PHYSCENE: Physically Interactable 3D Scene Synthesis for Embodied AI

Yandan Yang* Baoxiong Jia* Peiyuan Zhi Siyuan Huang
National Key Laboratory of General Artificial Intelligence, BIGAI
https://physcene.github.io

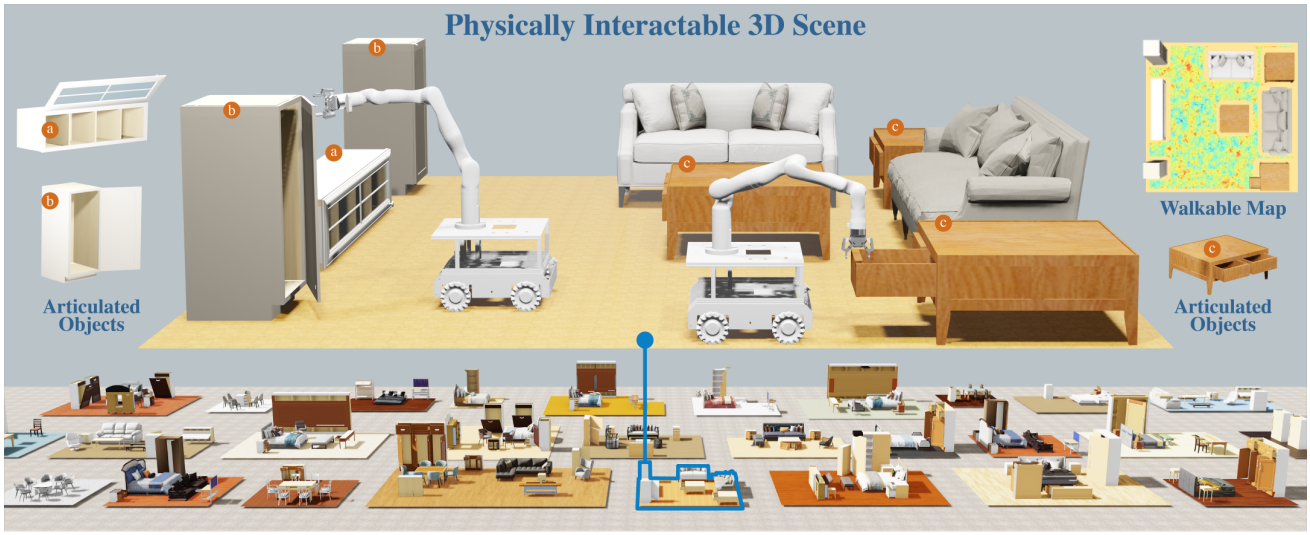

Figure 1. **Illustration of the PHYSCENE**, physically interactable scene synthesis method to generate interactive 3D scenes characterized by **realistic layouts, articulated objects, and rich physical interactivity** tailored for embodied agents.

Abstract

With recent developments in Embodied Artificial Intelligence (EAI) research, there has been a growing demand for high-quality, large-scale interactive scene generation. While prior methods in scene synthesis have prioritized the naturalness and realism of the generated scenes, the physical plausibility and interactivity of scenes have been largely left unexplored. To address this disparity, we introduce PHYSCENE*, a novel method dedicated to generating interactive 3D scenes characterized by realistic layouts, articulated objects, and rich physical interactivity tailored for embodied agents. Based on a conditional diffusion model for capturing scene layouts, we devise novel physics- and interactivity-based guidance mechanisms that integrate constraints from object collision, room layout, and object reachability. Through extensive experiments, we demonstrate that* PHYSCENE *effectively leverages these guidance functions for physically interactable scene synthesis, outperforming existing state-of-the-art scene synthesis methods by a large margin. Our findings suggest that the scenes generated by* PHYSCENE *hold considerable potential for facilitating diverse skill acquisition among agents within interactive environments, thereby catalyzing further advancements in embodied AI research.*

1. Introduction

The exploration of scene synthesis [7, 11, 14, 16, 30, 45, 54, 58, 62, 67] has constituted a persistent focus within the field of computer vision. Initially conceived to facilitate indoor design applications, scene synthesis aimed to create diverse 3D environments characterized by both realism and naturalness. However, with the advent of embodied artificial intelligence (EAI) [1, 12, 25, 27], the objectives of this task have taken on new dimensions. Simulated environments [9, 10, 33, 35, 50, 57], now supporting a plethora of intricate embodied tasks, have propelled the task of scene synthesis into an important data source that provides unlimited scenarios for agents to robustly learn skills like navigation [2, 34] and manipulation [18, 31, 48]. This trend underscores the growing importance of scene synthesis within

*indicates equal contribution.

the context of EAI research.

Nevertheless, achieving a seamless transition from conventional scene synthesis algorithms to those tailored for EAI presents significant challenges in scene generation. As many EAI tasks involve physics simulation [19, 36, 37, 39, 40, 65], the synthesized scenes must adhere to physical constraints while enabling a high degree of interactivity among objects (*e.g.*, articulated objects or fluids) and scene layout (*e.g.*, reachability of objects) to facilitate agent skill acquisition. These stringent interactivity requirements introduce several obstacles for scene synthesis algorithms. Limited by the quality of real-world scanned scenes [4, 8, 29], previous methods have primarily relied on manually created scenes [14, 15]. However, these datasets are designed with non-interactable objects, overlooking physical constraints, and are prone to violations of such constraints. Consequently, this poses a significant challenge for algorithms aiming to learn physically plausible arrangements of interactable objects. Beyond data-level hurdles, incorporating scene interactivity (*e.g.*, maintaining sufficient workspace, ensuring object reachability and interactivity) introduces non-trivial challenges in designing optimizable objectives that reflect such abstract concepts. These challenges emphasize the need for an effective scene synthesis algorithm that integrates the naturalness and realism of conventional synthesis algorithms while ensuring the physical plausibility and interactivity of scenes.

To address these challenges, we propose PHYSCENE, *a diffusion-based method embedded with physical commonsense for interactable scene synthesis.* Specifically, our approach builds on the efficacy of guided diffusion models [3, 23, 38, 51] to effectively learn scene distribution and guide the model in generating scenes that are both functionally interactive and physically plausible. To incorporate articulated objects into generated scenes, we utilize the shape and geometry features, bridging rigid-body objects from training scenes with existing articulated object datasets. To model physical plausibility and interactivity accurately, we impose three key constraints on the generated scenes: (1) **physical collision avoidance** between objects to enable simulation, (2) **object layouts** constrained on the floor plan to avoid inter-room conflicts, and (3) the **interactiveness and reachability** of each object when assuming an embodied agent of proper size need to navigate. We convert these constraints into guidance functions that can be easily integrated into the guided diffusion model. We further propose metrics considering the aforementioned constraints in our evaluation process for assessing all existing models. Through meticulously designed experiments, we demonstrate that PHYSCENE not only achieves state-of-the-art results on traditional scene synthesis metrics but also significantly enhances the physical plausibility and interactivity of generated scenes compared to existing methods. We hope this work can make a step forward in scalable indoor scene synthesis for EAI tasks, contributing to the broader landscape of EAI research.

In summary, our main contributions are:

- We propose PHYSCENE, a guided diffusion model, for physically interactable scene synthesis with realistic layouts and interactable objects.
- Through well-crafted designs of guidance functions, we convert constraints encompassing collision avoidance, room layout, and reachability into PHYSCENE in a simple and effective way to ensure the physical plausibility and interactivity of the generated scenes.
- By comparing with competitive baseline models, we show that PHYSCENE can not only achieve state-of-the-art results on traditional scene-synthesis metrics but also significantly outperforms existing methods for interactable scene synthesis on our delicately designed physical metrics, paving the way for new research topics bridging scene synthesis and EAI.

