned plane yields the initial scene $\mathcal{S}_0$ for Section 3.3. This initialization enforces plane contact, while finite-region support and interpenetration are addressed during subsequent optimization. Appendix B details the coordinate conversion, support-region extraction, and association procedure.

### 3.3 Scene Optimization

The optimizer starts from $\mathcal{S}_0$ and jointly refines the poses and scales of movable rigid objects. The optimized set $\mathcal{I}_{\text{opt}}$, of size $N_{\text{opt}}$, excludes wall-mounted and hanging objects, room-shell openings, and floor coverings such as rugs. These objects remain in the scene as visual and geometric context.

**Contact-constrained placement.** For each object $i \in \mathcal{I}_{\text{opt}}$ with initial scale $s_i^0$, we optimize two horizontal offsets, a yaw residual, and a log-scale residual, $q_i = (\Delta u_i, \Delta v_i, \Delta\psi_i, \sigma_i)$, with $s_i = s_i^0 e^{\sigma_i}$. Each movable object is assigned a horizontal support region. Its horizontal offsets determine the planar position. Let $h_i$ be the support height and $y_{\min,i}$ the mesh's lowest $y$-coordinate after scaling and rotation, before translation. Contact then fixes the vertical translation analytically:

$$t_{i,y} = h_i + g - y_{\min,i}, \tag{1}$$

where $g$ is a small contact gap. This preserves the prescribed separation as pose and scale change. Graph and region assignments remain fixed, while support heights and finite footprints follow their parents' current transforms.

**Joint objective.** All residuals $\boldsymbol{q} = \{q_i : i \in \mathcal{I}_{\text{opt}}\}$ are optimized under

$$\begin{aligned}\mathcal{L}(\boldsymbol{q}) = \lambda_{\text{mask}}\mathcal{L}_{\text{mask}} + \lambda_{\text{pc}}\mathcal{L}_{\text{pc}} + \lambda_{\text{col}}\mathcal{L}_{\text{col}} + \lambda_{\text{sup}}\mathcal{L}_{\text{sup}} \\ + \lambda_{\text{room}}\mathcal{L}_{\text{room}} + \lambda_{\text{prior}}\mathcal{L}_{\text{prior}}.\end{aligned} \tag{2}$$

**Visual alignment.** We render all objects together with differentiable rasterization (Laine et al., 2020), so the antialiased occupancy $\alpha_i(x;\boldsymbol{q})$ accounts for inter-object occlusion. Given the reference mask $M_i$ of each optimized object, the soft-IoU loss is

$$\mathcal{L}_{\text{mask}} = \frac{1}{N_{\text{opt}}} \sum_{i\in\mathcal{I}_{\text{opt}}} \left[1 - \frac{\sum_x \alpha_i(x)M_i(x) + \varepsilon_M}{\sum_x[\alpha_i(x) + M_i(x) - \alpha_i(x)M_i(x)] + \varepsilon_M}\right]. \tag{3}$$

Here $\varepsilon_M = 10^{-6}$. Let $P_i$ be the masked reference-depth points of object $i$ and $Q_i(\boldsymbol{q})$ its visible transformed-mesh samples. With $d(x, Y) = \min_{y\in Y} \|x - y\|_2$ and the common distance scale $d_0 = 50\,\text{mm}$, define $\rho_{d_0}(z) = \sqrt{1 + (z/d_0)^2} - 1$. The depth-alignment term is

$$\mathcal{L}_{\text{pc}} = \frac{1}{2N_{\text{opt}}} \sum_{i\in\mathcal{I}_{\text{opt}}} \left[\frac{1}{|P_i|}\sum_{p\in P_i} \rho_{d_0}(d(p, Q_i)) + \frac{1}{|Q_i|}\sum_{q\in Q_i} \rho_{d_0}(d(q, P_i))\right]. \tag{4}$$

**Interpenetration and placement constraints.** For a transformed probe $x \in X_a$ of object $a$ and the world-space signed-distance field $\phi_b^w$ of object $b$, define $r_{ab}(x) = [-\phi_b^w(x) - \epsilon_{\text{col}}]_+/d_0$, where $[z]_+ = \max(z, 0)$ and $\epsilon_{\text{col}} = 1\,\text{mm}$ is the penetration tolerance. The directed penetration penalty is

$$\ell_{ab}^{\text{sdf}} = \frac{1}{|X_a|}\sum_{x\in X_a} r_{ab}(x)^2 + \frac{1}{4}\max_{x\in X_a} r_{ab}(x)^2. \tag{5}$$

For independent objects moving on horizontal support planes, $\ell_{ab}^{\text{tan}}$ has the same mean-plus-maximum form but uses the normalized horizontal escape distance of penetrating probes. Because contact fixes vertical translation, this term measures the escape required along the available horizontal directions rather than a potentially shorter but infeasible vertical direction (Appendix C.3). Over these collision candidate pairs $\mathcal{P}_{\text{col}}$, the bidirectional collision loss is

$$\mathcal{L}_{\text{col}} = \frac{1}{2N_{\text{opt}}} \sum_{\{i,j\}\in\mathcal{P}_{\text{col}}} \sum_{(a,b)\in\{(i,j),(j,i)\}} \left(\ell_{ab}^{\text{sdf}} + \beta\ell_{ab}^{\text{tan}}\right). \tag{6}$$

Let $d_{\Omega,i}$ be the metric signed distance of the geometric mass center's horizontal projection to the assigned finite region, positive outside its boundary and negative inside it. With inward margin $m$ and smoothing scale $\tau$, the support loss is

$$\mathcal{L}_{\text{sup}} = \frac{1}{N_{\text{opt}}} \sum_{i\in\mathcal{I}_{\text{opt}}} \left[\frac{\tau}{d_0}\log\left(1 + \exp\left(\frac{d_{\Omega,i} + m}{\tau}\right)\right)\right]^2. \tag{7}$$

The softplus form smoothly penalizes center-of-mass projections approaching or crossing the finite support boundary, with $m$ providing an inward safety margin.

**Room bounds and initialization prior.** $\mathcal{L}_{\text{room}}$ penalizes object geometry extending beyond the estimated room boundaries. $\mathcal{L}_{\text{prior}}$ penalizes deviations from the initial configuration. Detailed formulations and parameter settings are provided in Appendix C.4.

**Optimization.** We initialize $\boldsymbol{q}_0 = 0$, run $T = 320$ Adam updates (Kingma & Ba, 2015) with cosine learning-rate decay, gradient-norm clipping, and hard parameter bounds, retaining the lowest complete-objective state among the initialization and all iterates:

$$k^\star = \arg\min_{0\le k\le T} \mathcal{L}(\boldsymbol{q}_k), \qquad \mathcal{S}_{\text{joint}} = \mathcal{S}(\boldsymbol{q}_{k^\star}). \tag{8}$$

We then restore $\mathcal{S}_{\text{joint}}$ and check for residual interpenetration using dense point-to-mesh signed-distance queries. If the audit finds a residual penetration, only the movable objects implicated by the detected pairs are active in the next round. We run at most three such local rounds, each with at most 80 Adam updates with the same loss terms and parameter bounds. The resulting scene is denoted by $\mathcal{S}_{\text{opt}}$. Algorithm 1 gives the complete procedure.

### 3.4 Simulator-Guided Agentic Stabilization

We initialize the accepted scene as $B_0 = \mathcal{S}_{\text{opt}}$ and obtain per-object motion and support feedback through a MuJoCo release audit (Todorov et al., 2012). At iteration $t$, the agent inspects this feedback and source/rendered views of the current accepted scene $B_t$, selects an unstable object or its support ancestor, and edits a working copy to form a candidate $\widetilde{B}_t$.

