TabletopGen: Tabletop Scene Generation and Interactive Simulation for Robotic Manipulation

Ziqian Wang[1,2], Yonghao He[3], Licheng Yang[1,2], Wei Zou[1,2], Hongxuan Ma[2], Liu Liu[4], Wei Sui[3], Yuxin Guo[1,2], and Hu Su[2⋆]

[1] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[2] State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS), Institute of Automation, Chinese Academy of Sciences, Beijing, China
[3] D-Robotics, Beijing, China
[4] Horizon Robotics, Beijing, China
{wangziqian2024,hu.su}@ia.ac.cn
https://d-robotics-ai-lab.github.io/TabletopGen.project/

Abstract. Simulation provides a low-cost, scalable pathway to large-scale robotic manipulation data collection. However, existing 3D scene generation methods can rarely be applied directly to manipulation data synthesis, as their generated scenes often lack instance-level interactivity and physical plausibility. Focusing on tabletop manipulation, we propose **TabletopGen**, a training-free and automated tabletop scene generation and interactive simulation engine. Starting from text or a single image, we first obtain independent 3D object models via generative instance extraction. Second, we introduce a novel pose and scale alignment approach that recovers a collision-free scene layout using a Differentiable Rotation Optimizer and a Top-View Spatial Alignment mechanism. Finally, we assemble the generated scene in a physics simulator with collision geometry, yielding a stable, interactable environment for synthesizing multimodal manipulation data. Extensive experiments and user studies demonstrate that TabletopGen achieves state-of-the-art performance in visual fidelity, layout accuracy, and physical plausibility. Furthermore, we validate the executability of the collected trajectories on a real robotic arm via zero-shot real-to-sim-to-real policy transfer, indicating that TabletopGen can serve as a reliable data engine for robotic manipulation data synthesis.

Keywords: 3D Scene Generation · Tabletop Environments · Interactive Simulation · Manipulation Data Synthesis

1 Introduction

The tabletop environment is the "last meter" of robotic manipulation: it serves as the fundamental platform for most fine-grained interaction tasks such as

⋆ Corresponding author.

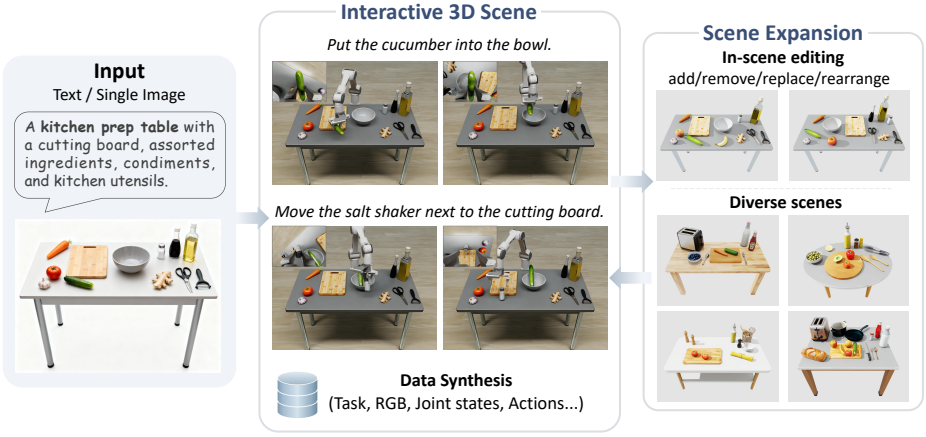

Fig. 1: We present **TabletopGen**, a training-free, fully automatic framework for generating tabletop simulation scenes. Given a text description or a single image, TabletopGen produces high-fidelity, detail-rich, instance-level interactive tabletop environments that can be directly used for robotic manipulation tasks. It supports multimodal data synthesis within the simulation and enables scene expansion—either via in-scene editing or by generating diverse scenes under the same description—thereby providing a richer coverage of simulated environments for large-scale data synthesis. We showcase more tabletop scenes of various types and styles in the supplementary.

grasping, placing, and assembly, and acts as a touchstone for evaluating robots' capabilities in complex manipulation [24, 45, 57]. With the rapid development of Embodied AI and Vision-Language-Action (VLA) models, learning manipulation policies increasingly relies on massive, highly generalizable interaction data [4, 6, 11, 28]. However, collecting data in real-world tabletop settings is often costly, inefficient, and difficult to scale up for broad coverage and generalization. In contrast, synthesizing data in simulation offers a highly promising route to obtain diverse interaction demonstrations in a low-cost and controllable manner [8, 47, 55]. Therefore, automatically generating high-fidelity, instance-interactive, physically plausible 3D tabletop scenes—and efficiently synthesizing manipulation data therein for planning and learning—is a critical step toward improving the generalization of robotic policies.

However, existing 3D scene generation methods often struggle to directly produce simulation-ready tabletop environments. Retrieval-based [7, 12, 24, 52] approaches acquire instance models from fixed asset libraries [13, 14, 16, 31], often causing insufficient diversity and geometric mismatch. Large language model (LLM)-based text-driven methods [2, 7, 17, 20, 43, 52] mostly focus on room-scale layouts, failing to handle the high-density, small-object functional arrangements and strong semantic constraints inherent to tabletop scenes. While single-image reconstruction methods [3, 10, 26, 32, 35] can generate visually similar scenes, they are limited by incomplete observation from a single viewpoint and the lack of physical priors (*e.g.*, treating the tabletop as a Z-support and XY-boundary).

As a result, they often produce instance meshes with fused or broken topology, as well as physically implausible layouts such as interpenetration and floating objects. Consequently, these scenes cannot be reliably imported into simulators for motion planning, as low-quality instances and physically invalid layouts compromise collision modeling and reachability, making it difficult to automatically generate large-scale, executable manipulation data.

To break this data bottleneck, we propose **TabletopGen**, a tabletop environment generation engine. As illustrated in Fig. 1, TabletopGen takes text or a single image as input and automatically produces diverse, instance-level, physically plausible tabletop scenes that can be directly used for motion planning and data synthesis. Our framework decouples the complex scene construction problem into three progressive stages. First, in the generative instance extraction stage, we extract and complete high-quality 2D instances from text-to-image synthesized or real-world images, and reconstruct them into 3D assets under a unified canonical coordinate system, thereby removing the reliance on fixed libraries and ensuring geometric completeness of instances. Next, to accurately estimate the rotation, translation, and scale of instances, we propose **a novel two-phase pose and scale alignment approach**: we use a Differentiable Rotation Optimizer (DRO) to estimate rotations, and introduce a Top-View Spatial Alignment (TSA) mechanism to provide physical priors and infer globally consistent translations and scales, yielding physically plausible layouts. Finally, we assemble the instances in a simulator and attach physical properties, thereby producing realistic and interactive 3D tabletop scenes that conform to real-world logic. Moreover, TabletopGen supports fast scene derivation by adding, removing, replacing, and rearranging objects, and can generate diverse scenes under the same text description. This enables low-cost scene scaling and broader simulation environment coverage for large-scale data synthesis.