2. Related Work

Indoor Scene Synthesis Indoor scene synthesis is formulated as a layout prediction problem, where each object is often represented by its 3D bounding box, semantic labels [14, 54], or shape features [51] for retrieving corresponding meshes from 3D asset libraries to the specific locations. To properly model the layout of objects in training datasets, current methods usually represent the arrangement of objects as a scene graph [7, 11, 62, 67] and utilize scene priors such as the spatial relationship between objects [45] and object category (co-)occurrence frequency [16, 58] for approximating the scene layout distribution. While generating new scenes, these works leverage iterative sampling or optimization methods to reject scenes that violate the designed scene priors for synthesizing scenes with desired properties [7, 13, 16, 45]. However, such methods are often limited by the efficacy of sampling or optimization algorithms. More recent works try to learn scene layout distributions with deep neural networks [26, 41, 42, 44, 54, 59, 64] to improve the generation efficiency.

For the quality evaluation of generated scenes, common metrics test model performance with perceptual quality scores (*e.g.*, FID [22], KID [5],*etc.*). However, these realism metrics do not address the physical plausibility and interactivity of generated scenes, which is crucial for adapting scenes into simulated environments. In fact, a commonly used scene synthesis dataset, 3D-FRONT dataset [14], exhibits frequent occurrence of these physically implausible layouts (as shown in Tab. 1). In addition, the interactivity of scenes for object manipulation and reachability is also understudied in prior works. ProcTHOR [10] has proposed a procedural generation pipeline for interactable scenes with rule-based constraints and statistical scene priors. Nonetheless, as pointed out by [32], these generated scenes suffer from the pre-defined priors, thus generating unrealistic scenes that are harmful to agent learning. To this end, we aim to bridge this gap in PHYSCENE, uniting efforts in scene synthesis and EAI to provide a pipeline that could suf-

fice for large-scale interactable scene synthesis while maintaining visual realism and naturalness.

Physical Plausibility and Interactivity in 3D Scenes Producing physically plausible generations in 3D scenes has been a long-standing problem for computer vision, given its subtleness in properly converting physical constraints into optimizable objectives. To tackle this challenge, various optimization-based approaches have been proposed for tasks such as scene-conditioned pose [21] and motion generation [52]. However, the study of physical plausibility for scene generation has been largely left untouched. Meanwhile, the modeling of interactivity of 3D scenes has been largely left untouched in existing works without proper definition. Some works [53] aim to define the level of scene interactivity via human and robot preferences in a scene rearrangement setting. Nonetheless, with their task-specific design, the optimization objectives are hard to be generalized to other settings. Therefore, PHYSCENE aims at addressing these obstacles and makes the first attempts to provide reasonable definitions of physical plausibility and scene interactivity in the context of scene synthesis.

Guided Diffusion Models Diffusion models [24, 38, 49] have shown promising results for generative AI [26, 28, 47] across various domains [43, 55, 56, 60, 63]. Through an iterative denoising process, diffusion models excel at handling high dimensional distributions without mode collapse. Such an iterative process also offers flexible ways to provide conditions [6, 46] and guidance [3, 23] that could effectively affect the inference of models. For example, SceneDiffuser [26] integrates a physics-based objective as conditional guidance for physically plausible planning and motion generation. PhysDiff [61] proposes a physics-based motion projection module to instill the laws of physics into the denoising diffusion process for motion generation. PHYSCENE takes insight from these powerful techniques and integrates physical and interactivity guidance as conditional guidance for scene synthesis. Compared to constrained sampling methods such as Markov Chain Monte Carlo (MCMC) [45, 53], diffusion guidance runs more efficiently during the inference stage. Meanwhile, in contrast to models that take in constraints as a learnable objective [51], our guidance functions can more effectively ensure the satisfaction of constraints during inference. To the best of our knowledge, PHYSCENE makes the first attempt to integrate a conditional diffusion model with physical plausibility and interactivity guidances to effectively generate physically interactable 3D scenes.

3. PHYSCENE

Physically interactable scene synthesis requires realistic layouts, articulated objects, and physical interactivity. However, integrating articulated objects into scenes trained solely with static objects presents data-level challenges.

Table 1. **Interactivity evaluation of scenes in the 3D-FRONT dataset.** These scenes exhibit a high rate of physical constraint violations including collision, layout, and interactivity. We provide detailed definitions of the metrics as explained in Sec. 4.

Data	Bedroom	Livingroom	Diningroom
$\text{Col}_{\text{obj}} \downarrow$	0.214	0.206	0.209
$\text{Col}_{\text{scene}} \downarrow$	0.42	0.625	0.57
$\text{R}_{\text{out}} \downarrow$	0.201	0.0584	0.159
$\text{R}_{\text{reach}} \uparrow$	0.850	0.841	0.876
$\text{R}_{\text{walkable}} \uparrow$	0.749	0.828	0.807

We outline our method for incorporating articulated objects into generated scenes in Sec. 3.1. We then detail the model structure and training process of PHYSCENE, where it learns prior layout knowledge from the dataset in Sec. 3.2. To ensure physical interactivity, we consider collision avoidance, room layout constraint, and agent interactiveness as three key constraints, and provide details in Sec. 3.3 on transforming them into guidance functions for posterior optimization during the inference process.

3.1. Object representation

The scene $\mathbf{x}$ is composed of N objects, noted as $\mathbf{x} = \{\mathbf{o}_1, ..., \mathbf{o}_\mathbf{N}\}$. Each object representation $\mathbf{o}_i = [\mathbf{c}_i, \mathbf{s}_i, \mathbf{r}_i, \mathbf{t}_i, \mathbf{f}_i]$ is composed of a semantic label $\mathbf{c}_i \in \mathbb{R}^C$ out of C categories, size $\mathbf{s}_i \in \mathbb{R}^3$, orientation $\mathbf{r}_i = (cos\theta_i, sin\theta_i) \in \mathbb{R}^2$, location $\mathbf{t}_i \in \mathbb{R}^3$ and 3D feature $\mathbf{f}_i \in \mathbb{R}^{32}$ encoded from the shape of the object. Notably, common approaches for scene synthesis retrieve objects using the predicted size $\mathbf{s}_i$ and label $\mathbf{c}_i$. However, such methods could not be applied across asset libraries. We therefore leverage the shape feature $\mathbf{f}_i$ as a critical indicator for object retrieval, especially considering the objects in available articulated object datasets are largely different from those in scene synthesis datasets. Specifically, we follow [51] and utilize a variational auto-encoder to embed object geometric features, transforming each 3D furniture model into a latent shape feature. For generating scenes with interactable objects, we consider object assets from: 1) 3D-FUTURE [15] which contains CAD models used in 3D-FRONT [14], and 2) GAPartNet [17] that includes various articulated objects. During inference, we use the latent encoded feature to find the best match of articulated objects in GAPartNet given the static objects in 3D-Front, thereby enabling the generation of scenes containing interactable objects.

