PHYSRECON: SIMULATION-IN-THE-LOOP SCENE RECONSTRUCTION FOR PHYSICALLY CREDIBLE EMBODIED DATA GENERATION FROM SINGLE IMAGES

Anonymous authors
Paper under double-blind review

ABSTRACT

Single-view 3D reconstruction and 6D pose estimation often produces visually plausible yet physically infeasible scenes, such as floating objects, unstable stacking and unreasonable object scales, while suffering from incomplete geometry caused by visual occlusion. To address these issues, we propose PhysRecon, a physics-guided closed-loop framework for physically credible single-image 3D scene reconstruction. Different from conventional purely visual reconstruction methods, our pipeline integrates occlusion-aware visual completion and physics-based iterative refinement. Specifically, we adopt an occlusion-aware inpainting module to recover complete object geometry under occluded observations. Instead of directly predicting object mass, our framework estimates physical density and calculates mass via reconstructed mesh volume, enabling physically consistent parameter modeling. We further design a failure-driven local sampling strategy combined with quasi-static physical simulation and physical violation auditing, which iteratively corrects unreasonable object scales and poses while freezing well-reconstructed objects. Evaluated on HomebrewedDB and LM-O benchmarks, our method outperforms state-of-the-art visual reconstruction baselines in both geometric reconstruction accuracy and 6D pose estimation. Extensive quantitative ablation and qualitative visualization further validate the effectiveness of each core component, demonstrating that PhysRecon can generate visually faithful and physically reasonable 3D scenes from a single RGB image.

1 INTRODUCTION

Single-image 3D reconstruction and 6D object pose estimation are fundamental tasks in computer vision, with wide applications in robotics, augmented reality, and physical simulation. Modern learning-based methods can recover object geometry and pose from a single RGB image, yet they suffer from two critical limitations. First, occluded object regions lead to incomplete or distorted mesh predictions, as the network cannot directly observe the hidden surfaces from the input view. Second, most existing reconstruction pipelines operate purely in the visual space without considering physical constraints. The recovered 3D scenes often contain physically implausible configurations: objects may float in the air, interpenetrate each other, or have unreasonable relative scales, even if the rendered mesh looks visually plausible. Such physically invalid reconstructions cannot be directly deployed in downstream physical simulation and robotic interaction tasks.

To tackle these challenges, we propose **PhysRecon**, a physically informed single-image 3D reconstruction framework. Instead of embedding all physical priors into a neural network, our framework adopts a *perception-then-physics-validation* paradigm. We first perform visual reconstruction with occlusion-aware inpainting to recover complete object geometry from a single RGB image. Then we perform pose and scale sampling, and leverage an external physics simulator as an agent for physical feasibility auditing. The simulator conducts quasi-static rollouts to detect physical violations, and feeds the physical validity signal back to refine the object scale and 6D pose. Different from prior work that directly regresses object mass, we estimate object density and compute mass from reconstructed volume, which brings more stable and physically consistent parameter estimation. Moreover, our pipeline supports curved contact surfaces instead of being limited to planar contact assumptions, and estimates physical properties for both objects and their contact interfaces.

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.

We validate PhysRecon on the HomebrewedDB and LM-O datasets. Experimental results show that our method achieves competitive geometric and pose estimation metrics, and significantly improves the physical plausibility of reconstructed scenes compared with SOTA baselines RecGen and Any6D. Extensive ablation studies confirm the contribution of each key component of our pipeline.

Our contributions are summarized as follows:

1. We propose a novel perception-and-simulation closed-loop framework PhysRecon for single-image physically valid 3D reconstruction. Instead of training a neural network with all physical constraints embedded, we use an external physics simulator as a physical checker to iteratively refine object pose and scale.
2. We design a physics-aware optimization pipeline for reconstruction. Our pipeline estimates object density (rather than mass directly) and computes mass from reconstructed volume; it supports curved contact surfaces and estimates physical attributes for both objects and contact interfaces, relaxing the common planar contact assumption.
3. We introduce an occlusion-aware inpainting module to recover plausible visual information for occluded regions, which effectively mitigates mesh fragmentation and broken geometry caused by occlusion and enables high-quality reconstruction for complex concave and hollow objects.

2 Related Work

Single-image generative 3D object reconstruction. Single-image-to-3D models produce object meshes or neural fields from RGB observations (Hong et al., 2024; Xu et al., 2024; Li et al., 2024). These approaches excel at generating plausible object shapes but typically output geometry in an arbitrary canonical coordinate frame without metric scale or global scene layout Li et al. (2026). They treat each object independently and do not reason about relative poses, mutual occlusion, or inter-object physical constraints. Recent feed-forward scene-level generation methods Bahmani et al. (2026); Wang et al. (2026); Gupta et al. (2026) lift single images into full 3D scenes. Nevertheless, they mostly optimize for visual realism rather than physical plausibility; floating objects and inter-penetration artifacts are common when exporting for embodied simulation. While prior works train neural networks jointly with physical losses Ma et al. (2026), our work keeps all perception backbone weights frozen. We delegate physical validation entirely to an external rigid-body simulator and perform only local pose-and-scale sampling, avoiding multi-constraint joint training.

Category-agnostic segmentation and amodal occlusion completion. Foundation-model-based automatic mask generators such as SAM and MobileSAM enable category-agnostic object proposal without pre-defined object lists Kirillov et al. (2023); Zhang et al. (2023). However, raw mask proposals contain fragments, background patches and overlapping regions; merging these proposals into valid object instances remains non-trivial Tai et al. (2025). Amodal segmentation and generative inpainting predict or hallucinate object regions behind occluders Ao et al. (2023); Lee & Park (2022). Most existing amodal pipelines produce expanded masks but do not tightly couple the completed silhouette as hard conditioning for downstream 3D reconstruction. Our PhysRecon builds upon this line: we perform true-occlusion reasoning to trigger generative inpainting, feed the expanded silhouette as the alpha-channel condition to object-wise reconstruction, and explicitly track geometric reliability meta-information visib_fract and alpha_over_mask for later simulation-in-the-loop refinement.