Thanks to instance independence and physically plausible layouts, the simulation scenes generated by TabletopGen can directly support data synthesis for manipulation policy learning. To validate the real-world utility of the synthesized demonstrations, we further conduct a zero-shot real-to-sim-to-real policy transfer experiment. Given a single photograph of a real tabletop, our framework reconstructs an instance-level simulated scene, synthesizes manipulation trajectories with domain randomization, and uses them to fine-tune a robot policy. The resulting policy is then directly deployed on a real robotic arm without any real-world data fine-tuning. Successful real-world executions across multiple tasks indicate that the generated assets, spatial layouts, and physical properties are sufficiently accurate to support downstream sim-to-real policy learning, validating TabletopGen as a practical tabletop manipulation data engine.

In summary, our main contributions are as follows:

- We present TabletopGen, an automated, training-free 3D tabletop scene generation framework. Starting from text or a single image, it generates simulation scenes that jointly achieve visual realism, instance-level interactivity, and physical plausibility, providing an efficient data synthesis route for policy learning in robotic manipulation tasks.

- We propose a novel pose and scale alignment method tailored for tabletop scenes. Via the decoupled DRO and TSA phases, we recover visually consistent and physically plausible tabletop layouts from the 2D reference.
- We conduct comprehensive system-level validation. Extensive experiments show that TabletopGen significantly outperforms existing methods in scene quality and physical validity. Furthermore, our zero-shot real-to-sim-to-real policy transfer experiments confirm the reliability of the generated scenes and synthetic trajectories for real-world robotic manipulation policy learning.

2 Related Works

Large-scale, high-quality robot data is crucial for embodied AI. For example, RT-1 [5] used approximately 130k trajectories collected by 13 robots over 17 months, and Octo [46] was trained on around 800k trajectories. The data for these studies are generally collected manually, which is costly, inefficient, and difficult to scale and generalize. With the advancement of physics simulation engines such as Isaac Sim [36] and SAPIEN [50], efficiently synthesizing data in simulation offers a promising solution to this challenge [58]. Meeting the demand for high-quality data generation requires building simulation scenes that are diverse and highly realistic. Related research primarily follows two directions: digital-twin reconstruction from real observations, and generative 3D scene synthesis.

2.1 Digital Twin and Real-to-Sim-to-Real

These methods [23, 33, 48, 49] reconstruct digital-twin scenes from real-world videos or image sequences, and collect manipulation demonstrations or optimize policies within them. By providing spatial geometry consistent with the real world, they can support high-precision Real-to-Sim-to-Real closed-loop deployment. However, their scene construction often relies on multi-view scanning or manual intervention. This strong dependence on real-world observations limits the diversity of simulated scenes, making it difficult to satisfy the need for massive scenes in embodied learning.

2.2 Generative 3D Scene Synthesis

To overcome the diversity bottleneck of real-scene inputs, an increasing number of works have begun exploring automatic 3D scene generation from text or single image, attempting to serve them as simulation environments for embodied AI.

Text-driven methods use LLMs for semantic and spatial reasoning to either directly generate 3D layouts [15,37,51], or first produce scene graphs or spatial constraints and then optimize layouts via solvers [2, 7, 17, 19, 32, 34, 43, 52]. These methods are typically designed for room-scale scenes and can support tasks like navigation and planning. However, they struggle to capture the high-density, semantics-constrained functional layouts of small objects that are unique to tabletops, and thus are not well-suited for fine-grained robotic manipulation.

Moreover, most methods rely on retrieving 3D assets from fixed libraries, which limits style consistency and asset flexibility. Recently, MesaTask [24] proposed a text-driven scene generation method targeting tabletop tasks, but it also relies on an asset library and places objects only on a rectangular plane, without modeling the full 3D structure of the table.

Image-driven methods aim to recover visually consistent 3D scenes from a single image. Retrieval-based approaches [12,18,21,60] are constrained by limited asset libraries and thus cannot precisely reconstruct target objects. Diffusion-based holistic generation methods [9, 26, 54] can end-to-end generate scenes consistent with the input view. Yet, due to incomplete single-view observations, they often produce fused or broken instance meshes. In addition, lacking dedicated tabletop datasets and physical priors, their generated scenes frequently exhibit physically implausible layouts such as interpenetration, floating objects, or objects falling outside the tabletop boundary. Compositional methods [3, 20, 22, 25, 35, 53] improve flexibility by per-instance generation and alignment; however, they typically rely on monocular depth back-projection for pose estimation. In dense tabletop layouts of small objects, camera calibration or depth estimation errors can be amplified and accumulated in scale and pose computation, and these methods also lack physical constraints during scene assembly, hindering globally consistent and physically feasible tabletop layouts.

Unlike the above methods, TabletopGen is a simulation data engine specifically designed for tabletop robotic manipulation. It ensures high-fidelity, instance-level 3D reconstruction from a single image, while leveraging the diversity of text-to-image generation to expand scenes infinitely. More importantly, we propose a novel decoupled pose and scale alignment approach: DRO precisely optimizes instance rotations using visual features, and TSA introduces physical priors via top-view and stacking relations to constrain scene boundaries and support relations. As a result, TabletopGen generates globally consistent and physically plausible tabletop simulation environments that can be directly used for manipulation data synthesis and zero-shot real-to-sim-to-real transfer.

3 Method

Given a text description T or an image I, our goal is to generate a tabletop scene $S = \{o_i\}_{i=1}^{n}$ that is interactive in a physics simulator. Each instance o_i is defined by its 3D model m_i, scale $s_i \in \mathbb{R}^3$, translation $t_i \in \mathbb{R}^3$, and rotation $r_i \in \mathbb{R}$ around the vertical axis. As shown in Fig. 2, we adopt a multi-stage architecture: we first generate canonicalized instance assets, then obtain a globally consistent layout via a decoupled pose and scale alignment approach, and finally assemble the scene in a simulator to support robotic manipulation data synthesis. Following prior works [20, 59], we argue that images provide more realistic and precise layout information than text. Thus, for a text input T, an LLM expands it into a detailed prompt that specifies the scene style, object set, and their functional layout, and a text-to-image model synthesizes a visually realistic and plausibly arranged reference image I_{ref}; image inputs serve directly as I_{ref}.