Ordinary edits use support-plane translation and normal-axis rotation, with height determined by contact; parent edits transport their support descendants. For loose objects such as cushions, local rigid-body settling can instead propose a pose. Scene graph edges and support bindings remain fixed. The executor checks pose bounds, room/support violations, pairwise penetration depths, mask IoU, and source-depth distance against preservation budgets relative to the fixed reference $B_0$, preventing cumulative drift across iterations.

A fresh full-scene release then evaluates the complete candidate. Relative to $B_t$, at least one unstable object must become strictly stable or reduce its normalized release motion by at least $10\%$. Objects already stable and support contacts retained in either $B_t$ or $B_0$ must be preserved. Thus physical improvement drives acceptance, while geometric and visual measures constrain the permitted changes.

If all checks pass, we accept $B_{t+1} = \widetilde{B}_t$; otherwise, we retain $B_{t+1} = B_t$ and return rejection feedback to the agent. The loop continues until all objects pass the benchmark stability criterion, the agent stops, or its budget is exhausted; unresolved failures may remain. Appendix D specifies the tools and thresholds.

## 4 Experimental Evaluation

### 4.1 Evaluation Protocol

**Datasets and baselines.** The benchmark contains $56$ single-image indoor scenes: $30$ synthetic images generated using GPT Image 2 (OpenAI, 2026c), $10$ 3D-FRONT views (Fu et al., 2021), and $16$ ScanNet++ images (Yeshwanth et al., 2023). We compare with six baselines: MIDI (Huang et al., 2025), Gen3DSR (Ardelean et al., 2025), SceneGen (Meng et al., 2026), 3D-RE-GEN (Sautter et al., 2025), direct SAM 3D assembly (SAM 3D Team et al., 2025), and REST3D (Ma et al., 2026). We report average valid results over scenes, omitting missing outputs.

**Metrics.** Reference-geometry evaluation uses source-view-visible surfaces on both 3D-FRONT and ScanNet++. Reference objects are selected from GT instance annotations with at least $50$ visible source-image pixels; architectural structures are excluded. After scene alignment, we evaluate Chamfer distance (CD, in m) and F-score $F_5$ at a 5 cm threshold (Knapitsch et al., 2017). We also report matched object-box IoU (B-IoU) of complete CAD-instance or annotated scan-instance bounds. Unmatched objects contribute zero to B-IoU; details are in Appendix E.

Physical metrics cover every method's valid submitted scenes. We report original-mesh collision involvement and stability, both normalized by the number of submitted objects. For each object, five-second stability requires translation at most 0.1 m, rotation at most $10^\circ$, and terminal linear speed at most 0.1 m/s. We also report mean object speeds at 0.25 and 5 s and displacement after 5 s. Wall/ceiling installations and openings inferred from ImScene's scene graph are exempt from gravity while retaining collision response. REST3D follows the same rule using its inferred scene tree. Baselines without a physical graph receive no inferred attachment labels: their submitted objects remain subject to gravity.

**Implementation.** The full objective uses weights $(1, 1, 60, 4, 30, 0.1)$ for mask, depth, collision, finite support, room, and prior, respectively. Adam uses seed $42$, learning rate $0.015$ with cosine decay, $320$ joint updates, and at most three $80$-update local rounds. For physical evaluation, a single scene-level similarity transform maps all submitted meshes to a common Z-up frame, centers the scene horizontally, places its global lowest vertex on the ground plane at $z = 0$, and uniformly scales the scene to a maximum horizontal span of $4$ m. The MuJoCo simulation (Todorov et al., 2012) includes every submitted object, with time step $1/960$ s, gravity $9.81$ m/s$^2$, density $350$ kg/m$^3$, and sliding friction $0.8$. All methods use the same CoACD convex decomposition rule (Wei et al.,

2022) and whole-object convex-hull mass and inertia. Only a common ground plane is added, not reconstructed wall or ceiling colliders. See Appendix E.

### 4.2 Quantitative Comparison

The main comparisons and loss ablations evaluate ImScene before agentic refinement; Section 4.5 evaluates the additional stabilization step. Table 1 reports valid scene counts ($N$), collision involvement, five-second stability, mean object speeds at $0.25$ and $5$ s, and mean displacement after $5$ s.

Table 1: Physical evaluation: collision involvement, stability, mean object speeds, and displacement.

| Method | $N$ | Coll. (%) ↓ | Stable (%) ↑ | $v_{0.25s}$ (m/s) ↓ | $v_{5s}$ (m/s) ↓ | $d_{5s}$ (m) ↓ |
|---|---|---|---|---|---|---|
| MIDI | 55 | 65.55 | 5.41 | 2.5268 | 0.1777 | 2.8231 |
| Gen3DSR | 55 | 73.09 | 0.36 | 2.7034 | 0.1532 | 3.4038 |
| SceneGen | 54 | 73.57 | 3.15 | 2.6441 | 0.1243 | 2.3319 |
| 3D-RE-GEN | 56 | 31.57 | 4.56 | 1.7421 | 0.0136 | 0.8992 |
| SAM3D | 56 | 43.46 | 10.62 | 1.1149 | 0.0424 | 0.7418 |
| REST3D | 56 | **4.41** | 86.19 | 0.0985 | 0.0357 | 0.2868 |
| ImScene | 56 | 9.28 | **89.49** | **0.0540** | **0.0060** | **0.0784** |

Table 1 shows that ImScene achieves the highest stability and the lowest object speeds and displacement, while REST3D has the lowest collision involvement.

Table 2: Geometric evaluation: Chamfer distance (CD), object-box IoU (B-IoU), and F-score at 5 cm ($F_5$).

| | *3D-FRONT* | | | | *ScanNet++* | | | |
|---|---|---|---|---|---|---|---|---|
| Method | $N$ | CD (m) ↓ | B-IoU ↑ | $F_5$ ↑ | $N$ | CD (m) ↓ | B-IoU ↑ | $F_5$ ↑ |
| MIDI | 9 | 0.5698 | 0.0866 | 0.3777 | 16 | 0.6326 | 0.0486 | 0.2534 |
| Gen3DSR | 10 | 0.3818 | 0.1581 | 0.4832 | 16 | 0.3167 | 0.1156 | 0.4294 |
| SceneGen | 9 | 0.4539 | 0.1309 | 0.3651 | 15 | 0.7544 | 0.0282 | 0.2144 |
| 3D-RE-GEN | 10 | 0.2724 | 0.2283 | 0.5434 | 16 | 0.2202 | 0.0833 | 0.5028 |
| SAM3D | 10 | 0.2595 | 0.3290 | 0.6246 | 16 | **0.2076** | 0.1050 | 0.5007 |
| REST3D | 10 | 0.2431 | 0.3632 | 0.5498 | 16 | 0.3108 | 0.0846 | 0.4553 |
| ImScene | 10 | **0.2253** | **0.3791** | **0.6975** | 16 | 0.2103 | **0.1177** | **0.5534** |

Overall, ImScene achieves the best results on most geometric metrics across both datasets (Table 2). On 3D-FRONT, ImScene obtains $F_5 = 0.6975$ and CD $0.2253$ m. On ScanNet++, its $F_5$ is $0.5534$ versus $0.5007$ for SAM 3D, while SAM 3D has lower CD. SAM 3D can align more closely with the reference geometry in some cases, but its higher collision involvement and lower stability indicate weaker physical reliability (Table 1).