6D object pose estimation under limited input. Traditional BOP-benchmark 6D pose estimation assumes known CAD models, object instance lists and often RGB-D inputs Hodan et al. (2018); Wen et al. (2024); Nguyen et al. (2024). Category-level pose estimation relaxes exact instance requirements but still relies on categorical priors Nguyen et al. (2023). More recent category-agnostic methods estimate object pose for arbitrary novel objects without given CAD templates Lee et al. (2025). Nevertheless, nearly all existing pose pipelines start from segmented object masks and solve pose given reconstructed or provided geometry. A key challenge for monocular RGB-only input is the well-known radial scale ambiguity: object scale and depth are coupled along viewing rays Zadaianchuk et al. (2026). Direct evaluation of individual scale or translation components is ill-defined. Our work follows this observation: we solve initial scale via closed-form from monocular depth, clamp subsequent scale updates within a bounded local window, and evaluate quality using observation-space metrics rather than raw parameter errors.

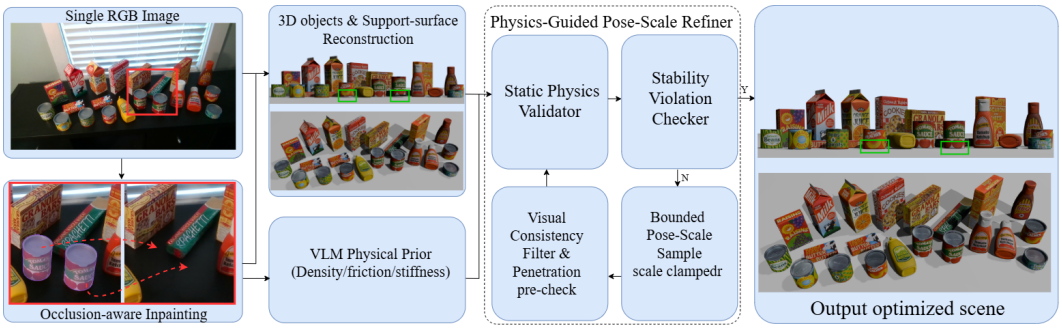


Figure 1: Overview of the PhysRecon pipeline. Starting from a single RGB image, we conduct occlusion-aware inpainting and then reconstruct 3D meshes for objects and planar supporting surfaces. A VLM provides physical priors (density, friction, stiffness); mass is derived as $m = \rho V$ from density and reconstructed volume. The downstream refinement is a failure-driven local sampling loop. The quasi-static physics oracle relaxes the scene to steady state, followed by a physical violation auditor. If invalid objects are detected, bounded truncated-Gaussian pose-scale sampling is applied to generate candidate configurations only for failed objects. Candidates go through penetration pre-filtering and visual consistency gate before re-entering physics simulation. The iteration stops once all physical and visual constraints are satisfied, outputting the physically valid 3D scene.

Physically-plausible and simulation-in-the-loop scene reconstruction for embodied AI. Embodied AI heavily depends on simulation-ready synthetic scenes He et al. (2026); Ling et al. (2026). Many reconstruction pipelines output visually appealing scenes yet yield physically invalid configurations when imported into simulators: floating objects, support penetration and unstable stacking are frequent failure modes Puig et al. (2024); Shen et al. (2021). Several recent works introduce physics-aware optimization or agentic loops to improve physical consistency. REST3D Ma et al. (2026) optimizes pose with differentiable physical energy terms. NeoWorld-Pro He et al. (2026) uses MLLM to generate scene programs and runs iterative physics-in-the-loop refinement for articulating objects. Scenethesis Ling et al. (2026) employs LLM-driven layout generation followed by contact-constrained pose optimization. Different from differentiable-physics training or end-to-end agent training, our approach treats Isaac-Sim strictly as a black-box validator oracle. We execute quasi-static rollout and violation audit, collect violation signals, and perform local Monte-Carlo sampling over pose-scale candidates while keeping mesh geometry fixed. We never adopt simulator-settled poses as outputs; visual alignment to input RGB remains a hard filter for every candidate. Unlike prior simulation-in-the-loop designs, we omit expensive outer-loop re-reconstruction and limit refinement strictly to an inner-loop over pose and scale, trading full expressive power for better computational tractability.

3 METHOD

3.1 OVERVIEW

Given a single RGB image as input, our PhysRecon framework first performs occlusion-aware image inpainting to complete occluded visual regions. The inpainted image is fed into a 3D reconstruction module to recover meshes for individual objects as well as planar supporting surfaces, where mesh geometry remains fixed in later optimization. Meanwhile, a vision-language model estimates physical priors including density, friction and stiffness for each object and contact surface. Instead of directly predicting mass, we compute mass via $m = \rho V$ using the reconstructed object volume V and predicted density ρ.

We then adopt a failure-driven local sampling loop for physics-aware refinement. We run quasi-static rollout on the current scene in the physics oracle, and the physical violation auditor identifies physically invalid objects. For each failed object, we generate candidate configurations via truncated Gaussian bounded sampling: we perturb its scale, translation and rotation around its current state within predefined bounds, while freezing all other objects in the scene. We sample N candidates per failed object to form the candidate set. Each candidate is filtered by a penetration pre-check to reject severe interpenetration, then passed through a visual consistency gate to discard candidates with

poor alignment to the input RGB image. Valid candidates are fed back into quasi-static simulation for the next iteration. The loop terminates when all objects pass the audit or the maximum iteration budget is exhausted, producing the final physically plausible 3D reconstruction. An illustration of the full pipeline is shown in Figure 1.