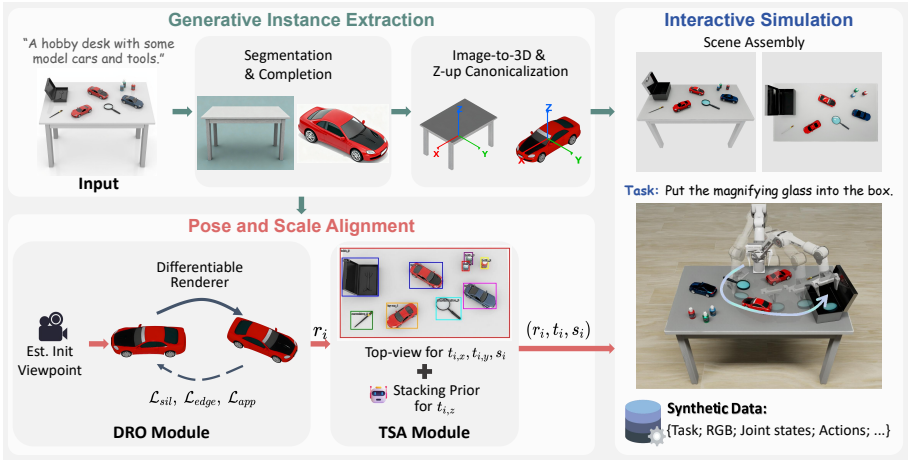

Fig. 2: Overview of our TabletopGen Framework. Given a text prompt (first converted into a reference image via Text-to-Image) or a single image, we proceed in three stages: (1) **Generative Instance Extraction** segments and completes 2D instances, and reconstructs Z-up canonicalized, directly placeable 3D assets. (2)**Pose and Scale Alignment** recovers a physically feasible layout: The **DRO (Differentiable Rotation Optimizer)** optimizes rotation with a tri-modal loss, while the **TSA (Top-view Spatial Alignment)** infers translation and scale using a synthesized top-view and stacking priors, together with an anchor object selected by RMA-Score. (3) **Interactive Simulation** assembles the scene in a physics simulator, performs motion planning for manipulation tasks and executes collision-free trajectories, thereby synthesizing multimodal interaction data.

3.1 Generative Instance Extraction

To enable instance-level interaction in simulation, we first extract and reconstruct geometrically complete and coordinate-canonical 3D assets from I_{ref}. We begin by using a multimodal large language model (MLLM) for open-vocabulary category recognition over tabletop scene objects (including the table itself), and combine it with GroundedSAM-v2 [41] to obtain instance masks. Because severe single-view occlusions often yield masks with holes or blurred boundaries, we discard traditional inpainting and instead utilize a multimodal generative model to complete and redraw each instance mask. The completed instance images are used to generate high-quality 3D meshes m_i' via Image-to-3D. However, the local coordinate system of m_i' is typically arbitrary, and directly using it can lead to unplaceable states such as being tilted or inverted. Therefore, guided by the visual cues from I_{ref} and the semantic priors from the MLLM, we apply rotation correction to align the local vertical axis of each m_i' to the world Z-up direction, yielding canonical assets m_i that can be directly placed in tabletop scenes.

3.2 Pose and Scale Alignment

The second and most critical stage is to estimate each instance's **rotation** r_i, **translation** t_i, and **scale** s_i, so that all canonical 3D models m_i can be assembled into a physically plausible scene consistent with I_{ref}. We propose a novel approach that decouples this problem into two phases: an instance's rotation around the vertical axis typically does not change the overall layout but can substantially alter its silhouette and texture distribution, impacting the visual consistency with the input. Therefore, our DRO phase precisely optimizes the rotation based on visual information. In contrast, translation and scale directly affect the spatial relations. Our TSA phase utilizes a top-view image to compensate for single-view observation and, combined with the constraints of physical priors, uniformly infers globally consistent translations and scales, thereby obtaining an executable physical layout.

Differentiable Rotation Optimizer (DRO). In this phase, we estimate r_i for each instance. Under the initial camera viewpoint (azimuth, elevation) estimated by MLLM, we render each m_i using a differentiable renderer and obtain a textured image $I_{render}(r_i)$, a soft silhouette $\hat{S}(r_i)$, and a soft edge map $\hat{E}(r_i)$ derived from $\hat{S}(r_i)$ using a Sobel filter. We then define a **tri-modal matching loss** $\mathcal{L}_{rot}$ to align the render with the target instance $I_{instance}$ and its mask S:

$$\mathcal{L}_{rot}(r_i) = \lambda_s \mathcal{L}_{sil}(r_i) + \lambda_e \mathcal{L}_{edge}(r_i) + \lambda_a \mathcal{L}_{app}(r_i) \tag{1}$$

where λ_s, λ_e and λ_a are weighting coefficients. The first term, $\mathcal{L}_{sil}$, is a soft IoU loss for shape consistency:

$$\mathcal{L}_{sil}(r_i) = 1 - \frac{\sum(\hat{S}(r_i) \cdot S)}{\sum(S + \hat{S}(r_i) - S \cdot \hat{S}(r_i))} \tag{2}$$

The second, $\mathcal{L}_{edge}(r_i)$, is a one-sided Chamfer loss that matches contours. It penalizes the distance from $\hat{E}(r_i)$ to the nearest ground-truth edge. A distance transform D_S is pre-computed from Canny edges of the target mask S. The loss is the weighted average distance over all pixels x:

$$\mathcal{L}_{edge}(r_i) = \frac{\sum_x D_S(x) \cdot \hat{E}(x, r_i)}{\sum_x \hat{E}(x, r_i)} \tag{3}$$

Finally, $\mathcal{L}_{app}$ is a perceptual feature loss that ensures appearance similarity with DINOv2 [38] features Φ:

$$\mathcal{L}_{app}(r_i) = ||\Phi(I_{render}(r_i)) - \Phi(I_{instance})||_2^2 \tag{4}$$

We find the optimal rotation $\hat{r}_i$ by minimizing this loss via gradient descent: $\hat{r}_i = \arg\min_{r_i} \mathcal{L}_{rot}(r_i)$. This process can also optionally refine the camera pose.

Top-View Spatial Alignment (TSA). In this phase, we incorporate two types of physical priors to achieve a physically plausible layout estimation: (1) we employ an MLLM to infer the stacking relations among instances for placement, thereby avoiding floating and vertical interpenetration; (2) we utilize a multimodal generative model to synthesize a top-view of the scene, I_{top}, and extract the 2D bounding boxes of all instances. This top-view prior naturally ensures that all objects are constrained within the valid tabletop boundaries.