3.2. Conditional Diffusion for Layout Modeling

With a data sample $\mathbf{x}_0$ representing the scene layout in the dataset, we gradually add Gaussian noise to $\mathbf{x}_0$ with a forward process $q(\mathbf{x}_{t+1}|\mathbf{x}_t)$ converting it into a Gaussian noise $\mathbf{x}_T$. Then a reverse denoising process $p_\theta(\mathbf{x}_t|\mathbf{x}_{t+1})$ is applied to recover the data from noise with learnable parameters θ. Additionally, we consider using the floor plan $\mathcal{F}$ as a condition for incorporating the workspace and room layout

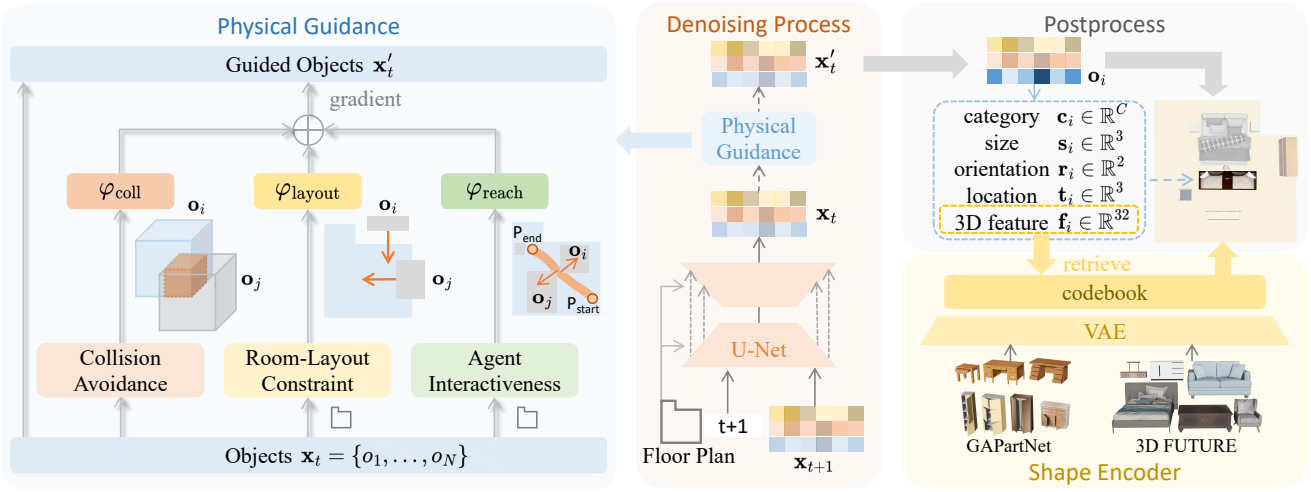

Figure 2. **Overview of PHYSCENE.** We leverage diffusion models for capturing scene layout distributions and apply three distinct guidance functions for improving the physical plausibility and interactivity of generated scenes.

constraints. In this case, we reconstruct $\mathbf{x}_0$ via:

$$p_\theta(\mathbf{x}_0|\mathcal{F}) = p(\mathbf{x}_T)\prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathcal{F}),$$
$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathcal{F}) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t, \mathcal{F}), \Sigma_\theta(\mathbf{x}_t, t, \mathcal{F})),$$

where $p_\theta(\mathbf{x}_0|\mathcal{F})$ denotes the probability of scene layout $\mathbf{x}_0$ given the conditional floor plan $\mathcal{F}$. As pointed out by previous works [24], this maximization of conditional probability $p_\theta(\mathbf{x}_0|\mathcal{F})$ could be equivalently formulated as a simplified objective of estimating the noise ϵ through:

$$\begin{aligned}\mathcal{L}_\theta(\mathbf{x}_0|\mathcal{F}) &= \mathbb{E}_{t,\boldsymbol{\epsilon},\mathbf{x}_0}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\hat{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\hat{\alpha}_t}\boldsymbol{\epsilon}, t, \mathcal{F})\|_2^2\right] \\ &= \mathbb{E}_{t,\boldsymbol{\epsilon},\mathbf{x}_0}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, \mathcal{F})\|_2^2\right],\end{aligned} \tag{1}$$

where $\hat{\alpha}_t$ is a pre-defined function of t in the forward process according to a noise schedule (see details in the *supplementary*). To learn this conditional model, we utilize a U-Net with attention blocks to model $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, \mathcal{F})$ with time embedding t and floor plan embedding $\mathcal{F}$ added as conditions within every U-Net layer.

3.3. Guidance for Physical Interactivity

Considering the physical constraints violations in scenes from existing training data (as shown in Tab. 1), we ensure the physical plausibility and interactivity of the generated scenes by guiding the conditional scene diffusion process with physic-based guidance functions. We start by first introducing guided sampling for diffusion models. Given constraint function $\varphi(\mathbf{x}, \mathcal{F})$, we formulate the guided inference problem as optimizing the probability of constraint satisfaction:

$$\begin{aligned}p(\mathbf{x}_0|\mathcal{F}, O=1) &\propto p_\theta(\mathbf{x}_0|\mathcal{F})p(O=1|\mathbf{x}_0, \mathcal{F}) \\ &\propto p_\theta(\mathbf{x}_0|\mathcal{F}) \cdot \exp\left(\varphi(\mathbf{x}_0, \mathcal{F})\right),\end{aligned} \tag{2}$$

where O is an optimality indicator checking if the conditional generated output $\mathbf{x}_t$ at denoising step t satisfies the constraints in $\varphi(\mathbf{x}, \mathcal{F})$. Similar to [26], we use the first order Taylor expansion around $\mathbf{x}_t = \boldsymbol{\mu}$ at timestep t to estimate the optimal condition in Eq. (2) with:

$$\begin{aligned}\log p(O=1|\mathbf{x}_t, \mathcal{F}) &\approx (\mathbf{x}_t - \boldsymbol{\mu})\mathbf{g} + C \\ \mathbf{g} &= \nabla_{\mathbf{x}_t} \log p(O=1|\mathbf{x}_t, \mathcal{F})|_{\mathbf{x}_t=\boldsymbol{\mu}} \\ &= \nabla_{\mathbf{x}_t} \varphi(\mathbf{x}_t, \mathcal{F})|_{\mathbf{x}_t=\boldsymbol{\mu}},\end{aligned} \tag{3}$$

where $\boldsymbol{\mu} = \boldsymbol{\mu}_\theta(\mathbf{x}_t, t, \mathcal{F})$, $\mathbf{g}$ is the first order gradient estimate at $\mathbf{x}_t = \boldsymbol{\mu}$ of $\log p(O=1|\mathbf{x}_t, \mathcal{F})$, and C is a constant. Therefore to generate a scene with constraints considered, we can modify the denoising process with a constraint perturbed Gaussian transition:

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathcal{F}, O=1) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu} + \lambda\boldsymbol{\Sigma}\mathbf{g}, \boldsymbol{\Sigma}), \tag{4}$$

where $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t, \mathcal{F})$ and λ is a scaling factor. Notably, the formulations in Eq. (2) and Eq. (4) leverage the predefined constraint functions $\varphi(\mathbf{x}, \mathcal{F})$ as a tilting function on the original scene layout distribution to handle constraints.

Under this formulation, we can easily combine the constraint functions into both learning and inference. Following Eq. (1), we can reformulate the optimization of objective with $\varphi(\mathbf{x}, \mathcal{F})$ through:

$$\mathcal{L}_\theta(\mathbf{x}_0|\mathcal{F}, O=1) = \mathbb{E}_{t,\boldsymbol{\epsilon},\mathbf{x}_0}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, \mathcal{F}) - \lambda\boldsymbol{\Sigma}\mathbf{g}\|_2^2\right] \tag{5}$$

In scene synthesis, the guidance functions $\varphi(\mathbf{x}_t, \mathcal{F})$ usually require real scene layouts for computing the violation constraints. Therefore, instead of optimizing for $\mathbf{x}_t$ which might not be meaningful for real scenes, we convert the guidance functions into $\varphi(\tilde{\mathbf{x}}_0^t, \mathcal{F})$ where $\tilde{\mathbf{x}}_0^t$ is the predicted scene layout given initialization $\mathbf{x}_t$. We summarize the guided learning and inference process of PHYSCENE in Algorithm 1.