### 4.3 Qualitative Comparison

Figure 3 and Appendix G compare ImScene with the six baselines. MIDI retains many object categories but exhibits irregular geometry; SceneGen changes the visible inventory or arrangement; and Gen3DSR produces distorted geometry. 3D-RE-GEN misidentifies objects on the floor. Direct SAM 3D assembles recognizable assets, but the red boxes highlight misplaced shelf or cabinet objects and inconsistent floor contacts. REST3D misses many objects inside cabinets and other storage furniture and produces objects at implausible scales. ImScene better preserves the layout and places small objects within their intended support regions.

Recently released GPT-6 Astra (OpenAI, 2026b) can generate high-fidelity 3D scenes. Figure 4 compares ImScene with GPT-6 Astra for a music room and an artist's studio. In these examples, the GPT-6 Astra outputs exhibit discrepancies in object arrangement relative to the input images.

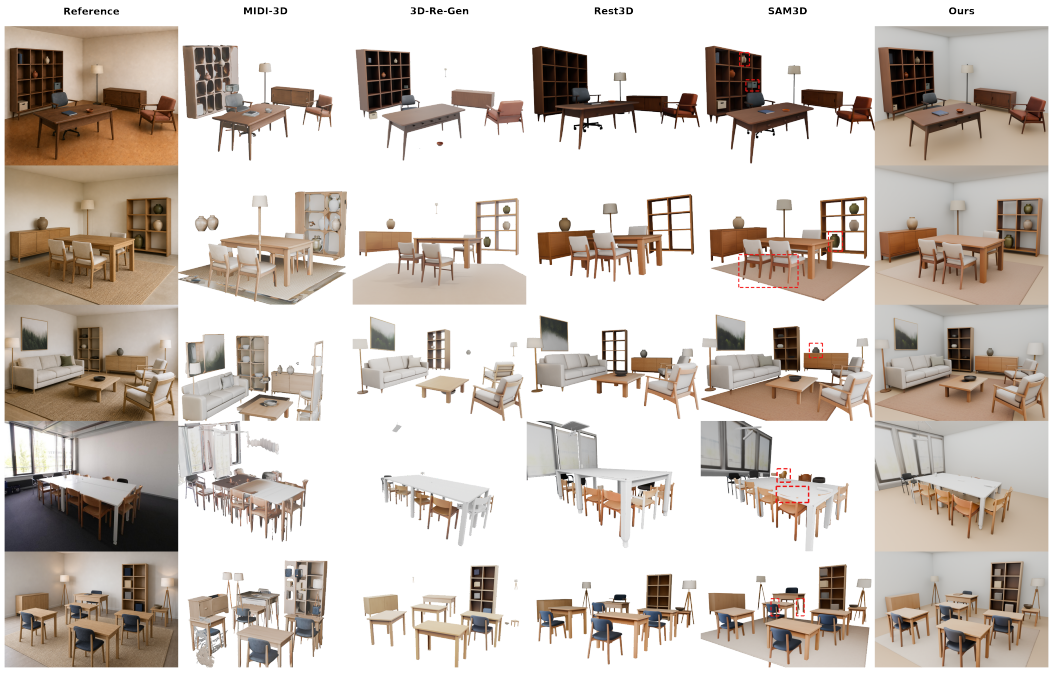


Figure 3: **Qualitative comparison across five indoor scenes.** Each row shows the reference, four selected baselines, and ImScene (right). The full comparison with all six baselines is in Appendix G. Red dashed boxes mark contact inconsistencies in SAM 3D.

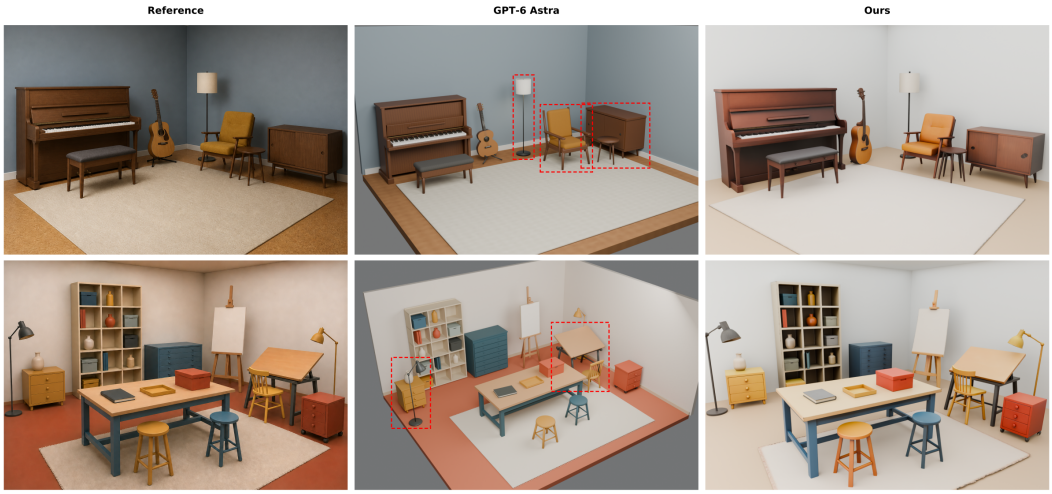


Figure 4: **Qualitative comparison with GPT-6 Astra.** GPT-6 Astra results are shown from their default output viewpoints. Red dashed boxes highlight layout discrepancies relative to the reference images, including incorrect orientations, placement against the wrong walls, and substantial shifts in relative positions.

### 4.4 Ablation Study

Table 3 reports a cumulative loss ablation, adding mask, depth, prior, finite-support, room, and collision terms in that order. Each optimized variant starts independently from $\mathcal{S}_0$. The final row, Full, includes all loss terms. Appendix F gives complementary single-term removals from Full.

Table 3: Cumulative loss ablation on 56 scenes and 963 objects. Geometry is evaluated on 10 3D-FRONT and 16 ScanNet++ scenes.

| | | | *3D-FRONT* | | *ScanNet++* | |
|---|---|---|---|---|---|---|
| Variant | Coll. (%) ↓ | Stable (%) ↑ | CD (m) ↓ | $F_5$ ↑ | CD (m) ↓ | $F_5$ ↑ |
| $\mathcal{S}_0$ | 30.32 | 74.85 | 0.2596 | 0.6100 | 0.2199 | 0.5337 |
| + mask | 31.13 | 76.19 | 0.2451 | 0.6446 | 0.2101 | 0.5691 |
| + depth | 25.39 | 78.85 | **0.2179** | **0.7302** | **0.2026** | **0.5924** |
| + prior | 26.81 | 77.81 | 0.2198 | 0.7234 | 0.2073 | 0.5744 |
| + support | 28.42 | 77.30 | 0.2182 | 0.7285 | 0.2081 | 0.5716 |
| + room | 28.46 | 77.63 | 0.2195 | 0.7192 | 0.2062 | 0.5633 |
| + collision (Full) | **9.28** | **89.49** | 0.2253 | 0.6975 | 0.2103 | 0.5534 |

Mask supervision improves source-view alignment, and adding depth gives the strongest geometric scores in this sequence. The prior, finite-support, and room terms constrain pose and scale drift, unsupported placement, and room-boundary violations, respectively; Appendix F isolates their effects.

Adding the collision term reduces collision involvement from $28.46\%$ to $9.28\%$ and raises stability from $77.63\%$ to $89.49\%$. This trades some geometric agreement for physical reliability: $F_5$ changes from $0.7192$ to $0.6975$ on 3D-FRONT and from $0.5633$ to $0.5534$ on ScanNet++.