Building upon this, we estimate the translation and scale. For each instance, let r_{img} and A_{px} denote the pixel aspect ratio and pixel area of its top-view bounding box, respectively. We query the MLLM for its commonsense physical size and apply the estimated r_i to obtain its physical aspect ratio r_{phys} projected onto the xy-plane. We then propose a Ratio-Matched and Area-Weighted Score (RMA-Score) to select a reliable scaling anchor:

$$\varepsilon_{ratio} = |\log r_{phys} - \log r_{img}|, \quad \text{RMA}(i) = \frac{A_{px}(i)}{1 + (\varepsilon_{ratio}/\tau)^2} \tag{5}$$

where τ is a tolerance hyperparameter. This score favors objects with large areas (for stability) and low aspect-ratio mismatch (to ensure geometric consistency). The instance with the highest RMA-Score is selected as the anchor to compute a global scaling factor α (meters/pixel), thereby precisely inferring the 3D scale s_i and the horizontal translation coordinates $(t_{i,x}, t_{i,y})$ for all instances. The vertical translation $t_{i,z}$ is then determined based on the stacking relations, ultimately yielding a physically plausible 3D layout $\{r_i, t_i, s_i\}$.

3.3 Interactive Simulation

We import each m_i into a physics simulator, Isaac Sim, and apply the corresponding $\{r_i, t_i, s_i\}$ transforms. To support realistic rigid-body dynamic interactions, we generate collision geometries for each instance via convex decomposition and enable gravity and friction parameters, thereby turning the scene into an interactive simulated tabletop environment. Furthermore, by adding, removing, replacing, or locally rearranging instances, we can expand the original scene into a massive number of variants for large-scale data synthesis.

Within the simulation environment, we support both manually specified manipulation tasks and the automatic generation of reasonable interaction tasks using an LLM based on the instances' semantic labels and physical properties. Given a task, we employ a motion planner [44] to automatically compute collision-free trajectories. During trajectory execution by the robotic arm, we can collect rich multimodal interaction data at low cost, including multi-view RGB images, joint states, actions, and more. Through this complete closed loop—from automated scene generation to automated data collection—TabletopGen provides a continuous stream of high-quality synthetic data for training VLA models with strong generalization capabilities.

4 Experiments

4.1 Setup

Implementation details. We use ChatGPT [1] for vision-language reasoning, Seedream [42] for image synthesis, and Hunyuan3D-3.0 [30] to generate 3D meshes. In DRO, we render with PyTorch3D [40] and run a coarse search over the $[0°, 360°)$ range at $5°$ step size to select the 8 candidate angles with the lowest loss at first. Each candidate is then refined using the Adam optimizer [29] (lr $= 3 \times 10^{-2}$, 140 steps; $\lambda_s = 0.5$, $\lambda_e = 0.5$, $\lambda_a = 2.0$). In TSA, we set the RMA-Score tolerance to $\tau = 0.25$. The prompts and step-wise intermediate outputs are provided in the supplementary materials.

Baselines. Since our tabletop scenes are primarily generated from 2D reference, we compare against representative single-image 3D scene generation methods: ACDC [12], Gen3DSR [3], and MIDI [26]. To illustrate our text-to-scene capability, we also compare with MesaTask [24] and Holodeck-table, a tabletop-adapted version of Holodeck [52] following the adaptation strategy used in MesaTask.

Metrics. Following prior works [20, 22, 24, 25, 34, 35, 53], we comprehensively evaluate both visual consistency and simulation usability: (1) LPIPS [56], DINOv2 [38], and CLIP [39] for perceptual similarity, visual consistency, and semantic alignment; (2) Physical plausibility via the proportion of colliding object pairs (Col_O) and the percentage of scenes containing collisions (Col_S); (3) Multi-dimensional GPT Evaluation using GPT-4o [27] to rate Visual Fidelity (VF), Image Alignment (IA) and Physical Plausibility (PP) on a 1-7 scale, and directly provide an Overall Ranking (OR) across methods.

For text-to-scene comparisons, where no unique reference image is available, we evaluate the generated scenes with GPT-4o under the same input text prompts. Specifically, we rate Consistency with Text (CwT), Object Quality and Recognizability (OQR), Layout Coherence and Realism (LCR), and Physical Plausibility (PP) on a 1-7 scale, and further report Col_O and Col_S.

4.2 Comparisons

We construct a diverse test set containing 78 samples to evaluate the robustness and generalization capability of our method. The test set is built from both synthesized images and real-world photographs, and spans multiple table shapes—square, round, and triangular. For functional categories, we follow the taxonomy used in MesaTask [24] and recommendations from ChatGPT, covering office tables, dining tables, workbenches, and crafting tables. By comparing TabletopGen with different baseline methods under identical inputs, we demonstrate its significant superiority in generating diverse tabletop scenes.

Table 1: Quantitative comparison with image-driven methods. Our method TabletopGen achieves the best overall performance across all metrics, showing substantially superior results in Visual & Perceptual quality, GPT-4o Evaluation, and near-zero Collision Rates compared to all baselines.

Method	Visual & Perceptual			GPT Evaluation					Collision Rate(%)	
	LPIPS↓	DINOv2↑	CLIP↑	VF↑	IA↑	PP↑	Avg.↑	OR↓	Col_O↓	Col_S↓
ACDC	0.5124	0.3775	0.6696	2.38	1.90	2.57	2.28	3.55	8.23	67.95
Gen3DSR	0.4891	0.5602	0.8636	2.92	4.17	3.32	3.47	3.02	16.88	85.90
MIDI	0.4559	0.7070	0.8867	4.22	4.48	4.30	4.33	2.32	17.39	98.72
Ours	**0.4483**	**0.8383**	**0.9077**	**6.06**	**6.30**	**6.22**	**6.19**	**1.08**	**0.42**	**7.69**

Table 2: Quantitative comparison with text-driven methods. Our method achieves substantially better text consistency, object quality, layout realism, and overall scene quality, while maintaining near-zero collision rates.

Method	GPT Evaluation					Collision Rate(%)	
	CwT↑	OQR↑	LCR↑	PP↑	Avg.↑	Col_O↓	Col_S↓
Holodeck-table	3.57	4.17	3.90	4.70	4.08	**0.00**	**0.00**
MesaTask	3.63	4.23	4.07	4.57	4.12	4.95	63.33
Ours	**6.63**	**6.10**	**6.07**	**6.27**	**6.27**	0.20	6.67

Quantitative Evaluation. Tab. 1 reports that our method achieves the best performance across all metrics in the three evaluation aspects, clearly outperforming all image-driven baselines. For **Visual & Perceptual quality**, TabletopGen attains the best LPIPS, DINOv2, and CLIP scores. We attribute this to our "instance-first, then alignment" design, which first ensures high-fidelity per-instance reconstruction and subsequently achieves precise pose and scale alignment. Under the **GPT Evaluation**, TabletopGen achieves the highest average score (6.19)—a 43% improvement over the second-best MIDI and the top Overall Ranking (OR). This demonstrates that our results are perceived as more realistic and semantically plausible. The **Collision Rate** shows the most significant performance gap. As previously analyzed, existing 3D scene generation methods struggle to capture the high-density tabletop layouts with complex spatial relations. Their Col_O typically lies around 8%–17%, and the scene-level Col_S reaches 67%–98%. In contrast, TabletopGen precisely recovers layouts via its DRO and TSA stages. Our Col_O is near zero (0.42%), and Col_S remains only 7.69%. This overwhelming advantage demonstrates our method's robustness and practical value in producing collision-free, interactive tabletop scenes.