Algorithm 1: Learnning and inference in PHYSCENE

Modules: Model $p_\theta(\cdot|\mathcal{F})$, guidance functions $\varphi(\cdot,\mathcal{F}) = \{\varphi_{\text{coll}}(\cdot), \varphi_{\text{layout}}(\cdot,\mathcal{F}), \varphi_{\text{reach}}(\cdot,\mathcal{F})\}$.

// constraint-guided learning

Input: 3D scene layout $\mathbf{x} = \{\mathbf{o}_1, ..., \mathbf{o}_N\}$ with floor plan $\mathcal{F}$, where N is a fixed number of objects.

repeat

$\mathbf{x}_0 \sim p(\mathbf{x}_0|\mathcal{F})$

$\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), t \sim \mathcal{U}(\{1, \cdots, T\})$

$\mathbf{x}_t = \sqrt{\hat{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\hat{\alpha}_t}\boldsymbol{\epsilon}$, $\tilde{\mathbf{x}}_0^t \sim p_\theta(\cdot|\mathcal{F})$

$\theta = \theta - \eta\nabla_\theta\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, \mathcal{F}) - \lambda\boldsymbol{\Sigma}\mathbf{g}\|_2^2$

until *converged*;

// one-step guided sampling

function sample $(\boldsymbol{\tau}^t, \varphi)$**:**

$\boldsymbol{\mu} = \boldsymbol{\mu}_\theta(\mathbf{x}_t, t, \mathcal{F})$, $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t, \mathcal{F})$

$\varphi(\mathbf{x}_t, \mathcal{F}) =$

$\gamma_1\varphi_{\text{coll}}(\mathbf{x}_t) + \gamma_2\varphi_{\text{layout}}(\mathbf{x}_t, \mathcal{F}) + \gamma_3\varphi_{\text{reach}}(\mathbf{x}_t, \mathcal{F})$

$\mathbf{x}_{t-1} = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu} + \lambda\boldsymbol{\Sigma}\nabla_{\mathbf{x}_t}\varphi(\mathbf{x}_t, \mathcal{F})|_{\mathbf{x}_t=\boldsymbol{\mu}}, \boldsymbol{\Sigma})$

return $\mathbf{x}_{t-1}$

// constraint-guided generation

Input: initial scene layout $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$

for $t = T, \cdots, 1$ **do**

// sampling with optimization

$\mathbf{x}_{t-1} =$ **sample**$(\mathbf{x}_t, \varphi)$

end

return $\mathbf{x}_0$

Based on this formulation, we further propose three physic-based guidances $\varphi_{\text{coll}}(\mathbf{x})$, $\varphi_{\text{layout}}(\mathbf{x}, \mathcal{F})$, and $\varphi_{\text{reach}}(\mathbf{x}, \mathcal{F})$ and integrate them into the inference process as illustrated in Algorithm 1. We detail the design of each guidance function as follows:

Collision Avoidance. We design a collision avoidance function to reduce object mesh collisions in the generated scene. Instead of calculating the collision mesh between objects, we use the predicted bounding boxes and object centers as effective approximates for estimating the collision score of objects. Specifically, we use $\boldsymbol{b}_i = [\mathbf{t}_i, \mathbf{r}_i, \mathbf{s}_i]$ to denote the 3D bounding box of object $\boldsymbol{o}_i$ including its location $\mathbf{t}_i$, orientation $\mathbf{r}_i$ and size $\mathbf{s}_i$. We use 3D IoU [66] to calculate the collision guidance objective via:

$$\varphi_{\text{coll}}(\mathbf{x}) = -\sum_{i,j,i\neq j} \textbf{IoU}_{3D}(\boldsymbol{b}_i, \boldsymbol{b}_j), \tag{6}$$

where $\textbf{IoU}_{3D}$ represents the 3D bounding box IoU between object bounding boxes. We sum the collisions of each pair of objects in scene $\boldsymbol{x}$ and take the negative value of the summation to penalize object collision.

Room-layout guidance An important goal of scalable scene synthesis is to generate interactable house-level scenes in which embodied agents can navigate and interact. To achieve this goal, we consider adding the room-layout guidance that penalizes the existence of objects which are outside of a pre-given floor plan. To consolidate this guidance function, we first extract a polygon of the room boundary given the floor plan $\mathcal{F}$. We then derive a set of W outside barriers for identifying the boundary, represented as bounding boxes of walls $\{\mathbf{b}_w^{\text{wall}}\}_{w=1}^W$ with infinite thickness. We use a similar IoU score between objects and walls for room-layout guidance following:

$$\varphi_{\text{layout}}(\mathbf{x}|\mathcal{F}) = -\sum_{i=1}^{N}\sum_{j=1}^{W} \textbf{IoU}_{3D}(\boldsymbol{b}_i, \boldsymbol{b}_j^{\text{wall}}). \tag{7}$$

Reachability guidance For an embodied agent, the synthesized scene should allow it to traverse the entire room and interact with all objects successfully. Notably, the synthesized room is often separated into several disjoint connected regions in scenes with improper layouts. Based on this key observation, we aim to adjust the object locations that most significantly affect this connectivity between regions. More specifically, considering an embodied agent represented by its bounding box $\boldsymbol{b}^{\text{agent}}$, we first map the generated scenes to a 2D room mask and calculate the walkable area in this scene considering the agent's size. Next, we employ Gaussian distributions on each positioned object in the scene to form a cost map for traversing the scene. Intuitively, points closer to objects will have higher costs. With the cost map, we plan the shortest path between the center of the two largest connection regions using the A* algorithm [20]. The resulting path indicates the least effort path to traverse between these two regions. We then select L agent positions on this shortest path with bounding boxes $\{\boldsymbol{b}_1^{\text{agent}}, ..., \boldsymbol{b}_L^{\text{agent}}\}$ for applying the guidance function. The reachability guidance can therefore be calculated via:

$$\varphi_{\text{reach}}(\mathbf{x}|\mathcal{F}) = -\sum_{i=1}^{N}\sum_{j=1}^{L} \textbf{IoU}_{3D}(\boldsymbol{b}_i, \boldsymbol{b}_j^{\text{agent}}). \tag{8}$$

Notably, we can extend the current method to incorporate interaction constraints to ensure the articulated object interaction (*e.g.*, grasping, opening) as well as complex rigid object interaction (*e.g.*, sit) with some simple modifications. More details are provided in the *supplementary*.

4. Experiment

Dataset For experimental comparisons, we train our diffusion model on the 3D-FRONT dataset [14] which contains 6813 houses with 14629 rooms. Each room is manually decorated with high-quality furniture objects from the 3D-FUTURE dataset [15]. Following the setting of DiffuScene [51] and ATISS [42], we use 4041 bedrooms, 900 dining rooms, and 813 living rooms for training and testing. In addition, we use both the 3D-FUTURE dataset [15] and GAPartNet [17] for object retrieval. Among them, GAPartNet [17] has abundant interactive assets, containing 1166 articulated objects from 27 object categories. We utilize articulated objects in the table and storage furniture category, such as wardrobe and table, to retrieve related objects in

Table 2. **Quantitative comparison on unconditional scene synthesis trained on 3D-Front.** We compare PHYSCENE with ATISS and Diffuscene on common perceptual quality scores FID, SCA, CKL, as well as physical plausibility measured in collision rate **Col**.