3.2 Perception Pipeline: Initial Scene Reconstruction and Physical Prior Estimation

Occlusion-aware Image Inpainting. We take a single RGB image $\boldsymbol{I} \in \mathbb{R}^{H\times W\times 3}$ as input. Occluded regions are detected via instance segmentation masks. An occlusion-aware inpainting model fills missing pixel values inside masked occluded regions, producing the completed image $\hat{\boldsymbol{I}}$. This step reduces reconstruction ambiguity caused by self-occlusion and inter-object overlap. The inpainted image $\hat{\boldsymbol{I}}$ is used as the input for the 3D reconstruction module.

Object and Planar Supporting Surface Reconstruction. From $\hat{\boldsymbol{I}}$, we reconstruct triangle meshes $\{M_o\}_{o\in\mathcal{O}}$ for each object and fit planar supporting surfaces. Each mesh stores fixed vertex positions and face topology. Throughout the subsequent iterative refinement, mesh geometry (vertex coordinates and connectivity) is fixed. Only per-object scale s, translation $\boldsymbol{t} \in \mathbb{R}^3$, and rotation $R \in SO(3)$ are optimized. The camera intrinsics K and extrinsics $[\boldsymbol{R}_{cam}, \boldsymbol{t}_{cam}]$ are estimated together with reconstruction and kept constant in all re-projection checks.

VLM Physical Prior Estimation. A vision-language model takes the input RGB image and per-object crop images as prompts to predict density ρ_o, contact friction coefficient μ_o, and contact stiffness k_o for each object o. The VLM also estimates physical parameters for planar supporting surfaces. We avoid direct mass prediction. Object mass is computed from density and mesh volume:

$$m_o = \rho_o V_o \tag{1}$$

where V_o is the volume of mesh M_o, precomputed once at initialization and unchanged during iterations, since mesh geometry is fixed. The set of physical parameters $\{\rho_o, \mu_o, k_o, m_o\}$ is passed to the physics oracle.

3.3 Failure-Driven Local Sampling Refinement Loop

The core of PhysRecon is an iterative closed-loop refinement. Instead of sampling all objects jointly, we only update objects flagged as physically invalid by the auditor.

3.3.1 Quasi-Static Physics Oracle

Given the current scene configuration $\mathcal{S} = \{(s_o, \boldsymbol{t}_o, R_o, M_o, m_o, \mu_o, k_o)\}_{o\in\mathcal{O}}$ and planar supports, the physics oracle performs quasi-static rollout. Gravity is ramped up gradually from zero to $\boldsymbol{g} = (0, 0, -9.81)\,\text{m/s}^2$. The simulation runs until all rigid bodies reach steady state or reaches a maximum simulation time budget T_{sim}. A steady-state condition is defined as:

$$\|\boldsymbol{v}_o\| < \epsilon_v, \quad \|\boldsymbol{\omega}_o\| < \epsilon_\omega, \quad \forall o \in \mathcal{O} \tag{2}$$

where $\boldsymbol{v}_o$ is linear velocity, $\boldsymbol{\omega}_o$ angular velocity, $\epsilon_v, \epsilon_\omega$ are small velocity thresholds. The steady-state scene $\mathcal{S}_{\text{steady}}$ is forwarded to the physical violation auditor.

3.3.2 Physical Violation Auditor

The auditor marks object o as invalid if **either** of the two criteria is violated:

1. **Static stability criterion**: If $\|\boldsymbol{v}_o\| \geq \epsilon_v$ or $\|\boldsymbol{\omega}_o\| \geq \epsilon_\omega$, the object continues sliding or tipping and is invalid.
2. **Visual re-projection criterion**: Render $\mathcal{S}_{\text{steady}}$ with the fixed camera $K, [\boldsymbol{R}_{cam}, \boldsymbol{t}_{cam}]$ to obtain rendered mask $\hat{\mathcal{M}}_o$. Compute mask IoU against the input image instance mask $\mathcal{M}_o$:

$$\text{IoU}_o = \frac{|\hat{\mathcal{M}}_o \cap \mathcal{M}_o|}{|\hat{\mathcal{M}}_o \cup \mathcal{M}_o|}. \tag{3}$$

If $\text{IoU}_o < \tau_{\text{iou}}$, the object's pose deviates from image observation and is marked invalid.

Let $\mathcal{O}_{\text{fail}} \subseteq \mathcal{O}$ denote the set of failed objects. If $\mathcal{O}_{\text{fail}} = \emptyset$, iteration terminates and output $\mathcal{S}_{\text{steady}}$. Otherwise, proceed to bounded pose-scale sampling.

3.3.3 BOUNDED POSE-SCALE CANDIDATE SAMPLING (IMPLEMENTATION DETAILS)

For each $o \in \mathcal{O}_{\text{fail}}$, its state vector is

$$\boldsymbol{x}_o = \left[s_o, \boldsymbol{t}_o, R_o\right], \tag{4}$$

with sampling centroid $\boldsymbol{x}_{0,o} = \left[s_{0,o}, \boldsymbol{t}_{0,o}, R_{0,o}\right]$ taken from the steady state of the previous iteration. We use **truncated Gaussian sampling** for each component independently:

$$s \sim \mathcal{N}(s_{0,o}, \sigma_s^2), \qquad s \in [s_{0,o} - \Delta_s,\ s_{0,o} + \Delta_s],\ s > 0 \tag{5}$$

$$t_i \sim \mathcal{N}(t_{0,o,i}, \sigma_t^2), \qquad t_i \in [t_{0,o,i} - \Delta_t,\ t_{0,o,i} + \Delta_t], \quad i \in \{x, y, z\} \tag{6}$$

$$\theta_j \sim \mathcal{N}(0, \sigma_\theta^2), \qquad \theta_j \in [-\Delta_\theta, \Delta_\theta], \quad j \in \{\text{roll}, \text{pitch}, \text{yaw}\} \tag{7}$$

The sampled Euler angles $(\theta_x, \theta_y, \theta_z)$ are converted to a rotation matrix R_{pert}. The new object rotation is $R_{\text{new}} = R_{\text{pert}} R_{0,o}$, then projected onto $SO(3)$ via SVD orthogonalization to ensure $R_{\text{new}}^\top R_{\text{new}} = I$ and $\det(R_{\text{new}}) = 1$.