For text-to-scene generation, we evaluate 30 text-scene pairs covering diverse tabletop types. As shown in Tab. 2, TabletopGen achieves clearly higher GPT-4o scores than MesaTask and Holodeck-table, improving the average rating from 4.12/4.08 to 6.27. Holodeck-table obtains zero collisions mainly through hard constraints on 2D bounding boxes, which often leads to sparse and less realis-

tic layouts, while MesaTask still suffers from frequent scene-level collisions. In contrast, TabletopGen maintains near-zero mesh-level collisions while preserving realistic and task-consistent tabletop arrangements.

User Study. We further conducted a comprehensive user study to assess perceptual quality and human preference, involving 128 participants. Each participant was randomly assigned 8 scenes; to ensure fairness, the display order of results from different methods was fully randomized. Participants rated the generated scenes on three criteria—Visual Fidelity (VF), Image Alignment (IA), and Physical Plausibility (PP)—using a 7-point scale. As shown in Tab. 3, our method significantly outperforms all baselines on every criterion. Our average score of 5.56 was substantially higher than the second-best method (3.57). Most importantly, when asked for an Overall Preference (OP), our results were selected in 83.13% of cases, confirming that our generated scenes are perceived as more realistic and physically plausible by human evaluators.

Table 3: User study results. Our method TabletopGen attains the highest mean scores on Visual Fidelity (VF), Image Alignment (IA), and Physical Plausibility (PP), and the best overall preference (OP), being selected in 83.13% of cases.

Method	**VF↑**	**IA↑**	**PP↑**	**Avg.↑**	**OP(%)**
ACDC	3.00	2.46	2.87	2.78	2.38
Gen3DSR	2.95	3.25	3.19	3.13	2.88
MIDI	3.82	3.57	3.32	3.57	11.61
Ours	**5.62**	**5.50**	**5.56**	**5.56**	**83.13**

Qualitative Evaluation. Fig. 3 compares our method with **single-image** baselines on both synthesized images and real-world photos. Thanks to our precise per-instance generation and robust pose and scale alignment, TabletopGen produces consistent object counts and categories, finer geometry with unified style, semantically coherent layouts, and collision-free placement. In contrast, the retrieval-based method (*e.g.*, ACDC) is limited by its fixed asset library, failing to match specific object styles and shapes. Generative reconstruction methods (*e.g.*, Gen3DSR and MIDI) struggle with occlusions, causing incomplete instance generation, object interpenetration, and layout inaccuracies. This finding aligns with the high collision rates reported in Tab. 1. Overall, TabletopGen delivers more realistic, simulation-ready tabletops with diverse types.

We also demonstrate our **text-to-scene** capability by comparing with MesaTask in Fig. 4. MesaTask retrieves assets and places them on a fixed plane, limiting diversity and fidelity and occasionally causing minor collisions(*e.g.*, the remote control and the can in the bottom row). In contrast, TabletopGen generates

Fig. 3: Qualitative comparison on synthesized inputs (top) and real-world photos (bottom). Our method, TabletopGen, consistently outperforms all baselines. TabletopGen demonstrates strong adaptability across diverse tabletop types, delivering more realistic appearances, finer instance models, more coherent object counts and layouts, and collision-free placement.

the entire scene, including the stylistically consistent table, delivering more realistic visual appearance, richer instance counts, more reasonable arrangements, and collision-free layouts.

4.3 Ablations

We ablate the two key stages, **DRO** and **TSA**, of the pose and scale alignment approach by replacing them with a naive MLLM-only baseline: when a stage is removed, we replace its specialized function by directly prompting ChatGPT to estimate the parameters from the reference image. As reported in Tab. 4, removing either stage degrades visual & perceptual scores and sharply increases collisions, while removing both leads to the largest failure. Fig. 5 visually confirms these findings. The "w/o DRO" baseline fails to recover correct rotations, while the "w/o TSA" baseline produces implausible placements and scale drift. The "w/o both" compounds these errors, causing the final scene to suffer from severe collisions (*e.g.*, the books interpenetrating) and occlusions (*e.g.*, the monitor moving forward and blocking other objects). In contrast, our full model's

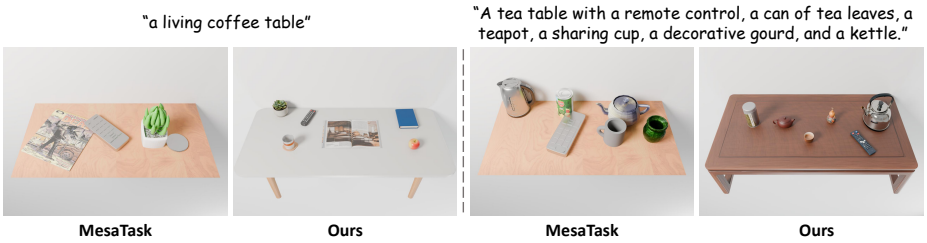

Fig. 4: Qualitative comparison under the same input texts. Compared with the recent tabletop generation method MesaTask, TabletopGen generates more complete and realistic scenes, including stylistically consistent tables, more detailed instance models, and more semantically and physically reasonable layouts.