### 4.5 Simulator-Guided Agentic Refinement

We evaluate simulator-guided agentic refinement as a complementary step to geometric optimization, focusing on stability gains and preservation of the reconstructed layout (Appendix D).

Table 4: Effect of agentic refinement on collision involvement, stability, and geometry.

| | | | *3D-FRONT* | | *ScanNet++* | |
|---|---|---|---|---|---|---|
| Variant | Coll. (%) ↓ | Stable (%) ↑ | CD (m) ↓ | $F_5$ ↑ | CD (m) ↓ | $F_5$ ↑ |
| Full | 9.28 | 89.49 | 0.2253 | 0.6975 | 0.2103 | **0.5534** |
| Full + Agent | **8.56** | **90.80** | **0.2252** | **0.6979** | **0.2103** | 0.5533 |

Table 4 shows a $1.31$-percentage-point stability gain and a $0.72$-point decrease in collision involvement. Across the two reference subsets, the maximum absolute changes in mean CD, $F_5$, and B-IoU are $1.5\times10^{-4}$ m, $3.2\times10^{-4}$, and $1.3\times10^{-4}$, respectively. Simulator-guided agentic refinement thus complements geometric optimization by improving stability and reducing collision involvement, while largely preserving the reconstructed layout and its consistency with the reference geometry and input image.

## 5 Conclusion and Limitations

We presented ImScene, an agentic framework that reconstructs editable and physically coherent 3D scenes from a single image by combining object-inventory correction, physical scene-graph reasoning, surface-grounded joint optimization, and simulator-guided repair. Experiments show that this integration substantially reduces inter-object collisions and improves dynamic stability while retaining the observed scene arrangement. ImScene nevertheless inherits the ambiguity of monocular depth, generated geometry, and inferred physical relations. Moreover, its current rigid-body formulation does not model deformable materials or articulated mechanisms, limiting interactions with objects such as fabrics, cushions, drawers, and doors. Extending the framework to deformable and articulated representations with richer physical properties would enable more diverse and realistic downstream interactions.

## AI USE STATEMENT

GPT Image 2 was used to generate the 30 synthetic input images in the evaluation benchmark, and ChatGPT (OpenAI, 2026a) was used to help polish the writing.

## REPRODUCIBILITY STATEMENT

Section 3 and Appendices A–D specify the scene representation, initialization, optimization, and agentic refinement. Section 4 reports the benchmark composition, baselines, evaluation metrics, optimization settings, and simulation parameters; Appendix E gives the evaluation protocol and metric definitions.

## 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.

Dominik Bauer, Timothy Patten, and Markus Vincze. SporeAgent: Reinforced scene-level plausibility for object pose refinement. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 654–662, 2022.

Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R. Richter, and Vladlen Koltun. Depth Pro: Sharp monocular metric depth in less than a second. In *International Conference on Learning Representations (ICLR)*, 2025.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025.

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 *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10913–10922, 2021.

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)*, pp. 5134–5143, 2017.

Diederik P. Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics*, 36(4), 2017.

Samuli Laine, Janne Hellsten, Tero Karras, Yeongho Seol, Jaakko Lehtinen, and Timo Aila. Modular primitives for high-performance differentiable rendering. *ACM Transactions on Graphics*, 39(6): 1–14, 2020.

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.

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. In *International Conference on Learning Representations (ICLR)*, 2026.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *Computer Vision – ECCV 2024*, pp. 38–55, 2024.

Xiaoxuan Ma, Jiashun Wang, Nicolás 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.

Soroush Nasiriany, Fei Xia, Wenhao Yu, Ted Xiao, Jacky Liang, Ishita Dasgupta, Annie Xie, Danny Driess, Ayzaan Wahid, Zhuo Xu, Quan Vuong, Tingnan Zhang, Tsang-Wei Edward Lee, Kuang-Huei Lee, Peng Xu, Sean Kirmani, Yuke Zhu, Andy Zeng, Karol Hausman, Nicolas Heess, Chelsea Finn, Sergey Levine, and Brian Ichter. PIVOT: Iterative visual prompting elicits actionable knowledge for VLMs. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 37321–37341, 2024.

Junfeng Ni, Yixin Chen, Bohan Jing, Nan Jiang, Bin Wang, Bo Dai, Puhao Li, Yixin Zhu, Song-Chun Zhu, and Siyuan Huang. PhyRecon: Physically plausible neural scene reconstruction. In *Advances in Neural Information Processing Systems*, volume 37, pp. 25747–25780, 2024.

Yinyu Nie, Xiaoguang Han, Shihui Guo, Yujian Zheng, Jian Chang, and Jian Jun Zhang. Total3DUnderstanding: Joint layout, object pose and mesh reconstruction for indoor scenes from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 55–64, 2020.

OpenAI. ChatGPT, 2026a. URL `https://chatgpt.com/`.

OpenAI. GPT-6 Astra System Card, 2026b. URL `https://deploymentsafety.openai.com/gpt-6-astra`.

OpenAI. GPT-Image-2. OpenAI API model documentation, 2026c. URL `https://developers.openai.com/api/docs/models/gpt-image-2`.

Nicholas Pfaff, Thomas Cohn, Sergey Zakharov, Rick Cory, and Russ Tedrake. SceneSmith: Agentic generation of simulation-ready indoor scenes. In *International Conference on Machine Learning (ICML)*, 2026.

Hou In Derek Pun, Hou In Ivan Tam, Austin T. Wang, Xiaoliang Huo, Angel X. Chang, and Manolis Savva. HSM: Hierarchical scene motifs for multi-scale indoor scene generation. In *Proceedings of the IEEE Conference on 3D Vision (3DV)*, pp. 1356–1367, 2026.

Minghan Qin, Yuang Wang, Xiuyu Yang, Yushi Long, Yujian Zhang, Ruihuan Wang, Kai Ye, Yangang Zhang, and Hang Li. Lucida: Parse, generate, and place for composable real-to-sim scene modeling. *arXiv preprint arXiv:2608.30821*, 2026.

Qwen Team. Qwen3.8-Max: A new bar for coding and cowork, August 2026. URL `https://qwen.ai/blog?id=qwen3.8`.

SAM 3D Team, Xingyu Chen, Fu-Jen Chu, Pierre Gleize, Kevin J Liang, Alexander Sax, Hao Tang, Weiyao Wang, Michelle Guo, Thibaut Hardin, Xiang Li, Aohan Lin, Jiawei Liu, Ziqi Ma, Anushka Sagar, Bowen Song, Xiaodong Wang, Jianing Yang, Bowen Zhang, Piotr Dollár, Georgia Gkioxari, Matt Feiszli, and Jitendra Malik. SAM 3D: 3Dfy anything in images. *arXiv preprint arXiv:2511.16624*, 2025.

Tobias Sautter, Jan-Niklas Dihlmann, and Hendrik P. A. Lensch. 3D-RE-GEN: 3D reconstruction of indoor scenes with a generative framework. *arXiv preprint arXiv:2512.17459*, 2025.

Emanuel Todorov, Tom Erez, and Yuval Tassa. MuJoCo: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5026–5033, 2012.

Xinyue Wei, Minghua Liu, Zhan Ling, and Hao Su. Approximate convex decomposition for 3D meshes with collision-aware concavity and tree search. *ACM Transactions on Graphics*, 41(4): 1–18, 2022.

Hongchi Xia, Xuan Li, Zhaoshuo Li, Qianli Ma, Jiashu Xu, Ming-Yu Liu, Yin Cui, Tsung-Yi Lin, Wei-Chiu Ma, Shenlong Wang, Shuran Song, and Fangyin Wei. SAGE: Scalable agentic 3D scene generation for embodied AI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22358–22368, 2026.