Hyperparameters: We set $\Delta_s = 0.15, \Delta_t = 0.08\,\text{m}, \Delta_\theta = 10^\circ$. The Gaussian standard deviations are set to $\sigma_s = \Delta_s/3, \sigma_t = \Delta_t/3, \sigma_\theta = \Delta_\theta/3$, such that the truncation bounds correspond to $\pm 3\sigma$. We sample $N = 12$ candidate states per failed object. When generating a candidate scene, **all non-failed objects keep their state unchanged**. Only the failed object's $(s, \boldsymbol{t}, R)$ is replaced by newly sampled values.

3.3.4 PENETRATION PRE-CHECK AND VISUAL CONSISTENCY GATE

Penetration Pre-Check. For each candidate scene, we compute pairwise penetration depth $d_p(o_i, o_j)$ between object meshes and $d_p(o, \text{plane})$ between objects and planar supporting surfaces. A candidate is discarded if:

$$\max\left(\max_{i,j} d_p(o_i, o_j), \max_{o} d_p(o, \text{plane})\right) > \tau_{\text{pen}} \tag{8}$$

where τ_{pen} is penetration threshold. This step filters invalid configurations before simulation to avoid ill-conditioned contact solving and reduce simulation cost.

Visual Consistency Gate. For candidates passing penetration test, we re-project the candidate scene to the input view and compute per-object mask IoU. A candidate is rejected if $\text{IoU}_o < \tau_{\text{iou,candidate}}$. The remaining valid candidates are fed to the quasi-static physics oracle for next iteration.

The iterative loop terminates under two conditions: (1) $\mathcal{O}_{\text{fail}} = \emptyset$; (2) reaching maximum iteration count $I_{\max}$. When hitting $I_{\max}$, we select the scene with highest average IoU and zero physical violations as final output.

4 EXPERIMENTS

4.1 DATASETS AND EVALUATION PROTOCOL

We evaluate our PhysRecon framework on two real-world object pose and reconstruction benchmarks: **HomebrewedDB** Kaskman et al. (2019) and **LM-O** Brachmann et al. (2014). We strictly follow the RecGen evaluation protocol Zadaianchuk et al. (2026). For fair comparison, all competing methods are evaluated on **identical per-(frame,object)** target lists. We run all methods locally with official released weights and compute metrics under the identical evaluation driver, rather than copying reported numbers from prior papers. All metrics are computed on the full target set without zero-padding for failed samples.

Evaluation Metrics. We adopt the standard metrics from RecGen protocol:

- CD_{norm}: Normalized Chamfer Distance (↓), measuring geometric mesh reconstruction error.
- **ADD-SB**: Symmetry-aware Average Distance of Points (↓), the core pose estimation metric for symmetric objects.
- **ADD-SB@τ**: Recall at threshold τ, higher values indicate better pose accuracy. We report @0.1 and @0.05.
- **DRE@0.05**: Depth Recall Error at threshold 0.05 (↑), evaluating depth alignment between reconstructed mesh and observation.

Table 1: Results on HomebrewedDB. Best in bold.

Method	CD_{norm}(↓)	ADD-SB(↓)	ADD-SB@0.1(↑)	ADD-SB@0.05(↑)	DRE@0.05(↑)
Any6D	0.071	0.115	68.27%	36.55%	36.55%
RecGen	0.034	0.050	**96.39**%	73.90%	51.00%
Ours	**0.026**	**0.043**	95.18%	**75.10**%	**51.00**%

Table 2: Results on LM-O. Best in bold.

Method	CD_{norm}(↓)	ADD-SB(↓)	ADD-SB@0.1(↑)	ADD-SB@0.05(↑)	DRE@0.05(↑)
Any6D	0.075	0.115	59.81%	25.23%	35.51%
RecGen	0.059	0.075	87.04%	**60.19**%	37.96%
Ours	**0.042**	**0.055**	**91.67**%	57.41%	**43.52**%

Baselines. We compare against two state-of-the-art single-view reconstruction & pose estimation methods:

1. **Any6D (InstantMesh)** Lee et al. (2025): Single-view 6D pose and mesh reconstruction baseline.
2. **RecGen (1-view)** Zadaianchuk et al. (2026): RecGen single-view reconstruction model, the primary baseline for our protocol.

4.2 Implementation Details

For perception preprocessing, we adopt an occlusion-aware image inpainting module to fill occluded regions masked by instance segmentation. The 3D reconstruction module outputs object meshes and planar supporting surfaces; mesh topology and vertex geometry remain fixed during the iterative physics refinement loop. We use a VLM to predict per-object physical priors: density, friction coefficient and contact stiffness. Object mass is computed via $m = \rho V$, where V is precomputed mesh volume. Our quasi-static physics oracle is built on NVIDIA Isaac Sim. Gravity is gradually ramped to $9.81\,\mathrm{m/s^2}$. The physics violation auditor uses linear/angular velocity thresholds and mask IoU to mark physically invalid objects. For bounded pose-scale sampling, we use truncated Gaussian perturbation over scale, translation and Euler rotation angles. We sample $N = 12$ candidate configurations for each failed object, while keeping all other objects fixed. We filter candidates by penetration pre-check and visual consistency gate before feeding valid candidates into the physics rollout. The maximum iteration budget is set to $I_{\max} = 15$. All hyperparameters are fixed across both datasets, without dataset-specific tuning.