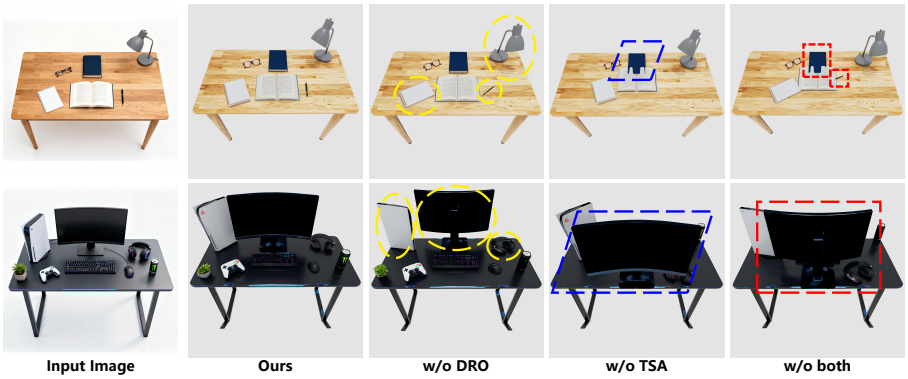

Fig. 5: Qualitative ablation study on pose and scale alignment components. Compared to our full model (Ours), removing DRO yields incorrect rotations (yellow circles), removing TSA causes misplacements (blue parallelograms), and removing both amplifies these errors, resulting in severe occlusions and collisions (red rectangles).

result (Ours) closely matches the input image and presents a collision-free layout. This demonstrates that both DRO and TSA are essential for achieving a coherent and physically plausible scene reconstruction.

4.4 Data Engine Demonstrations

Zero-shot Real-to-Sim-to-Real Policy Transfer. To further validate that TabletopGen can support downstream policy learning, we train manipulation policies using only synthetic demonstrations and directly transfer them to the real world. Starting from a real tabletop photo containing fruits, containers, and distractor objects, we generate an instance-level simulated scene and attach collision geometry and physical properties. A motion planner is then used to compute collision-free trajectories for an AgileX PiPER arm to execute pick-and-place tasks in simulation. We consider three tasks with increasing difficulty: placing a mango onto a plate, placing a smaller strawberry into a mug, and stack-

Table 4: Quantitative ablation study on pose and scale alignment components. Our full model (DRO+TSA) attains the best visual & perceptual scores and the lowest collision rates; removing either component degrades performance, and removing both leads to large increases in collisions.

Method	Visual & Perceptual			Collision Rate (%)	
	LPIPS↓	DINOv2↑	CLIP↑	Col_O↓	Col_S↓
Ours (full)	**0.4483**	**0.8383**	**0.9077**	**0.42**	**7.69**
w/o DRO	0.4523	0.8261	0.9012	1.27	16.67
w/o TSA	0.4799	0.8041	0.8954	5.50	61.54
w/o both	0.4811	0.7897	0.8922	5.41	62.82

Fig. 6: Zero-shot real-to-sim-to-real policy learning. In a simulated scene reconstructed from a real tabletop photo, we synthesize 2,000 trajectories for each of three tasks with different difficulty levels to fine-tune a $\pi_{0.5}$ policy, and directly deploy it on a physical AgileX PiPER arm without any real-world data fine-tuning. The high success rates in both simulation and the real world indicate that TabletopGen can generate high-quality and usable data for robot manipulation policy learning.

ing two small blocks on a plate. For each task, we apply domain randomization to the background lighting, arm pose, camera pose and focal length, and object poses, and synthesize 2,000 trajectories to fine-tune a $\pi_{0.5}$ policy [28], without using any real-world data for fine-tuning. We then evaluate each task for 50 trials in both simulation and the original real-world scene. As shown in Fig. 6, the resulting policies achieve high success rates in both simulated and real-world executions. The larger sim-to-real gap in Task 3 is expected, since stacking is highly sensitive to friction, contact modeling, and release timing, where small errors may cause the blocks to slip or collapse. Overall, these results show that TabletopGen can provide realistic and usable synthetic data for sim-to-real robot manipulation policy learning.

In-scene Editing for Scene Expansion. A key capability of a strong data engine is to efficiently expand data diversity. Since TabletopGen constructs scenes from independent 3D instances, the generated simulation environments are naturally editable and extensible. As shown in Fig. 7, we derive multiple scene variants by adding, removing, replacing, or rearranging objects, without re-running the full pipeline. This editing preserves the overall scene semantics while produc-

Fig. 7: In-scene Editing. Starting from the original generated scene, we demonstrate four types of in-scene editing: Add, Remove, Replace, and Rearrange. These edits can quickly derive multiple scene variants, providing broader environment coverage for subsequent task generation and interactive data synthesis.

ing diverse simulation environments, thereby broadening the coverage of downstream task generation and interaction data synthesis at a low additional cost.

5 Conclusion

In this paper, we present **TabletopGen**, a training-free, fully automatic tabletop scene generation and interactive simulation engine. It converts text or a single image into high-fidelity, instance-level interactable, and physically plausible 3D tabletop scenes, which can be utilized for multimodal data synthesis in robotic manipulation tasks. We decouple 2D-to-3D layout recovery into two stages: a Differentiable Rotation Optimizer (DRO) to accurately recover per-instance orientations, and a Top-view Spatial Alignment (TSA) module that injects tabletop-specific physical priors to infer globally consistent translations and metric scales, yielding collision-free and executable layouts. Extensive quantitative evaluations and a large-scale user study show that TabletopGen achieves state-of-the-art performance in visual quality, layout accuracy, and physical validity. Moreover, we demonstrate zero-shot real-to-sim-to-real policy transfer, showing that policies fine-tuned only on synthetic trajectories generated by TabletopGen can be directly deployed for real-world robotic manipulation. Finally, thanks to efficient in-scene editing and text-driven scene expansion, TabletopGen provides a scalable approach to deriving diverse simulation environments at low cost, offering a practical route to alleviate the data bottleneck in embodied AI.