Method	Bedroom					Living Room					Dining Room				
	FID ↓	SCA ↓	CKL ↓	Col_{obj} ↓	$\text{Col}_{\text{scene}}$ ↓	FID ↓	SCA ↓	CKL ↓	Col_{obj} ↓	$\text{Col}_{\text{scene}}$ ↓	FID ↓	SCA ↓	CKL ↓	Col_{obj} ↓	$\text{Col}_{\text{scene}}$ ↓
ATISS [42]	36.92	**49.24**	0.0036	0.255	0.50	55.76	53.33	0.0016	0.372	0.870	41.89	58.20	0.0028	0.483	0.91
DiffuScene [51]	28.63	51.33	0.0031	0.238	0.42	54.36	**50.24**	**0.0010**	0.183	0.570	**37.68**	**57.60**	0.0031	0.253	0.63
PhyScene (Ours)	**28.56**	55.71	**0.0030**	**0.187**	**0.35**	**40.67**	56.20	0.0015	**0.130**	**0.477**	37.88	58.74	**0.0022**	**0.134**	**0.40**

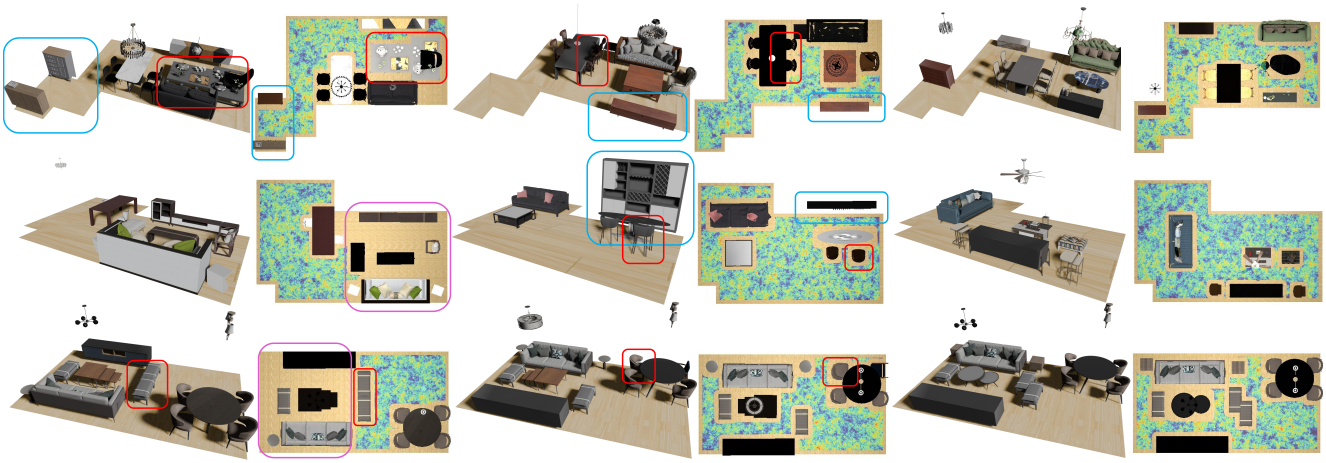

Figure 3. **Visualization of floor-plan conditioned scene synthesis between PhyScene, ATISS, and DiffuScene.** The **red**, **purple**, and **blue** boxes highlight collisions between objects, objects outside the floor plan, and unreachable areas to the embodied agent, respectively.

generated scenes, and provide the full object category mapping between datasets in the *supplementary*.

Baseline We mainly consider two state-of-the-art scene synthesis methods as baselines: 1) ATISS [42], a transformer-based model that predicts the 3D object bounding box in an autoregressive manner, and 2) DiffuScene [51], a diffusion-based model that learns 3D objects layout without floor plan constraint. We test these baselines in both unconditional synthesis and floor-plan-conditioned synthesis settings to compare our proposed PHYSCENE model.

Metric To evaluate the realism and diversity of the synthesized scenes, we follow the previous works and calculate Fréchet Inception Distance [22] (FID), Kernel Inception Distance [5] (KID $\times 0.001$), Scene Classification Accuracy (SCA), and Category KL divergence (CKL $\times 0.01$) on 1000 synthesized scenes. In addition, we check the collision rate between each pair of objects in the generated scene using their CAD models. We use $\mathbf{Col}_{\text{obj}}$ to denote the percentage of objects that collide with other objects in the generated scene, $\mathbf{Col}_{\text{scene}}$ to denote the ratio of scenes that possess object collisions over all generated scenes. Since the CAD models in the 3D-FUTURE dataset are usually not watertight, we apply re-meshing for each object mesh before evaluation. To evaluate the violation of the floor plan layout, we mark the rate of objects outside the floor plan as $\mathbf{R}_{\text{out}}$. Finally, we calculate the average reachable rate of objects in the scene $\mathbf{R}_{\text{reach}}$ starting from a random starting point on the floor plan. We calculate the average ratio of the largest connected walkable area over all walkable areas in the room, denoted as $\mathbf{R}_{\text{walkable}}$, to evaluate the reachability and interactivity of the generated scenes.

4.1. Unconditioned Scene Synthesis

We provide quantitative evaluation results in Tab. 2. As shown in Tab. 2, PHYSCENE achieves state-of-the-art results on almost all metrics, especially with a significant improvement on physical plausibility metrics such as $\mathbf{Col}_{\text{obj}}$ and $\mathbf{Col}_{\text{scene}}$. This result quantitatively proves that PHYSCENE effectively produces improved scene layouts with reduced collision rates while achieving better visual plausibility. Notably, diffusion-based scene-synthesis models (*i.e.*, DiffuScene and PHYSCENE) exhibit superior performance in collision avoidance compared to ATISS. This affirms the advantage of employing diffusion models as the primary generative model for scene synthesis, given their robust performance and adaptability in integrating guidance functions. We provide qualitative results in Fig. 3, demonstrating that our model successfully generates scenes with significantly fewer instances of physical constraint violations due to object collisions while maintaining high levels of naturalness and diversity.

Table 3. **Floor-conditioned Scene Synthesis.** We compare PHYSCENE with ATISS and DiffuScene on common perceptual quality scores FID, KID, SCA, CKL, as well as physical plausibility metrics $\mathbf{Col}_{\text{obj}}$, $\mathbf{Col}_{\text{scene}}$, $\mathbf{R}_{\text{out}}$, $\mathbf{R}_{\text{reach}}$, $\mathbf{R}_{\text{walkable}}$.