4.3 Main Quantitative Results

Table 1 reports reconstruction and pose metrics on HomebrewedDB. Our PhysRecon achieves the best performance on CD_{norm} (0.026) and ADD-SB (0.043). Compared with RecGen, our method reduces Chamfer distance by 23.5% and ADD-SB error by 14.0%. Our ADD-SB@0.1 recall (95.18%) is slightly lower than RecGen, while ADD-SB@0.05 improves to 75.10%. DRE@0.05 is matched at 51.00%. Any6D shows substantially worse performance on all metrics, demonstrating the advantage of our physics-guided refinement.

Table 2 presents results on LM-O. Our method consistently outperforms both baselines on CD_{norm} (0.042) and ADD-SB (0.055). We obtain 91.67% ADD-SB@0.1 recall, surpassing RecGen (87.04%). On ADD-SB@0.05, our result (57.41%) is close to RecGen's 60.19%. Our DRE@0.05 reaches 43.52%, which outperforms both competitors. These results validate that our failure-driven local sampling loop improves geometric reconstruction and 6D pose estimation across two benchmarks.

4.4 Ablation Study

We perform ablation experiments to analyze the contribution of occlusion-aware image inpainting, reported in Table 3. The ablation uses paired identical samples: 59 targets from HomebrewedDB and 114 targets from LM-O. Within each dataset, samples are fully matched between "enabled" and "disabled" settings. We bold the better result within each dataset pair.



Figure 2: Qualitative comparison of 3D reconstruction and pose estimation results against baseline methods (RecGen, Any6D) on single-object samples from HomebrewedDB and LM-O. Rows from top to bottom: input RGB image, predictions of our PhysRecon, RecGen, Any6D, and ground-truth mesh. Our method recovers more accurate object geometry and correct object pose, especially for textured boxes and curved objects. The red box highlights a challenging case: competing baselines produce distorted mesh geometry and wrong orientation for the raisin box, while PhysRecon produces a reconstruction closely aligned with the ground truth. Compared with baselines, our physically-constrained pipeline yields meshes with better geometric fidelity and physically plausible poses.

Table 3: Ablation on occlusion-aware inpainting. Best per pair bolded.

Dataset	Inpaint	CD_n(↓)	A-SB(↓)	A@0.1(↑)	A@0.05(↑)	DRE(↑)
HomebrewedDB	Enabled	**0.031**	**0.045**	**93.22%**	67.80%	**42.37%**
HomebrewedDB	Disabled	0.035	0.046	91.53%	**69.49%**	40.68%
LM-O	Enabled	**0.041**	**0.052**	93.86%	63.16%	**42.98%**
LM-O	Disabled	0.041	0.054	93.86%	63.16%	42.11%

$CD_n = CD_{norm}$, A-SB=ADD-SB, A@τ=ADD-SB@τ, DRE=DRE@0.05

On HomebrewedDB, enabling image inpainting improves four out of five metrics: CD_{norm} reduces by 13.1%, ADD-SB reduces by 2.0%, ADD-SB@0.1 and DRE@0.05 increase by 1.69 percentage points. The only degradation appears in ADD-SB@0.05 (-1.69 pp). Under paired sign test, only CD_{norm} on HomebrewedDB is statistically significant ($p = 0.004$, 95% CI $[+0.0000, +0.0082]$, sign test $41 : 18$). For LM-O, the performance gap is much smaller. ADD-SB@0.1 and ADD-SB@0.05 remain unchanged. The DRE@0.05 metric gains 0.88 percentage points with inpainting enabled. All confidence intervals on LM-O cross zero, showing no statistically significant improvement. We conclude that occlusion-aware inpainting yields moderate but not universally consistent gains in numerical reconstruction metrics.

We also conduct a qualitative ablation to validate the contribution of our occlusion-aware inpainting module, as visualized in Figure 3. When removing the inpainting branch, the reconstruction



Figure 3: Qualitative ablation study on occlusion-aware image inpainting module. Rows from top to bottom: input RGB image, reconstruction results of full PhysRecon (with inpainting), PhysRecon without inpainting module, and ground-truth mesh. Without the occlusion-aware inpainting, the reconstructed meshes exhibit severe geometric artifacts and broken surfaces in occluded regions, especially for the curved cherry can and butter carton. By incorporating inpainting to recover the visual information hidden by occlusion, our full pipeline produces complete, clean object meshes that closely match ground-truth geometry. This demonstrates that the inpainting component effectively mitigates reconstruction degradation caused by heavy occlusion.

pipeline cannot observe the occluded object regions from the input RGB image. This leads to incomplete, broken meshes and distorted surfaces, most notably on the cherry can and butter carton where occlusion hides large portions of object geometry. With occlusion-aware inpainting enabled, the missing visual information is hallucinated reasonably, allowing our reconstruction backbone to recover complete and intact object meshes. The qualitative comparison confirms that the inpainting module is critical to repairing mesh artifacts under heavy occlusion and preserves high-fidelity geometry consistent with the ground truth.

4.5 Qualitative Results

We provide qualitative visualizations to demonstrate two core strengths of PhysRecon: improved mesh reconstruction quality and physical plausibility after simulation-based refinement.

First, we compare our PhysRecon against two strong baselines, RecGen and Any6D, on reconstruction and pose estimation tasks, as shown in Figure 2. Compared with the competitors, our method reconstructs more faithful object geometry and more accurate 6D poses. Across different categories including rigid toys, curved bottles and textured cartons, our method consistently reduces geometric distortion and pose misalignment.