5.1 Limitations and Future Work.

TabletopGen currently focuses on typical tabletop scenes, and future work can explore extending it to more complex support structures, such as multi-level shelves and cabinets. Moreover, given the data-engine capability demonstrated by TabletopGen, our future focus will be to use this framework to synthesize large-scale, highly diverse simulated interaction data for policy training, thereby enhancing robots' generalization ability in complex real-world manipulation tasks.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Aguina-Kang, R., Gumin, M., Han, D.H., Morris, S., Yoo, S.J., Ganeshan, A., Jones, R.K., Wei, Q.A., Fu, K., Ritchie, D.: Open-universe indoor scene generation using llm program synthesis and uncurated object databases. arXiv preprint arXiv:2403.09675 (2024)
3. Ardelean, A., Özer, M., Egger, B.: Gen3dsr: Generalizable 3d scene reconstruction via divide and conquer from a single view. In: 2025 International Conference on 3D Vision (3DV). pp. 616–626. IEEE (2025)
4. Black, K., Brown, N., Driess, D., Esmail, A., Equi, M., Finn, C., Fusai, N., Groom, L., Hausman, K., Ichter, B., et al.: π_0: A vision-language-action flow model for general robot control. arXiv preprint arXiv:2410.24164 (2024)
5. Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Dabis, J., Finn, C., Gopalakrishnan, K., Hausman, K., Herzog, A., Hsu, J., et al.: Rt-1: Robotics transformer for real-world control at scale. arXiv preprint arXiv:2212.06817 (2022)
6. Bu, Q., Cai, J., Chen, L., Cui, X., Ding, Y., Feng, S., Gao, S., He, X., Hu, X., Huang, X., et al.: Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. arXiv preprint arXiv:2503.06669 (2025)
7. Çelen, A., Han, G., Schindler, K., Van Gool, L., Armeni, I., Obukhov, A., Wang, X.: I-design: Personalized llm interior designer. In: European Conference on Computer Vision. pp. 217–234. Springer (2024)
8. Chen, T., Chen, Z., Chen, B., Cai, Z., Liu, Y., Liang, Q., Li, Z., Lin, X., Ge, Y., Gu, Z., et al.: Robotwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation. arXiv preprint arXiv:2506.18088 (2025)
9. Chen, Y., Nguyen, H.T., Voleti, V., Jampani, V., Jiang, H.: Housecrafter: Lifting floorplans to 3d scenes with 2d diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 28440–28450 (2025)
10. Chung, J., Lee, S., Nam, H., Lee, J., Lee, K.M.: Luciddreamer: Domain-free generation of 3d gaussian splatting scenes. arXiv preprint arXiv:2311.13384 (2023)
11. Collaboration, O.E., O'Neill, A., Rehman, A., Gupta, A., Maddukuri, A., Gupta, A., Padalkar, A., Lee, A., Pooley, A., Gupta, A., et al.: Open x-embodiment: Robotic learning datasets and rt-x models. arXiv preprint arXiv:2310.08864 **1**(2) (2023)
12. Dai, T., Wong, J., Jiang, Y., Wang, C., Gokmen, C., Zhang, R., Wu, J., Fei-Fei, L.: Automated creation of digital cousins for robust policy learning. In: Conference on Robot Learning (CoRL) (2024)
13. Deitke, M., Liu, R., Wallingford, M., Ngo, H., Michel, O., Kusupati, A., Fan, A., Laforte, C., Voleti, V., Gadre, S.Y., VanderBilt, E., Kembhavi, A., Vondrick, C., Gkioxari, G., Ehsani, K., Schmidt, L., Farhadi, A.: Objaverse-xl: A universe of 10m+ 3d objects. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems. vol. 36, pp. 35799–35813. Curran Associates, Inc. (2023)
14. Deitke, M., Schwenk, D., Salvador, J., Weihs, L., Michel, O., VanderBilt, E., Schmidt, L., Ehsani, K., Kembhavi, A., Farhadi, A.: Objaverse: A universe of annotated 3d objects. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13142–13153 (2023)

15. Feng, W., Zhu, W., Fu, T.J., Jampani, V., Akula, A., He, X., Basu, S., Wang, X.E., Wang, W.Y.: Layoutgpt: Compositional visual planning and generation with large language models. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems. vol. 36, pp. 18225–18250. Curran Associates, Inc. (2023)
16. Fu, H., Jia, R., Gao, L., Gong, M., Zhao, B., Maybank, S., Tao, D.: 3d-future: 3d furniture shape with texture. International Journal of Computer Vision **129**(12), 3313–3337 (2021)
17. Fu, R., Wen, Z., Liu, Z., Sridhar, S.: Anyhome: Open-vocabulary generation of structured and textured 3d homes. In: European Conference on Computer Vision. pp. 52–70. Springer (2024)
18. Gao, D., Rozenberszki, D., Leutenegger, S., Dai, A.: Diffcad: Weakly-supervised probabilistic cad model retrieval and alignment from an rgb image. ACM Transactions on Graphics (TOG) **43**(4), 1–15 (2024)
19. Gao, G., Liu, W., Chen, A., Geiger, A., Schölkopf, B.: Graphdreamer: Compositional 3d scene synthesis from scene graph s. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
20. Gu, Z., Cui, Y., Li, Z., Wei, F., Ge, Y., Gu, J., Liu, M.Y., Davis, A., Ding, Y.: Artiscene: Language-driven artistic 3d scene generation through image intermediary. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 2891–2901 (2025)
21. Gümeli, C., Dai, A., Nießner, M.: Roca: Robust cad model retrieval and alignment from a single image. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4022–4031 (2022)
22. Han, H., Yang, R., Liao, H., Xing, J., Xu, Z., Yu, X., Zha, J., Li, X., Li, W.: Reparo: Compositional 3d assets generation with differentiable 3d layout alignment. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 25367–25377 (2025)
23. Han, X., Liu, M., Chen, Y., Yu, J., Lyu, X., Tian, Y., Wang, B., Zhang, W., Pang, J.: Re^{3}sim: Generating high-fidelity simulation data via 3d-photorealistic real-to-sim for robotic manipulation. arXiv preprint arXiv:2502.08645 (2025)
24. Hao, J., Liang, N., Luo, Z., Xu, X., Zhong, W., Yi, R., Jin, Y., Lyu, Z., Zheng, F., Ma, L., et al.: Mesatask: Towards task-driven tabletop scene generation via 3d spatial reasoning. arXiv preprint arXiv:2509.22281 (2025)
25. Hu, Y., Liu, S., Yang, X., Wang, X.: Flash sculptor: Modular 3d worlds from objects. arXiv preprint arXiv:2504.06178 (2025)
26. Huang, Z., Guo, Y.C., An, X., Yang, Y., Li, Y., Zou, Z.X., Liang, D., Liu, X., Cao, Y.P., Sheng, L.: Midi: Multi-instance diffusion for single image to 3d scene generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 23646–23657 (2025)
27. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
28. Intelligence, P., Black, K., Brown, N., Darpinian, J., Dhabalia, K., Driess, D., Esmail, A., Equi, M., Finn, C., Fusai, N., et al.: $\pi_{0.5}$: a vision-language-action model with open-world generalization. arXiv preprint arXiv:2504.16054 (2025)
29. Kingma, D.P.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
30. Lai, Z., Zhao, Y., Liu, H., Zhao, Z., Lin, Q., Shi, H., Yang, X., Yang, M., Yang, S., Feng, Y., et al.: Hunyuan3d 2.5: Towards high-fidelity 3d assets generation with ultimate details. arXiv preprint arXiv:2506.16504 (2025)