Room Type	Method	FID ↓	KID ↓	SCA ↓	CKL ↓	$\mathbf{Col}_{\text{obj}}$ ↓	$\mathbf{Col}_{\text{scene}}$ ↓	$\mathbf{R}_{\text{out}}$ ↓	$\mathbf{R}_{\text{walkable}}$ ↑	$\mathbf{R}_{\text{reach}}$ ↑
Bedroom	ATISS	30.19	0.0010	49.14	0.0028	0.248	0.46	0.286	0.839	0.736
	DiffuScene	**25.00**	**0.0004**	51.78	0.0031	0.228	0.43	0.272	0.827	0.755
	PhyScene (Ours)	25.52	0.0006	**50.10**	**0.0025**	**0.187**	**0.36**	**0.245**	**0.865**	**0.762**
Living Room	ATISS	45.66	0.0035	**51.64**	0.0016	0.316	0.85	**0.136**	0.814	**0.791**
	DiffuScene	**38.69**	**0.0012**	54.06	0.0017	0.198	0.69	0.238	0.790	0.756
	PhyScene (Ours)	43.33	0.0031	53.50	**0.0015**	**0.191**	**0.63**	0.219	**0.815**	0.771
Dining Room	ATISS	41.66	0.0039	64.57	0.0040	0.591	0.96	**0.132**	**0.874**	**0.848**
	DiffuScene	**38.31**	**0.0020**	60.19	**0.0013**	0.160	0.55	0.244	0.787	0.847
	PhyScene (Ours)	39.90	0.0026	**60.00**	**0.0013**	**0.151**	**0.53**	0.217	0.852	0.789

Figure 4. **Generated scenes with articulated objects.** We visualize the opening sequence of articulated objects (left) and the generated scenes with texture (right).

Table 4. **Articulated Object Embedding.** We compare PHYSCENE with ATISS and DiffuScene on physical plausibility.

Method	$\mathbf{Col}_{\text{obj}}$ ↓	$\mathbf{Col}_{\text{scene}}$ ↓	$\mathbf{R}_{\text{out}}$ ↓	$\mathbf{R}_{\text{reach}}$ ↑
ATISS	0.360	0.86	**0.154**	**0.758**
DiffuScene	0.262	0.78	0.237	0.702
PhyScene (Ours)	**0.251**	**0.76**	0.229	0.755

Table 5. **Ablation study on the use of guidance functions.** Our final result balances the effectiveness of three guidances.

Collision	Layout	Interact	$\mathbf{Col}_{\text{obj}}$ ↓	$\mathbf{R}_{\text{out}}$ ↓	$\mathbf{R}_{\text{walkable}}$ ↑	$\mathbf{R}_{\text{reach}}$ ↑
			0.200	0.240	0.808	0.763
✓			0.111	0.354	0.832	0.793
	✓		0.279	0.110	0.774	0.742
		✓	0.239	0.260	0.927	0.813
✓	✓	✓	0.191	0.219	0.815	0.771

4.2. Floor-conditioned Scene Synthesis

We provide comparisons between PHYSCENE and baseline models in terms of both visual and physical metrics in Tab. 3. PHYSCENE surpasses baselines in collision metrics and the CKL score. Additionally, compared to DiffuScene, our model consistently exhibits performance improvements across all physical interactability metrics, highlighting the effectiveness of our physical guidance functions in enhancing the generation process of diffusion-based models with physical constraints. It is noteworthy that, except for the Bedroom setting, ATISS achieves favorable results on floor plan violation ($\mathbf{R}_{\text{out}}$) and reachability metric ($\mathbf{R}_{\text{reach}}$). We attribute this to its prioritization of floor plan constraints over collision avoidance within the scene. We provide qualitative visualization of all models' generations in Fig. 3.

4.3. Scene Synthesis with Articulated Objects

To generate scenes with articulated objects, we utilize the predicted scene layout along with object features to retrieve articulate objects. Recognizing the spatial requirements for interacting with articulated objects, we compute 3D bounding boxes for these objects, considering their joints being manipulated to the fullest extent, and use these expanded bounding boxes for guidance calculation. We show the quantitative results of our guided substitution for articulated objects in the Living Room setting in Tab. 4. Results show the collision rate with articulated objects is much higher than that with rigid objects (compared with the collision rate shown in Tab. 3). And our model shows a great improvement over previous methods. We visualize the qualitative results of our guided substitution for articulated objects in Fig. 4 and leave more visualizations in the *supplementary*.

4.4. Ablation Study on Guidance

We conduct ablative studies on our proposed guidance functions in the Living Room setting in Tab. 5. Given that these guidance functions serve different roles in layout optimization, they may exhibit potential conflicts with each other. As shown in Tab. 5, the $\mathbf{Col}_{\text{obj}}$ and $\mathbf{R}_{\text{out}}$ metrics have a negative impact on each other because the collision guidance pushes objects apart while the floor plan guidance pushes objects closer to fit in the scene. However, we managed to strike

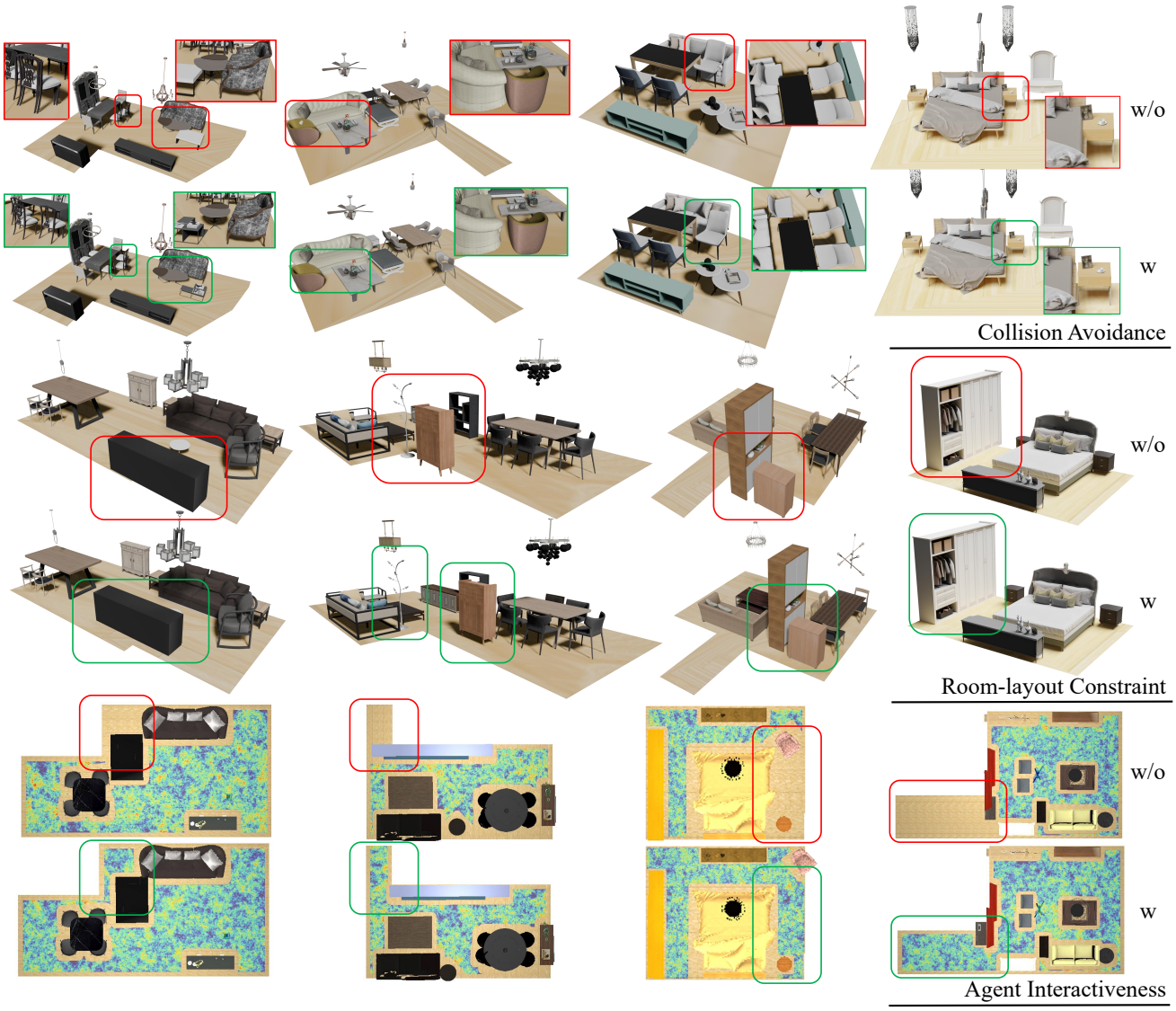

Figure 5. **Ablation on Guidance.** Results of different guidance with floor-plan conditions. For each ablation on guidance functions, we show four generated scenes (four columns) without guidance in the first row and mark the violation of constraints in **red** boxes. The second row shows the improvement after considering guidance functions in **green** boxes.

a balance among these guidances, leading to improvements on each corresponding metric. We provide qualitative visualizations illustrating the effect of each guidance in Fig. 5.