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):1–19, 2025.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023.

Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. ScanNet++: A high-fidelity dataset of 3D indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12–22, 2023.

# A SCENE UNDERSTANDING

We use Qwen3.8-Max (Qwen Team, 2026) for all VLM calls throughout ImScene, from instance detection and correction to physical-relation inference and simulator-guided refinement.

## A.1 ITERATIVE INSTANCE CORRECTION

Detection correction uses at most four rounds of source-image and labeled-overlay inspection. New detections are deduplicated before receiving persistent IDs; the final merge uses mask unions for object parts assigned to one asset. Nonoverlapping ID labels are also used for relation inference.

Table 5 reports instance detection results on the 26 images from 3D-FRONT and ScanNet++ with ground-truth instance mask annotations. For this evaluation, we include only ground-truth instances whose visible masks cover at least 0.5% of the image. ImScene achieves the highest Recall@0.5 and mask mIoU, with both metrics averaged across images. Despite using a detection agent, REST3D still misses many objects inside cabinets and other storage furniture, as shown in the qualitative comparisons in Appendix G. Our direct SAM 3D baseline takes instance masks as input and is therefore not evaluated as a separate detector.

Table 5: Detection comparison for visible ground-truth masks covering at least 0.5% of the image.

| Metric | MIDI | Gen3DSR | SceneGen | 3D-RE-GEN | REST3D | SAM3D | ImScene |
|---|---|---|---|---|---|---|---|
| Recall@0.5 (%) ↑ | 63.77 | 73.21 | 75.59 | 79.41 | 67.80 | N/A | **81.16** |
| Mask mIoU (%) ↑ | 58.29 | 64.63 | 67.07 | 67.92 | 58.89 | N/A | **71.10** |

## A.2 VLM PHYSICAL SCENE GRAPH

The VLM receives the source RGB image, a nonoverlapping instance-ID overlay, and a catalog of object identifiers (Figure 5). The image provides visual context, while the overlay disambiguates instances. For each object, it returns one primary parent, a relation from Table 6, and a brief visual justification. Parents must be detected objects or admissible room surfaces, which root the hierarchy. Occluded contacts are inferred from object semantics and visible context; an item on an undetected internal shelf is assigned `inside` the enclosing cabinet.

Table 6: Physical-relation vocabulary. Surface extraction subsequently grounds each accepted edge in a finite region of the parent geometry.

| Relation | Admissible parent | Interpretation |
|---|---|---|
| `on` | Object or ground | The child rests on an exposed support surface. |
| `inside` | Object | The child occupies a container or an internal shelf of the parent. |
| `mounted_on` | Wall | The child is rigidly attached to a vertical room surface. |
| `hang` | Ceiling | The child is suspended from the ceiling. |

Validation checks coverage, IDs, relation-parent compatibility, uniqueness, and acyclicity. An invalid graph and its structural violations are returned to the VLM for a complete corrected graph, within a budget of three requests. If no response passes validation, the response with the most usable edges is repaired: retain the first valid edge per source, redirect one edge per cycle to the floor, and assign objects without a valid parent to the floor. The final hierarchy determines parent-before-child placement and guides the surface binding in Appendix B.3.

# B METRIC INITIALIZATION AND SURFACE BINDING

## B.1 CAMERA AND LAYOUT COORDINATES

The camera frame uses $+X$ to the right, $+Y$ upward, and $-Z$ along the viewing direction. For pixel-center coordinates $(u, v) = (x + \frac{1}{2}, y + \frac{1}{2})$, metric depth $z = D(u, v)$, and camera intrinsics

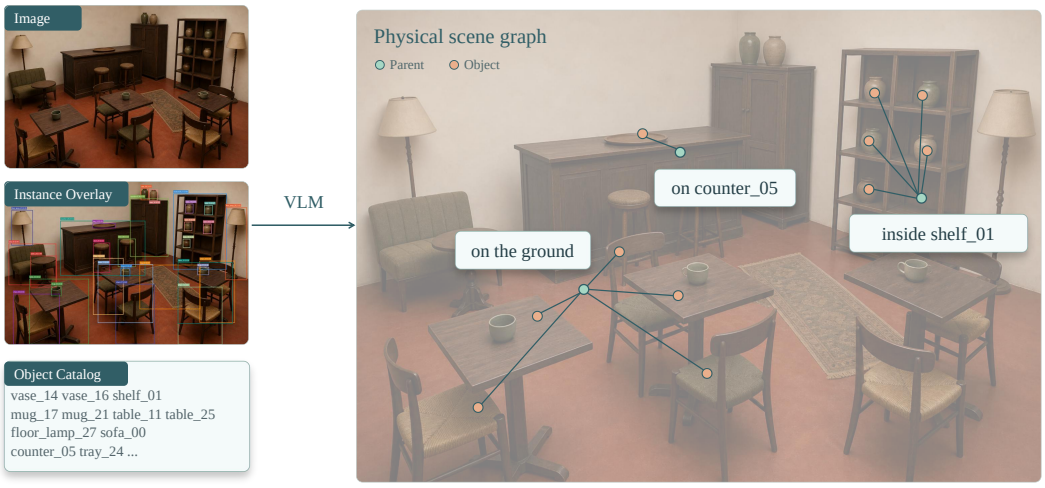


Figure 5: **VLM physical-relation inference.** The VLM receives the RGB image, a nonoverlapping instance-ID overlay, and the object catalog, then proposes one primary parent and relation type for each object.

$K = (f_x, f_y, c_x, c_y)$, backprojection gives

$$x^C(u, v, z) = [(u - c_x)z/f_x \quad (c_y - v)z/f_y \quad -z]^\top . \tag{9}$$

Depth Pro supplies $D$ and focal length; the principal point is the image center. Gravity, yaw, and finite room bounds follow Section 3.2. For each SAM 3D output, let $\widehat{V}_i$, $\widehat{t}_i$, $\widehat{R}_i$, and $\widehat{s}_i$ denote its mesh, camera-frame translation, row-vector rotation, and uniform scale. The exported $Y$-up mesh is mapped to the model's internal $Z$-up basis by

$$C_{\text{asset}} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{bmatrix}. \tag{10}$$

Let $o_i$ be the axis-aligned bounding-box center of $\widehat{V}_i$ and $V_i = \{\widehat{v} - o_i : \widehat{v} \in \widehat{V}_i\}$ its centered mesh. The camera-basis conversion $C_{\text{cam}} = \text{diag}(-1, 1, -1)$ maps the model's camera axes to our camera axes.

Let $R_{LC} \in SO(3)$ be the camera-to-layout rotation obtained by gravity alignment followed by room-yaw alignment. For the backprojected floor points $P_f^C$, the floor height $\widehat{h}_f$ is the 10th percentile of their vertical coordinates after rotation. If floor evidence is insufficient, $\widehat{h}_f$ is instead the estimated lower room bound from ceiling, wall, or visible-depth evidence and the room-height prior. We place the layout floor at $Y = 0$ without horizontally recentering the camera, giving

$$d_L = -\widehat{h}_f e_y = (0, -\widehat{h}_f, 0)^\top, \qquad x^L = R_{LC}x^C + d_L, \tag{11}$$

where $e_y = (0, 1, 0)^\top$. Thus $d_L$ is also the camera center in layout coordinates. The centered mesh pose in layout coordinates is

$$\begin{aligned} R_i^m &= R_{LC}C_{\text{cam}}\widehat{R}_i^\top C_{\text{asset}}, \\ t_i^m &= R_{LC}C_{\text{cam}}\widehat{t}_i + d_L + R_i^m(\widehat{s}_i o_i), \qquad s_i^m = \widehat{s}_i. \end{aligned} \tag{12}$$

The transpose converts row-vector rotations; the center-offset term preserves the centered mesh's position. The renderer uses the inverse of Eq. 11 for the source camera, preserving the joint camera/asset projection. The resulting assembly is $\mathcal{S}_{\text{metric}} = \{(V_i, t_i^m, R_i^m, s_i^m)\}_{i=1}^N$.