31. Li, C., Zhang, R., Wong, J., Gokmen, C., Srivastava, S., Martín-Martín, R., Wang, C., Levine, G., Lingelbach, M., Sun, J., et al.: Behavior-1k: A benchmark for embodied ai with 1,000 everyday activities and realistic simulation. In: Conference on Robot Learning. pp. 80–93. PMLR (2023)
32. Li, H., Shi, H., Zhang, W., Wu, W., Liao, Y., Wang, L., Lee, L.h., Zhou, P.Y.: Dreamscene: 3d gaussian-based text-to-3d scene generation via formation pattern sampling. In: European Conference on Computer Vision. pp. 214–230. Springer (2024)
33. Li, P., Geng, H., Crate, J., Han, Y., Zhang, J., Wang, F., Cheng, C.T., Dong, R., Wang, Y.J., Lou, H., et al.: Rose: Reconstructing objects, scenes, and trajectories from casual videos for robotic manipulation. In: NeurIPS 2025 Workshop on Bridging Language, Agent, and World Models for Reasoning and Planning (2025)
34. Ling, L., Lin, C.H., Lin, T.Y., Ding, Y., Zeng, Y., Sheng, Y., Ge, Y., Liu, M.Y., Bera, A., Li, Z.: Scenethesis: A language and vision agentic framework for 3d scene generation. arXiv preprint arXiv:2505.02836 (2025)
35. Meng, Y., Wu, H., Zhang, Y., Xie, W.: Scenegen: Single-image 3d scene generation in one feedforward pass. In: 2026 International Conference on 3D Vision (3DV). pp. 543–553. IEEE (2026)
36. NVIDIA: Isaac Sim 4.5.0 (2025)
37. Öcal, B.M., Tatarchenko, M., Karaoğlu, S., Gevers, T.: Sceneteller: Language-to-3d scene generation. In: European Conference on Computer Vision. pp. 362–378. Springer (2024)
38. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
39. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
40. Ravi, N., Reizenstein, J., Novotny, D., Gordon, T., Lo, W.Y., Johnson, J., Gkioxari, G.: Accelerating 3d deep learning with pytorch3d. arXiv preprint arXiv:2007.08501 (2020)
41. Ren, T., Liu, S., Zeng, A., Lin, J., Li, K., Cao, H., Chen, J., Huang, X., Chen, Y., Yan, F., et al.: Grounded sam: Assembling open-world models for diverse visual tasks. arXiv preprint arXiv:2401.14159 (2024)
42. Seedream, T., Chen, Y., Gao, Y., Gong, L., Guo, M., Guo, Q., Guo, Z., Hou, X., Huang, W., Huang, Y., et al.: Seedream 4.0: Toward next-generation multimodal image generation. arXiv preprint arXiv:2509.20427 (2025)
43. Sun, F.Y., Liu, W., Gu, S., Lim, D., Bhat, G., Tombari, F., Li, M., Haber, N., Wu, J.: Layoutvlm: Differentiable optimization of 3d layout via vision-language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 29469–29478 (2025)
44. Sundaralingam, B., Hari, S.K.S., Fishman, A., Garrett, C., Van Wyk, K., Blukis, V., Millane, A., Oleynikova, H., Handa, A., Ramos, F., et al.: Curobo: Parallelized collision-free robot motion generation. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 8112–8119. IEEE (2023)
45. Tang, Z., Sundaralingam, B., Tremblay, J., Wen, B., Yuan, Y., Tyree, S., Loop, C., Schwing, A., Birchfield, S.: RGB-only reconstruction of tabletop scenes for collision-free manipulator control. In: ICRA (2023)

46. Team, O.M., Ghosh, D., Walke, H., Pertsch, K., Black, K., Mees, O., Dasari, S., Hejna, J., Kreiman, T., Xu, C., et al.: Octo: An open-source generalist robot policy. arXiv preprint arXiv:2405.12213 (2024)
47. Tian, Y., Yang, Y., Xie, Y., Cai, Z., Shi, X., Gao, N., Liu, H., Jiang, X., Qiu, Z., Yuan, F., et al.: Interndata-a1: Pioneering high-fidelity synthetic data for pre-training generalist policy. arXiv preprint arXiv:2511.16651 (2025)
48. Torne, M., Jain, A., Yuan, J., Macha, V., Ankile, L., Simeonov, A., Agrawal, P., Gupta, A.: Robot learning with super-linear scaling. arXiv preprint arXiv:2412.01770 (2024)
49. Torne, M., Simeonov, A., Li, Z., Chan, A., Chen, T., Gupta, A., Agrawal, P.: Reconciling reality through simulation: A real-to-sim-to-real approach for robust manipulation. arXiv preprint arXiv:2403.03949 (2024)
50. Xiang, F., Qin, Y., Mo, K., Xia, Y., Zhu, H., Liu, F., Liu, M., Jiang, H., Yuan, Y., Wang, H., et al.: Sapien: A simulated part-based interactive environment. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11097–11107 (2020)
51. Yang, Y., Lu, J., Zhao, Z., Luo, Z., Yu, J.J., Sanchez, V., Zheng, F.: Llplace: The 3d indoor scene layout generation and editing via large language model. arXiv preprint arXiv:2406.03866 (2024)
52. Yang, Y., Sun, F.Y., Weihs, L., VanderBilt, E., Herrasti, A., Han, W., Wu, J., Haber, N., Krishna, R., Liu, L., et al.: Holodeck: Language guided generation of 3d embodied ai environments. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16227–16237 (2024)
53. Yao, K., Zhang, L., Yan, X., Zeng, Y., Zhang, Q., Xu, L., Yang, W., Gu, J., Yu, J.: Cast: Component-aligned 3d scene reconstruction from an rgb image. ACM Transactions on Graphics (TOG) **44**(4), 1–19 (2025)
54. Yu, H.X., Duan, H., Herrmann, C., Freeman, W.T., Wu, J.: Wonderworld: Interactive 3d scene generation from a single image. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5916–5926 (2025)
55. Yu, H., Jia, B., Chen, Y., Yang, Y., Li, P., Su, R., Li, J., Li, Q., Liang, W., Zhu, S.C., et al.: Metascenes: Towards automated replica creation for real-world 3d scans. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 1667–1679 (2025)
56. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
57. Zhang, S., Wicke, P., Şenel, L.K., Figueredo, L., Naceri, A., Haddadin, S., Plank, B., Schütze, H.: Lohoravens: A long-horizon language-conditioned benchmark for robotic tabletop manipulation. arXiv preprint arXiv:2310.12020 (2023)
58. Zhao, W., Queralta, J.P., Westerlund, T.: Sim-to-real transfer in deep reinforcement learning for robotics: a survey. In: 2020 IEEE symposium series on computational intelligence (SSCI). pp. 737–744. IEEE (2020)
59. Zhou, J., Li, X., Qi, L., Yang, M.H.: Layout-your-3d: Controllable and precise 3d generation with 2d blueprint. arXiv preprint arXiv:2410.15391 (2024)
60. Zook, A., Sun, F.Y., Spjut, J., Blukis, V., Birchfield, S., Tremblay, J.: Grs: Generating robotic simulation tasks from real-world images. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 594–603 (2025)