5. Conclusion

In this paper, we introduce PHYSCENE, a guided conditional diffusion model for physically interactable scene synthesis. To ensure the physical plausibility and interactivity of the generated scenes, we devise novel guidance functions converting constraints on object collision, room layout, and interactivity to guidance within each inference step in the diffusion process. Our experimental results demonstrate consistent performance improvement over state-of-the-art baseline models on physical plausibility and interactivity metrics, showcasing the effectiveness of our designed guidance functions and the generation pipeline.

Future work Due to data limitations, PHYSCENE is presently restricted to considering only limited room types, without incorporating small objects. This limitation poses a significant obstacle to the applicability of these scenes in embodied AI tasks, particularly those involving small object manipulation such as pick and place tasks. We leave this area as an important focus for future research.

Acknowledgement We thank Ms. Zhen Chen from BIGAI for refining the figures, and all colleagues from the BIGAI TongVerse project for fruitful discussions and help on simulation developments. We would also like to thank the anonymous reviewers for their constructive feedback.

References

[1] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, et al. Do as i can, not as i say: Grounding language in robotic affordances. In *Conference on Robot Learning (CoRL)*, 2022. 1

[2] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1

[3] Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 3

[4] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, et al. Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2

[5] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018. 2, 6

[6] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[7] Angel Chang, Manolis Savva, and Christopher D Manning. Learning spatial knowledge for text to 3d scene generation. In *Proceedings of the conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014. 1, 2

[8] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[9] Matt Deitke, Winson Han, Alvaro Herrasti, Aniruddha Kembhavi, Eric Kolve, Roozbeh Mottaghi, Jordi Salvador, Dustin Schwenk, Eli VanderBilt, Matthew Wallingford, et al. Robothor: An open simulation-to-real embodied ai platform. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1

[10] Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Kiana Ehsani, Jordi Salvador, Winson Han, Eric Kolve, Aniruddha Kembhavi, and Roozbeh Mottaghi. Procthor: Large-scale embodied ai using procedural generation. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 1, 2

[11] Helisa Dhamo, Fabian Manhardt, Nassir Navab, and Federico Tombari. Graph-to-3d: End-to-end generation and manipulation of 3d scenes using scene graphs. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2021. 1, 2

[12] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. In *Proceedings of International Conference on Machine Learning (ICML)*, 2023. 1

[13] Matthew Fisher, Manolis Savva, Yangyan Li, Pat Hanrahan, and Matthias Nießner. Activity-centric scene synthesis for functional 3d scene modeling. *ACM Transactions on Graphics (TOG)*, 34(6):1–13, 2015. 2

[14] Huan Fu, Bowen Cai, Lin Gao, Ling-Xiao Zhang, Jiaming Wang, Cao Li, Qixun Zeng, Chengyue Sun, Rongfei Jia, Binqiang Zhao, et al. 3d-front: 3d furnished rooms with layouts and semantics. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2021. 1, 2, 3, 5

[15] Huan Fu, Rongfei Jia, Lin Gao, Mingming Gong, Binqiang Zhao, Steve Maybank, and Dacheng Tao. 3d-future: 3d furniture shape with texture. *International Journal of Computer Vision (IJCV)*, 129:3313–3337, 2021. 2, 3, 5

[16] Qiang Fu, Xiaowu Chen, Xiaotian Wang, Sijia Wen, Bin Zhou, and Hongbo Fu. Adaptive synthesis of indoor scenes via activity-associated object relation graphs. *ACM Transactions on Graphics (TOG)*, 36(6):1–13, 2017. 1, 2

[17] Haoran Geng, Helin Xu, Chengyang Zhao, Chao Xu, Li Yi, Siyuan Huang, and He Wang. Gapartnet: Cross-category domain-generalizable object perception and manipulation via generalizable and actionable parts. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3, 5

[18] Ran Gong, Jiangyong Huang, Yizhou Zhao, Haoran Geng, Xiaofeng Gao, Qingyang Wu, Wensi Ai, Ziheng Zhou, Demetri Terzopoulos, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. Arnold: A benchmark for language-grounded task learning with continuous states in realistic 3d scenes. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2023. 1

[19] Jiayuan Gu, Fanbo Xiang, Xuanlin Li, Zhan Ling, Xiqiang Liu, Tongzhou Mu, Yihe Tang, Stone Tao, Xinyue Wei, Yunchao Yao, et al. Maniskill2: A unified benchmark for generalizable manipulation skills. *arXiv preprint arXiv:2302.04659*, 2023. 2

[20] Peter Hart, Nils Nilsson, and Bertram Raphael. A formal basis for the heuristic determination of minimum cost paths. *IEEE Transactions on Systems Science and Cybernetics*, 4 (2):100–107, 1968. 5

[21] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2019. 3

[22] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 2, 6

[23] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 2, 3

[24] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 3, 4

[25] Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. An embodied generalist agent in 3d world. *arXiv preprint arXiv:2311.12871*, 2023. 1

[26] Siyuan Huang, Zan Wang, Puhao Li, Baoxiong Jia, Tengyu Liu, Yixin Zhu, Wei Liang, and Song-Chun Zhu. Diffusion-based generation, optimization, and planning in 3d scenes. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 3, 4

[27] Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, et al. Inner monologue: Embodied reasoning through planning with language models. In *Conference on Robot Learning (CoRL)*, 2022. 1

[28] Naoto Inoue, Kotaro Kikuchi, Edgar Simo-Serra, Mayu Otani, and Kota Yamaguchi. Layoutdm: Discrete diffusion model for controllable layout generation. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[29] Baoxiong Jia, Yixin Chen, Huangyue Yu, Yan Wang, Xuesong Niu, Tengyu Liu, Qing Li, and Siyuan Huang. Sceneverse: Scaling 3d vision-language learning for grounded scene understanding. *arXiv preprint arXiv:2401.09340*, 2024. 2

[30] Chenfanfu Jiang, Siyuan Qi, Yixin Zhu, Siyuan Huang, Jenny Lin, Lap-Fai Yu, Demetri Terzopoulos, and Song-Chun Zhu. Configurable 3d scene synthesis and 2d image rendering with per-pixel ground truth using stochastic grammars. *International Journal of Computer Vision (IJCV)*, pages 920–941, 2018. 1

[31] Yunfan Jiang, Agrim Gupta, Zichen Zhang, Guanzhi Wang, Yongqiang Dou, Yanjun Chen, Li Fei-Fei, Anima Anandkumar, Yuke Zhu, and Linxi Fan. Vima: General robot manipulation with multimodal prompts. *arXiv*, 2022. 1