Second, we visualize the effect of our physics validation and refinement loop. While our main numerical metrics measure geometric alignment against ground truth mesh and pose, these metrics do not capture physical feasibility. Our physics-aware sampling corrects physically invalid configurations: floating objects, unstable stacked objects, and inter-object penetration. The quasi-static rollout inside PhysRecon adjusts object scale and pose so objects rest stably on planar supporting surfaces, eliminating impossible geometries that pass visual mask checks but violate gravity and contact constraints. These physically corrected cases are visualized in Figure 4.

Figure 4 illustrates the progressive behavior of our pipeline across three example multi-object scenes. Raw reconstructed meshes from the perception backbone produce object geometries that

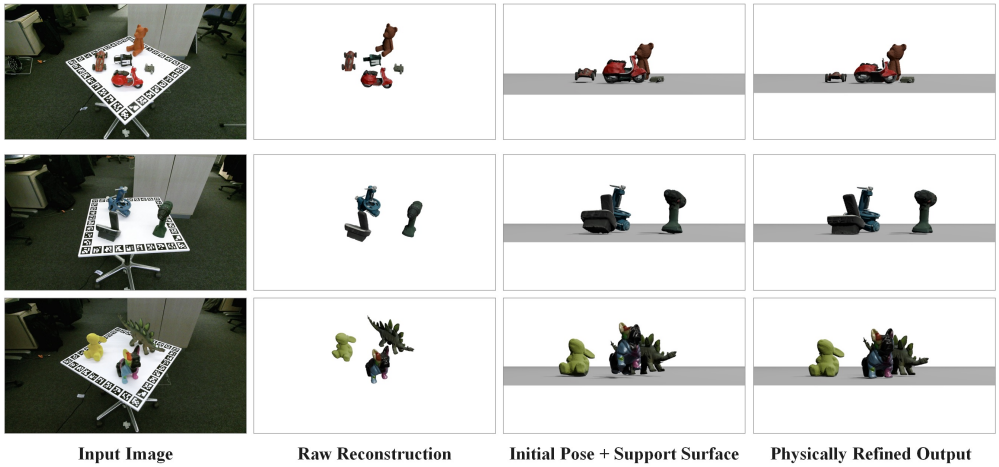


Figure 4: Qualitative comparison showing the effect of our physics refinement pipeline. From left to right: (1) Input RGB image of multi-object scenes; (2) Raw initial 3D reconstruction output, where objects float in space without contact constraints; (3) Initial pose estimates with estimated supporting surface, where objects are placed onto the surface but remain physically unstable; (4) Our final result after physics-based pose-scale sampling and quasi-static rollout. The physics module resolves floating artifacts and unstable configurations, adjusting object pose and scale so all objects rest stably on the support surface while preserving visual consistency with the input image.

float freely in 3D space with no contact information. After adding the estimated support surface, objects are projected onto the surface, yet many configurations are still physically infeasible (unbalanced tilting, unstable stacking). Our physics-aware pose-scale sampling and quasi-static rollout further adjust each object's scale and pose to eliminate floating and penetration artifacts. The final configuration satisfies gravity and contact constraints while maintaining alignment with the visual observation from the input RGB image.

5 Conclusion

In this work, we presented PhysRecon, a physically informed single-image 3D reconstruction framework. Unlike existing approaches that embed physical constraints directly into neural networks, our method adopts a perception-then-physics-validation paradigm. We first employ an occlusion-aware inpainting module to recover complete object geometry from a single RGB input. Then we perform pose and scale sampling within an inner loop, and use an external physics simulator to conduct quasi-static rollout and physical violation audit. Instead of directly regressing mass values, our framework estimates object density and computes mass from reconstructed mesh volume. Furthermore, our design removes the restrictive planar contact assumption, supporting curved contact surfaces and jointly estimating physical properties for both objects and their contact interfaces.

We evaluate PhysRecon on standard benchmarks, comparing against strong baselines including RecGen and Any6D. Quantitative metrics demonstrate that our method maintains competitive geometric reconstruction and pose estimation accuracy, while substantially improving the physical plausibility of reconstructed scenes. Ablation studies validate the effectiveness of our occlusion-aware inpainting module and the physics-based pose-scale optimization pipeline. Supplementary qualitative results in the appendix further verify the generalization ability of our method across diverse object categories, including complex concave, hollow, and mechanical objects.

Our work also has several limitations. The current framework assumes rigid objects and does not handle deformable bodies. In addition, the physical simulation validation introduces extra computational overhead compared with pure feed-forward reconstruction models. Future work may extend PhysRecon to handle deformable objects, reduce simulation overhead, and extend the physical audit to more complex multi-object dynamic interactions.

AI USE STATEMENT

This work leverages artificial intelligence tools **only for grammatical correction, wording polishing, and fluent expression optimization** of the manuscript text. No AI model or tool is involved in the core research process, including but not limited to research idea design, theoretical derivation, algorithm design, experimental scheme formulation, data collection and processing, model training and testing, result analysis, and conclusion summarization. All intellectual contributions, core innovations, experimental designs, and analytical conclusions in this paper are independently completed by the authors. The use of AI tools does not affect the originality, integrity, and academic rigor of the research work.

ETHICS STATEMENT

This study strictly abides by academic ethics and ICLR ethical guidelines throughout the entire research and manuscript preparation process. There is no data forgery, tampering, plagiarism, duplicate submission, or improper academic behavior in this work. All referenced literatures, methods, and experimental materials are clearly cited and attributed in the manuscript. The research does not involve human subjects, sensitive personal data, animal experiments, or any potentially harmful, discriminatory, or risky content that violates social ethics and academic norms. All authors have confirmed the rationality and compliance of the research content, and undertake full responsibility for the academic integrity and ethical compliance of this work.