### B.2 Finite Support Regions

Following HSM and SceneSmith (Pun et al., 2026; Pfaff et al., 2026), we derive support regions from upward-facing, approximately horizontal parent-mesh triangles, not a parent bounding box.

We group triangles by height and union their planar projections. This retains holes and disconnected components, so distinct shelves are not bridged by a convex hull.

For movable objects, each horizontal region is stored in parent-local coordinates as

$$\Pi_k = (h_k, \Omega_k), \tag{13}$$

where $h_k$ is local height and $\Omega_k \subset \mathbb{R}^2$ is the finite $XZ$ footprint, possibly with holes. The floor is represented as a horizontal support region, with a finite footprint defined by the estimated room boundaries. Regions move with their parent during optimization.

### B.3 Relation-to-Surface Association

For an object parent, we select the region most compatible with the initialized child. Let $h_b$ be the child's bottom height in parent coordinates, $h_k$ the candidate layer height, and $d_k$ the signed lateral distance of the initialized child center to $\Omega_k$ (positive outside). With parent scale $s_p^0$, the association score is

$$C_k = s_p^0 |h_k - h_b| + 2 s_p^0 [d_k]_+. \tag{14}$$

The minimum-score admissible region is selected; its footprint for child $i$ is denoted $\Omega_i = \Omega_{k(i)}$ in the optimization loss. A one-time feasibility check may replace a parent whose initialized horizontal gap to the child exceeds $0.5$ m, but only if a larger, nearby object has an admissible mesh region and the replacement remains acyclic. Parent and region assignments are then frozen.

## C Joint Optimization Details

### C.1 Parameters and Contact Evaluation

Using the residuals defined in Section 3.3, let $(t_p, R_p, s_p)$ be the current parent transform, or $(0, I, 1)$ for a room parent. For region height $\widetilde{h}_i$ and initial parent-local anchor $(u_i^0, v_i^0)$, the world anchor is

$$a_i = t_p + s_p R_p (u_i, \widetilde{h}_i, v_i)^\top, \qquad (u_i, v_i) = (u_i^0, v_i^0) + (\Delta u_i, \Delta v_i). \tag{15}$$

For the horizontal supports considered here, $h_i = [a_i]_y$ is the current support height. Rotation is composed in the parent frame as

$$R_i = R_p R_y(-\Delta\psi_i) R_{i|p}^0, \tag{16}$$

where $R_i^0$ and $R_p^0$ are the child and parent orientations in $\mathcal{S}_0$, $R_{i|p}^0 = (R_p^0)^\top R_i^0$ is their relative orientation, and $R_y$ denotes rotation about the layout $Y$ axis. Scale is uniform, $s_i = s_i^0 \exp(\sigma_i)$. For the centered mesh $V_i$, the quantity in Eq. 1 is $y_{\min,i} = \min_{v \in V_i} [s_i R_i v]_y$. Objects are evaluated in parent-before-child order, so parent changes propagate to descendants. Equation 1 then eliminates vertical translation at every pose and scale; finite-region validity remains a separate loss. Wall-mounted objects and objects suspended from the ceiling are projected onto the nearest wall and ceiling surfaces, respectively, while rugs are excluded from collision computations.

### C.2 Visual Evidence

The mask loss renders all assets jointly (Laine et al., 2020); instance ownership and occlusion therefore determine $\alpha_i$. In the depth loss, $P_i$ contains masked metric-depth points after trimming boundary-depth outliers. A transformed mesh sample enters $Q_i$ only if it projects to a pixel owned by $i$ and agrees with the rendered depth. Visibility selection is discrete, but the retained 3D samples remain differentiable with respect to pose. Fixed full-surface samples are used if no sample passes the visibility test.

### C.3 Collision Evaluation

Broad-phase bounds discard disjoint pairs. Surface and interior probes supply $X_a$ in Eq. 5. The tangent term measures horizontal escape when contact constraints eliminate vertical translation. For each probe $x \in X_a$ inside $b$, cast 16 evenly spaced horizontal rays and select the nearest exit point

$h_{ab}(x)$ among their first intersections with the mesh. Let $n_{ab}^{\parallel}(x)$ be the normal of the intersected triangle, projected onto the horizontal plane, normalized, and oriented toward the exit ray. The local contact-cut distance and penalty are

$$\begin{aligned} e_{ab}^{\tan}(x) &= \left[(h_{ab}(x)-x)^{\top} n_{ab}^{\parallel}(x)\right]_+, \\ r_{ab}^{\tan}(x) &= \frac{[e_{ab}^{\tan}(x)-\epsilon_{\mathrm{col}}]_+}{d_0}, \qquad \epsilon_{\mathrm{col}} = 1\,\mathrm{mm}, \\ \ell_{ab}^{\tan} &= \frac{1}{|X_a|}\sum_{x\in X_a} r_{ab}^{\tan}(x)^2 + \frac{1}{4}\max_{x\in X_a} r_{ab}^{\tan}(x)^2. \end{aligned} \tag{17}$$

For nonpenetrating probes or rays without an exit, $e_{ab}^{\tan} = 0$. The selected exit and normal are held fixed within a gradient step and refreshed at the next evaluation, giving a locally differentiable contact cut rather than an exact planar SDF. This term is applied only to independent objects with compatible horizontal translation planes; it is omitted for support ancestor-descendant pairs. Its relative weight in Eq. 6 is $\beta = 0.2$.

### C.4 Finite Support, Room Bounds, and Initialization Prior

For a closed mesh, the support point is its uniform-density center of mass; for an open mesh, we use the area-weighted surface center. We evaluate the support point's horizontal projection relative to $\Omega_i$ using $d_{\Omega,i}$, in the parent's current local frame. Positive distance means outside $\Omega_i$, including inside one of its holes. The inward margin is $m = 10\,\mathrm{mm}$ and the softplus transition is $\tau = 5\,\mathrm{mm}$. This is a differentiable center-containment proxy; simulation separately tests dynamic stability.

**Room bounds.** Let $l, h \in \mathbb{R}^3$ be the room's lower and upper corners and $x_i(v)$ the world-space position of vertex $v \in V_i$. The nonnegative violation vector $b(x) = \max(l - x, x - h, 0)$ is defined componentwise, with $b_k$ its $k$th coordinate. We penalize its mean and worst value:

$$\mathcal{L}_{\mathrm{room}} = \frac{1}{N_{\mathrm{opt}} d_0^2}\sum_{i\in\mathcal{I}_{\mathrm{opt}}}\left[\frac{1}{3|V_i|}\sum_{v\in V_i}\|b(x_i(v))\|_2^2 + \frac{1}{4}\max_{v\in V_i,k} b_k(x_i(v))^2\right]. \tag{18}$$

**Initialization prior.** Let $\delta_i = s_p(u_i - u_i^0, 0, v_i - v_i^0)^{\top}$ be horizontal displacement relative to the current parent and $\bar{R}_i = R_p R_{i|p}^0$ the initial child orientation transported by that parent. We regularize translation, orientation, and scale using reference tolerances $d_{\mathrm{reg}}$, $\theta_{\mathrm{reg}}$, and $s_{\mathrm{reg}} > 1$:

$$\mathcal{L}_{\mathrm{prior}} = \frac{1}{N_{\mathrm{opt}}}\sum_{i\in\mathcal{I}_{\mathrm{opt}}}\left[\frac{\|\delta_i\|_2^2}{d_{\mathrm{reg}}^2} + \frac{\|R_i - \bar{R}_i\|_F^2}{8\sin^2(\theta_{\mathrm{reg}}/2)} + \left(\frac{\log(s_i/s_i^0)}{\log s_{\mathrm{reg}}}\right)^2\right]. \tag{19}$$

Here $\|\cdot\|_F$ is the Frobenius norm, $s_i^0$ is the initial uniform scale, and the reference tolerances are $d_{\mathrm{reg}} = 0.15\,\mathrm{m}$, $\theta_{\mathrm{reg}} = 15^\circ$, and $s_{\mathrm{reg}} = 1.2$.