[32] Mukul Khanna, Yongsen Mao, Hanxiao Jiang, Sanjay Haresh, Brennan Schacklett, Dhruv Batra, Alexander Clegg, Eric Undersander, Angel X Chang, and Manolis Savva. Habitat synthetic scenes dataset (hssd-200): An analysis of 3d scene scale and realism tradeoffs for objectgoal navigation. *arXiv preprint arXiv:2306.11290*, 2023. 2

[33] Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Matt Deitke, Kiana Ehsani, Daniel Gordon, Yuke Zhu, et al. Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474*, 2017. 1

[34] Jacob Krantz, Erik Wijmans, Arjun Majumdar, Dhruv Batra, and Stefan Lee. Beyond the nav-graph: Vision-and-language navigation in continuous environments. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2020. 1

[35] Chengshu Li, Fei Xia, Roberto Martín-Martín, Michael Lingelbach, Sanjana Srivastava, Bokui Shen, Kent Vainio, Cem Gokmen, Gokul Dharan, Tanish Jain, et al. igibson 2.0: Object-centric simulation for robot learning of everyday household tasks. *arXiv preprint arXiv:2108.03272*, 2021. 1

[36] Chengshu Li, Ruohan Zhang, Josiah Wong, Cem Gokmen, Sanjana Srivastava, Roberto Martín-Martín, Chen Wang, Gabrael Levine, Michael Lingelbach, Jiankai Sun, et al. Behavior-1k: A benchmark for embodied ai with 1,000 everyday activities and realistic simulation. In *Conference on Robot Learning*, 2023. 2

[37] Xingyu Lin, Yufei Wang, Jake Olkin, and David Held. Softgym: Benchmarking deep reinforcement learning for deformable object manipulation. In *Conference on Robot Learning*, 2021. 2

[38] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv preprint arXiv:2211.01095*, 2022. 2, 3

[39] Mayank Mittal, Calvin Yu, Qinxi Yu, Jingzhou Liu, Nikita Rudin, David Hoeller, Jia Lin Yuan, Ritvik Singh, Yunrong Guo, Hammad Mazhar, et al. Orbit: A unified simulation framework for interactive robot learning environments. *IEEE Robotics and Automation Letters*, 2023. 2

[40] Tongzhou Mu, Zhan Ling, Fanbo Xiang, Derek Yang, Xuanlin Li, Stone Tao, Zhiao Huang, Zhiwei Jia, and Hao Su. Maniskill: Generalizable manipulation skill benchmark with large-scale demonstrations. *arXiv preprint arXiv:2107.14483*, 2021. 2

[41] Yinyu Nie, Angela Dai, Xiaoguang Han, and Matthias Nießner. Learning 3d scene priors with 2d supervision. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[42] Despoina Paschalidou, Amlan Kar, Maria Shugrina, Karsten Kreis, Andreas Geiger, and Sanja Fidler. Atiss: Autoregressive transformers for indoor scene synthesis. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2, 5, 6

[43] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 3

[44] Pulak Purkait, Christopher Zach, and Ian Reid. Sg-vae: Scene grammar variational autoencoder to generate new indoor scenes. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2020. 2

[45] Siyuan Qi, Yixin Zhu, Siyuan Huang, Chenfanfu Jiang, and Song-Chun Zhu. Human-centric indoor scene synthesis using stochastic grammar. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 2, 3

[46] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 3

[47] Ludan Ruan, Yiyang Ma, Huan Yang, Huiguo He, Bei Liu, Jianlong Fu, Nicholas Jing Yuan, Qin Jin, and Baining Guo. Mm-diffusion: Learning multi-modal diffusion models for joint audio and video generation. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[48] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Cliport: What and where pathways for robotic manipulation. In *Conference on Robot Learning*, 2022. 1

[49] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 3

[50] Andrew Szot, Alexander Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Singh Chaplot, Oleksandr Maksymets, et al. Habitat 2.0: Training home assistants to rearrange their habitat. *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 34:251–266, 2021. 1

[51] Jiapeng Tang, Yinyu Nie, Lev Markhasin, Angela Dai, Justus Thies, and Matthias Nießner. Diffuscene: Denoising diffusion models for gerative indoor scene synthesis. In *Proceed-*

ings of Conference on Computer Vision and Pattern Recognition (CVPR), 2024. 2, 3, 5, 6

[52] Jiashun Wang, Huazhe Xu, Jingwei Xu, Sifei Liu, and Xiaolong Wang. Synthesizing long-term 3d human motion and interaction in 3d scenes. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[53] Weiqi Wang, Zihang Zhao, Ziyuan Jiao, Yixin Zhu, Song-Chun Zhu, and Hangxin Liu. Rearrange indoor scenes for human-robot co-activity. In *Proceedings of International Conference on Robotics and Automation (ICRA)*, 2023. 3

[54] Xinpeng Wang, Chandan Yeshwanth, and Matthias Nießner. Sceneformer: Indoor scene generation with transformers. In *Proceedings of International Conference on 3D Vision (3DV)*, 2021. 1, 2

[55] Zan Wang, Yixin Chen, Baoxiong Jia, Puhao Li, Jinlu Zhang, Jingze Zhang, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. Move as you say, interact as you can: Language-guided human motion generation with scene affordance. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[56] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2023. 3

[57] Fanbo Xiang, Yuzhe Qin, Kaichun Mo, Yikuan Xia, Hao Zhu, Fangchen Liu, Minghua Liu, Hanxiao Jiang, Yifu Yuan, He Wang, et al. Sapien: A simulated part-based interactive environment. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1

[58] Kun Xu, Kang Chen, Hongbo Fu, Wei-Lun Sun, and Shi-Min Hu. Sketch2scene: Sketch-based co-retrieval and co-placement of 3d models. *ACM Transactions on Graphics (TOG)*, 32(4):1–15, 2013. 1, 2

[59] Ming-Jia Yang, Yu-Xiao Guo, Bin Zhou, and Xin Tong. Indoor scene generation from a collection of semantic-segmented depth images. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2021. 2

[60] Peiyu Yu, Sirui Xie, Xiaojian Ma, Baoxiong Jia, Bo Pang, Ruiqi Gao, Yixin Zhu, Song-Chun Zhu, and Ying Nian Wu. Latent diffusion energy-based model for interpretable text modeling. In *Proceedings of International Conference on Machine Learning (ICML)*, 2022. 3

[61] Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. Physdiff: Physics-guided human motion diffusion model. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2023. 3

[62] Guangyao Zhai, Evin Pınar Örnek, Shun-Cheng Wu, Yan Di, Federico Tombari, Nassir Navab, and Benjamin Busam. Commonscenes: Generating commonsense 3d indoor scenes with scene graphs. *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 1, 2

[63] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2023. 3

[64] Zaiwei Zhang, Zhenpei Yang, Chongyang Ma, Linjie Luo, Alexander Huth, Etienne Vouga, and Qixing Huang. Deep generative modeling for scene synthesis via hybrid representations. *ACM Transactions on Graphics (TOG)*, 39(2):1–21, 2020. 2

[65] Kaizhi Zheng, Xiaotong Chen, Odest Chadwicke Jenkins, and Xin Wang. Vlmbench: A compositional benchmark for vision-and-language manipulation. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2

[66] Dingfu Zhou, Jin Fang, Xibin Song, Chenye Guan, Junbo Yin, Yuchao Dai, and Ruigang Yang. Iou loss for 2d/3d object detection. In *Proceedings of International Conference on 3D Vision (3DV)*, 2019. 5

[67] Yang Zhou, Zachary While, and Evangelos Kalogerakis. Scenegraphnet: Neural message passing for 3d indoor scene augmentation. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2019. 1, 2