REPRODUCIBILITY STATEMENT

All experiments, analyses, and conclusions presented in this paper are **fully reproducible**. The complete experimental settings, including model architecture details, hyperparameter configurations, training strategies, evaluation metrics, and experimental environment (software version, hardware configuration, operating system) are comprehensively and clearly described in the manuscript. There are no hidden experimental parameters, unstated operations, or unaccountable experimental processes. Researchers can completely replicate all experimental results and verify the corresponding conclusions of this paper by following the detailed steps provided in the manuscript. We ensure the stability and reliability of the experimental results, and all core findings are robust and repeatable under standard experimental conditions.

REFERENCES

Jiayang Ao, Qiuhong Ke, and Krista A. Ehinger. Image amodal completion: A survey. *Computer Vision and Image Understanding*, 229:103661, 2023.

Sherwin Bahmani, Tianchang Shen, Jiawei Ren, Jiahui Huang, Yifeng Jiang, Haithem Turki, Andrea Tagliasacchi, David B. Lindell, Zan Gojcic, Sanja Fidler, Huan Ling, Jun Gao, and Xuanchi Ren. Lyra: Generative 3d scene reconstruction via video diffusion model self-distillation. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=tIVCfVnIHo.

Eric Brachmann, Alexander Krull, Frank Michel, Stefan Gumhold, Jamie Shotton, and Carsten Rother. Learning 6d object pose estimation using 3d object coordinates. In *European Conference on Computer Vision (ECCV)*, pp. 536–551, 2014.

Vinayak Gupta, Chih-Hao Lin, Shenlong Wang, Anand Bhattad, and Jia-Bin Huang. Generalizable sparse-view 3d reconstruction from unconstrained images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 33217–33226, 2026.

Yumeng He, Yichen Song, Xiaotian Yang, Weijia Zhang, Zanwei Zhou, Junru Gong, Xiaokang Yang, and Yunbo Wang. Neoworld-pro: Programming interactive scenes from monocular images for embodied simulation. *arXiv preprint arXiv:2608.24212*, 2026.

Tomas Hodan, Frank Michel, Eric Brachmann, Wadim Kehl, Anders Glent Buch, Dirk Kraft, Bertram Drost, Joel Vidal, Stephan Ihrke, Xenophon Zabulis, Caner Sahin, Fabian Manhardt, Federico Tombari, Tae-Kyun Kim, Jiri Matas, and Carsten Rother. Bop: Benchmark for 6d object pose estimation. In *European Conference on Computer Vision (ECCV)*, pp. 19–34, 2018.

Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. In *International Conference on Learning Representations (ICLR)*, 2024.

Roman Kaskman, Sergey Zakharov, Ivan Shugurov, and Slobodan Ilic. Homebreweddb: Rgb-d dataset for 6d pose estimation of 3d objects. In *IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 2767–2776. IEEE, 2019.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4015–4026, 2023.

Hyunmin Lee and Jaesik Park. Instance-wise occlusion and depth orders in natural scenes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Taeyeop Lee, Bowen Wen, Minjun Kang, Gyuree Kang, In So Kweon, and Kuk-Jin Yoon. Any6d: Model-free 6d pose estimation of novel objects. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Dong-Yang Li, Wang Zhao, Yuxin Chen, Wenbo Hu, Meng-Hao Guo, Fang-Lue Zhang, Ying Shan, and Shi-Min Hu. Pixal3d: Pixel-aligned 3d generation from images. In *ACM SIGGRAPH Conference Papers*, 2026.

Jiahao Li, Hao Tan, Kai Zhang, Zexiang Xu, Fujun Luan, Yinghao Xu, Yicong Hong, Kalyan Sunkavalli, Greg Shakhnarovich, and Sai Bi. Instant3d: Fast text-to-3d with sparse-view generation and large reconstruction model. In *International Conference on Learning Representations (ICLR)*, 2024.

Lu Ling, Chen-Hsuan Lin, Tsung-Yi Lin, Yifan Ding, Yu Zeng, Yichen Sheng, Yunhao Ge, Ming-Yu Liu, Aniket Bera, and Max Li. Scenethesis: A language and vision agentic framework for 3d scene generation. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=SzhezVoaNB.

Xiaoxuan Ma, Jiashun Wang, Nicolas Ugrinovic, Yehonathan Litman, and Kris Kitani. Rest3d: Reconstructing physically stable 3d scenes from a single image. In *SIGGRAPH Asia*, 2026.

Van Nguyen Nguyen, Thibault Groueix, Georgy Ponimatkin, Vincent Lepetit, and Tomas Hodan. Cnos: A strong baseline for cad-based novel object segmentation. In *IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 2134–2140, 2023.

Van Nguyen Nguyen, Thibault Groueix, Mathieu Salzmann, and Vincent Lepetit. Gigapose: Fast and robust novel object pose estimation via one correspondence. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9903–9913, 2024.

Xavier Puig, Eric Undersander, Andrew Szot, Mikael Dallaire Cote, Tsung-Yen Yang, Ruslan Partsey, Ruta Desai, Alexander Clegg, Michal Hlavac, So Yeon Min, Vladimír Vondruš, Théophile Gervet, Vincent-Pierre Berges, John M. Turner, Oleksandr Maksymets, Zsolt Kira, Mrinal Kalakrishnan, Jitendra Malik, Devendra Singh Chaplot, Unnat Jain, Dhruv Batra, Akshara Rai, and Roozbeh Mottaghi. Habitat 3.0: A co-habitat for humans, avatars, and robots. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=4znwzG92CE.