### C.5 Optimization Schedule and Dense Audit

Algorithm 1 uses the quantities defined above, with $\mathcal{B}$ denoting room bounds and $\Pi$ the relation-to-region assignments. $\mathcal{C}_r$ is the audited collision-pair set, $\mathcal{A}_r$ its movable objects, and $\mathcal{W}_r$ the collision samples fixed within local round $r$.

**Algorithm 1: Metric projection and support-constrained scene optimization**

**Input:** $(I, \{M_i\}, D, K, \mathcal{G}, \mathcal{S}_{\text{init}})$.
**Output:** $(\mathcal{S}_{\text{opt}}, \mathcal{C}_{\text{final}})$.

$(T_{C\to L}, \mathcal{B}) \leftarrow \text{FitRoom}(I, D, K)$
$\mathcal{S}_{\text{metric}} \leftarrow \text{Convert}(\mathcal{S}_{\text{init}}, T_{C\to L})$
$\Pi \leftarrow \text{Bind}(\mathcal{G}, \mathcal{S}_{\text{metric}}, \mathcal{B})$
$\mathcal{S}_0 \leftarrow \text{Project}(\mathcal{S}_{\text{metric}}, \Pi)$; $\boldsymbol{q}_0 \leftarrow 0$
$\{\boldsymbol{q}_k\}_{k=1}^{320} \leftarrow \text{Adam}(\mathcal{L}, \boldsymbol{q}_0; \mathcal{I}_{\text{opt}})$
$k^\star \leftarrow \arg\min_{0\le k\le 320} \mathcal{L}(\boldsymbol{q}_k)$; $\boldsymbol{q}^\star \leftarrow \boldsymbol{q}_{k^\star}$; $\mathcal{S}_{\text{joint}} \leftarrow \mathcal{S}(\boldsymbol{q}^\star)$; $\mathcal{S} \leftarrow \mathcal{S}_{\text{joint}}$
**for** $r = 1, 2, 3$ **do**
  $\mathcal{C}_r \leftarrow \text{Audit}(\mathcal{S})$; **if** $\mathcal{C}_r = \varnothing$ **then break**
  $\mathcal{A}_r \leftarrow \mathcal{I}_{\text{opt}} \cap \bigcup_{\{i,j\}\in\mathcal{C}_r} \{i, j\}$
  $\mathcal{W}_r \leftarrow \text{Sample}(\mathcal{S})$; $\boldsymbol{q}_{r,0} \leftarrow \boldsymbol{q}^\star$
  $T_r = 80$; $\{\boldsymbol{q}_{r,k}\}_{k=1}^{T_r} \leftarrow \text{Adam}(\mathcal{L}(\,\cdot\,; \mathcal{W}_r), \boldsymbol{q}_{r,0}; \mathcal{A}_r)$
  $k_r^\star \leftarrow \arg\min_{0\le k\le T_r} \mathcal{L}(\boldsymbol{q}_{r,k}; \mathcal{W}_r)$; $\boldsymbol{q}^\star \leftarrow \boldsymbol{q}_{r,k_r^\star}$; $\mathcal{S} \leftarrow \mathcal{S}(\boldsymbol{q}^\star)$
**end for**
$\mathcal{S}_{\text{opt}} \leftarrow \mathcal{S}$; $\mathcal{C}_{\text{final}} \leftarrow \text{Audit}(\mathcal{S})$
**return** $(\mathcal{S}_{\text{opt}}, \mathcal{C}_{\text{final}})$

Each local round uses the complete objective of Eq. 2. Its fixed object-local probe set $\mathcal{W}_r$ combines 8192 area-weighted surface samples per mesh with all vertices and triangle centroids. Joint/local learning rates are $0.015/0.003$, with cosine decay and gradient-norm clipping at 10. The contact gap is $g = 1\,\text{mm}$; residual bounds are $\pm 0.5\,\text{m}$ horizontally and $\pm 40^\circ$ in yaw, with scale ratio $s_i/s_i^0 \in [0.65, 1.5]$.

# D Simulator-Guided Stabilization

## D.1 Tools and Edit Scope

The agent can inspect surfaces and coordinate frames, preview pose edits, and create working copies of candidate scenes. Each pose edit uses the inspected world, object-local, or support frame and pivot, with translations in metres and axis-angle rotations in degrees. Parent edits transport support descendants; a working copy permits up to 12 edits before a joint audit. Ordinary objects stay on the same horizontal support patch, with tangent translation, normal-axis rotation, and dependent height. Wall/ceiling installations, rugs, and objects lacking valid geometry or support bindings are ineligible for direct edits. Meshes, scales, and scene graph edges are immutable; every changed pose stays within 20 cm and $30^\circ$ of $B_0$.

For loose objects like pillows, local rigid-body settling proposes a pose with the surroundings held static. The first trial uses the complete current pose; at most two retries per object may shift the object's initial position by at most 5 cm or set sliding friction in $[0.2, 1.2]$. A local settling result must be verified by re-simulating the updated scene before acceptance.

## D.2 Audit, Acceptance, and Budgets

The paired experiment uses 5 s simulations with the Full baseline's frozen colliders, mass, inertia, gravity, installation policy, and normalization (Appendix E). Strict diagnostics use native metric units: displacement $\le 1$ cm, rotation $\le 5^\circ$, retained terminal contact with the declared support, and final-0.5 s RMS linear/angular speeds $\le 0.02$ m/s and 0.1 rad/s. For an unstable object, the motion score is $m_i = \max\{\Delta_i^{\text{pos}}/(0.01\,\text{m}), \Delta_i^{\text{ang}}/5^\circ\}$. Acceptance requires at least one strict recovery or a 10% motion-score decrease relative to $B_t$, while preserving strict and benchmark-stable objects and retained support from both $B_t$ and $B_0$. A motion decrease alone does not count as a recovered benchmark-stable object.

All preservation budgets below are cumulative relative to $B_0$ and apply to settling too. The camera, source references, object set, and support bindings remain fixed. Inherited violations within these budgets alone are not grounds to reject a repair.

Table 7: Geometric and visual preservation budgets for agentic refinement, measured relative to the initial scene $B_0$.

| Quantity | Maximum regression |
|---|---|
| Per-object room/support distance violation | $+5$ mm |
| Per-pair penetration depth (convex proxies) | $+2$ mm |
| Mean mask IoU (one-pixel boundary tolerance / raw) | $-0.01$ / $-0.005$ |
| Visible source-depth CD (per object / mean) | $+5$ mm / $+1$ mm |

Source-depth CD averages the two directional mean Euclidean distances to the input-depth observations, in metres. Intermediate working-copy edits may violate visual budgets, but the complete candidate must pass before acceptance. Rejection retains $B_t$. The run stops when the agent decides to stop, all objects pass the primary benchmark criterion, or the maximum number of tool calls is reached.

# E Evaluation Details

## E.1 Reference Geometry

On 3D-FRONT and ScanNet++, CD and $F_5$ evaluate object surfaces visible from the input-image viewpoint. Reference visibility is determined by ScanNet++'s annotated camera parameters or the virtual camera parameters used to render our manually selected 3D-FRONT views. Reference objects are selected from GT annotations, excluding architectural structures and invalid labels, with at least $50$ visible image pixels per object. All methods use the same reference objects, independently of their detection masks.

Reference and reconstructed points are sampled from depth maps in the input view, with at most $30000$ points per set. Predictions exclude clearly labelled architectural objects; objects with uncertain labels are retained.

Before scoring, we align each reconstruction to its reference with a single scene-level similarity transform for scale, rotation, and translation. Image correspondences initialize the transform, which is refined by robust ICP; global registration is used when too few correspondences are available. The resulting transform is shared by all objects in the scene.

For reference/predicted samples $P, Q$ and $d(x, Y) = \min_{y \in Y} \|x - y\|_2$,

$$\mathrm{CD}(P, Q) = \frac{1}{|P|} \sum_{p \in P} d(p, Q) + \frac{1}{|Q|} \sum_{q \in Q} d(q, P). \tag{20}$$

CD is measured in metres and uses unsquared Euclidean distances. $F_5$ is the harmonic mean of the two fractions of points with a nearest neighbour strictly closer than $0.05$ m. B-IoU forms each predicted object's AABB from all its registered mesh vertices, and each selected reference object's AABB from its full CAD mesh (FRONT) or all annotated scan vertices (ScanNet++). Using a class-agnostic Hungarian assignment $\mathcal{A}$ that maximizes summed box IoU, we report

$$\text{B-IoU} = \frac{\sum_{(i,j) \in \mathcal{A}} \mathrm{IoU}(B_i, \widehat{B}_j)}{\max(N_{\mathrm{pred}}, N_{\mathrm{GT}})}. \tag{21}$$

Unmatched instances contribute zero. All metrics are computed per scene and then averaged with equal scene weights, with valid counts reported per method.

## E.2 Physical Model and Metrics

All submitted meshes are exported to right-handed $Z$-up coordinates. For their joint bounds $\boldsymbol{\ell}, \boldsymbol{u}$, evaluation applies the same scene-level transformation to every vertex of every object:

$$s_{\mathrm{eval}} = \frac{4\,\mathrm{m}}{\max(u_x - \ell_x, u_y - \ell_y)}, \qquad \boldsymbol{v}_{\mathrm{eval}} = s_{\mathrm{eval}} \left( \boldsymbol{v} - \begin{bmatrix} (\ell_x + u_x)/2 \\ (\ell_y + u_y)/2 \\ \ell_z \end{bmatrix} \right). \tag{22}$$

A global translation aligns the scene's lowest vertex with the ground plane at $z = 0$. MuJoCo settings follow Section 4. The model has a common ground plane, free objects, and no reconstructed wall/ceiling colliders. CoACD uses at most 32 parts, or 256 for storage objects, with the same selection rule for every method; whole-object convex hulls determine mass and inertia.

Collision involvement is evaluated before simulation using the original meshes. BVH queries detect triangle intersections, while sampled signed-distance queries estimate penetration depths. Let $\mathcal{P}_c$ contain pairs exceeding 1 mm penetration or having an open-mesh intersection without quantified depth. Rug-involving pairs are excluded, but all objects $\mathcal{O}$ remain in the denominator:

$$C_{\text{obj}} = \frac{|\bigcup_{(i,j)\in\mathcal{P}_c}\{i,j\}|}{|\mathcal{O}|}. \tag{23}$$

The five-second stability indicator of object $i$ is

$$\eta_i = \mathbf{1}[\Delta_i^{\text{pos}} \le 0.1\,\text{m}]\,\mathbf{1}[\Delta_i^{\text{ang}} \le 10^\circ]\,\mathbf{1}[v_i(5\,\text{s}) \le 0.1\,\text{m/s}]. \tag{24}$$

Displacement and rotation are measured from the release pose; $v_i$ is terminal linear speed. For each scene, stability is the fraction of objects with $\eta_i = 1$. We report the average of these fractions across scenes, giving each scene equal weight.

## F Individual Loss Ablations

To isolate each term's role within the complete objective, we remove one loss at a time and retain the other five (Table 8). These experiments complement the cumulative study in Section 4.4 by measuring the contribution of each term when all other terms are present. We evaluate the same 56 scenes with the geometric and physical metrics defined in Appendix E.

Table 8: Individual-loss removal on 56 scenes and 963 submitted objects per variant. Full includes all six loss terms; each remaining row removes one term.

| | | | *3D-FRONT* | | *ScanNet++* | |
|---|---|---|---|---|---|---|
| Variant | Coll. (%) ↓ | Stable (%) ↑ | CD ↓ | $F_5$ ↑ | CD ↓ | $F_5$ ↑ |
| Full | 9.28 | 89.49 | 0.2253 | 0.6975 | 0.2103 | 0.5534 |
| w/o mask | 8.13 | 89.04 | 0.2254 | 0.6988 | 0.2101 | 0.5638 |
| w/o depth | **6.52** | 89.51 | 0.2663 | 0.5612 | 0.2234 | 0.5200 |
| w/o collision | 29.09 | 77.54 | **0.2200** | **0.7208** | **0.2054** | **0.5641** |
| w/o finite support | 7.19 | 88.87 | 0.2224 | 0.7067 | 0.2083 | 0.5531 |
| w/o room | 8.21 | **89.52** | 0.2233 | 0.7074 | 0.2091 | 0.5603 |
| w/o prior | 7.94 | 88.95 | 0.2226 | 0.7083 | 0.2081 | 0.5601 |

Mask supervision improves source-view silhouette agreement: removing it lowers mean mask IoU by 0.0035. Depth is essential for metric geometric alignment. Removing depth increases CD from 0.2253 to 0.2663 m and reduces $F_5$ from 0.6975 to 0.5612 on 3D-FRONT. On ScanNet++, CD increases from 0.2103 to 0.2234 m and $F_5$ falls from 0.5534 to 0.5200.

The prior controls deviation from the initialized layout. Removing it increases mean center displacement and rotation from initialization by 1.71 cm and 9.52°, respectively, while reducing stability from 89.49% to 88.95%. Finite support enforces placement within the assigned supporting regions, while the room term enforces global containment. Removing these terms increases support-domain violations by 5.58 percentage points and mean maximum room-boundary violation by 103.97 mm, respectively. These diagnostics establish their roles in preserving layout structure and valid placement.

Collision is the dominant term for physical reliability. Removing it raises collision involvement by 19.81 percentage points and lowers stability by 11.95 points. The lower CD and higher $F_5$ of this variant also expose the geometry-physics trade-off: the image-aligned variant retains intersecting objects, and explicit collision resolution directly targets these incompatible placements.

## G Additional Qualitative Results

Figure 6 extends Figure 3 to include all six baselines. Figures 7 and 8 compare the same methods across ten additional indoor scenes.




Figure 6: **Full qualitative comparison for the five scenes in Figure 3.** Each column shows the reference, six baselines, and ImScene (bottom). Red dashed boxes mark contact inconsistencies in SAM 3D.




Figure 7: **More qualitative results (I).** Comparisons across five indoor scenes.




Figure 8: **More qualitative results (II).** Comparisons across five additional indoor scenes.