Bokui Shen, Fei Xia, Chengshu Li, Roberto Martín-Martín, Linxi Fan, Guanzhi Wang, Claudia Pérez-D'Arpino, Shyamal Buch, Sanjana Srivastava, Lyne P. Tchapmi, Micael E. Tchapmi, Kent Vainio, Josiah Wong, Li Fei-Fei, and Silvio Savarese. igibson 1.0: A simulation environment for interactive tasks in large realistic scenes. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 7520–7527, 2021.

Wei-En Tai, Yu-Lin Shih, Cheng Sun, Yu-Chiang Frank Wang, and Hwann-Tzong Chen. Segment anything, even occluded. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29385–29394, 2025.

Pengfei Wang, Liyi Chen, Zhiyuan Ma, Yanjun Guo, Guowen Zhang, and Lei Zhang. One2scene: Geometric consistent explorable 3d scene generation from a single image. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=iEelSbUsSy.

Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. Foundationpose: Unified 6d pose estimation and tracking of novel objects. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17868–17879, 2024.

Jiale Xu, Weihao Cheng, Yiming Gao, Xintao Wang, Shenghua Gao, and Ying Shan. Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. *arXiv preprint arXiv:2404.07191*, 2024.

Andrii Zadaianchuk, Leonardo Barcellona, Lennard Schuenemann, Christian Gumbsch, Zehao Wang, Muhammad Zubair Irshad, Fabien Despinoy, Rahaf Aljundi, Stratis Gavves, and Sergey Zakharov. Reconstruction by generation: 3d multi-object scene reconstruction from sparse observations. *arXiv preprint arXiv:2604.27106*, 2026.

Chaoning Zhang, Dongshen Han, Yu Qiao, Jung Uk Kim, Sung-Ho Bae, Seungkyu Lee, and Choong Seon Hong. Faster segment anything: Towards lightweight sam for mobile applications. *arXiv preprint arXiv:2306.14289*, 2023.

A Appendix

A.1 Additional Multi-view Qualitative Results

Figure 5 provides supplementary multi-view qualitative reconstruction results on more test samples. The first and fourth rows show the original input RGB images with masked target objects. For each input scene, we render the reconstructed mesh from two separate camera viewpoints to inspect the geometric consistency of the recovered 3D model. Our PhysRecon is capable of reconstructing complete meshes for a variety of rigid objects. The reconstructed geometry remains consistent across different viewing angles, without severe mesh distortion or pose drift. These examples further verify that our pipeline can reliably recover object shape and 6D pose for both curved and box-shaped objects.

A.2 Extra Ablation Visualizations for Occlusion-aware Inpainting

Figure 6 presents additional qualitative ablation examples to further verify the effect of our occlusion-aware inpainting module. The samples here focus on challenging concave and hollow objects, which are difficult to reconstruct when parts of the surface are occluded in the input view. Without inpainting, occluded regions cannot be properly inferred, resulting in broken meshes, missing geometry and non-watertight surfaces on hollow ring-like objects. By recovering plausible visual information for occluded areas, the inpainting branch enables our full PhysRecon to reconstruct complete and continuous mesh surfaces, matching the ground-truth shape. Together with the main-text ablation visualization, these cases provide more evidence that the inpainting module improves reconstruction quality for complex geometries under occlusion.

A.3 Additional Qualitative Comparisons against Baselines

Figure 7 presents extra qualitative examples comparing PhysRecon with RecGen and Any6D over a diverse set of object instances. The samples cover challenging categories including mechanical assemblies, hand power tools, telecommunication devices and non-rigid-looking toy sculptures. RecGen and Any6D tend to produce distorted mesh geometry and misaligned poses on objects with complex topological structures. Benefiting from occlusion-aware inpainting and physics-based pose-scale sampling, our PhysRecon reconstructs object shapes and orientations that are much closer to the ground truth across all these test cases. Together with the main-text baseline comparison figure, these visual results provide more comprehensive evidence of the superiority of our method.



Figure 5: Additional multi-view qualitative results of PhysRecon, presented in the appendix. Rows 1 and 4: input RGB images with overlaid object masks, denoting the target objects for 3D reconstruction. Rows 2 and 3 (corresponding to row 1 scenes): two distinct viewpoints of the reconstructed 3D meshes from our PhysRecon. Rows 5 and 6 (corresponding to row 4 scenes): two additional viewpoints of reconstructed meshes for the second batch of test scenes. These supplementary visualizations demonstrate that our method can recover high-fidelity object geometry and correct object pose across diverse object categories, including toys, tools and daily containers. The reconstructed meshes maintain consistent shape and texture under different viewing angles, validating the stability of our reconstruction pipeline.



Figure 6: Additional qualitative ablation results for the occlusion-aware inpainting module (Appendix). Rows from top to bottom: input RGB images, reconstructions from full PhysRecon with occlusion-aware inpainting, reconstructions from PhysRecon without the inpainting repair branch, and ground-truth meshes. For samples with heavily occluded concave and hollow geometries, removing the inpainting module leads to severe mesh fragmentation, holes and incomplete surfaces, especially visible on the hollow ring-shaped objects. With our inpainting component, the pipeline can recover complete closed surfaces for these challenging concave objects, yielding geometry much closer to ground truth. These supplementary examples further validate the necessity of occlusion-aware inpainting for reconstructing complex shapes under occlusion.



Figure 7: Additional qualitative comparisons against baseline methods (Appendix). Rows from top to bottom: input RGB images, reconstructions from our PhysRecon, RecGen, Any6D, and ground-truth meshes. This figure includes a broader range of object categories such as mechanical parts, hand tools, electronics and irregular toys. Compared with RecGen and Any6D, our PhysRecon produces more faithful geometry and accurate 6D poses for objects with complex topology. Baselines frequently suffer from distorted shapes and incorrect orientation for objects with intricate structures. These supplementary examples further demonstrate the consistent advantage of our physically constrained reconstruction pipeline over competing approaches.