This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Diffusion-based Generation, Optimization, and Planning in 3D Scenes

Siyuan Huang[1*✉], Zan Wang[1,2*], Puhao Li[1,3], Baoxiong Jia[1]
Tengyu Liu[1], Yixin Zhu[4], Wei Liang[2,5✉], Song-Chun Zhu[1,3,4]

[1] National Key Laboratory of General Artificial Intelligence, BIGAI
[2] School of Computer Science & Technology, Beijing Institute of Technology
[3] Dept. of Automation, Tsinghua University [4] Institute for AI, Peking University
[5] Yangtze Delta Region Academy of Beijing Institute of Technology, Jiaxing

https://scenediffuser.github.io

Figure 1. **Illustration of the SceneDiffuser**, applicable to various scene-conditioned 3D tasks: (a) human pose generation, (b) human motion generation, (c) dexterous grasp generation, (d) path planning for 3D navigation with goals, and (e) motion planning for robot arms.

Abstract

We introduce the SceneDiffuser, a conditional generative model for 3D scene understanding. SceneDiffuser provides a unified model for solving scene-conditioned generation, optimization, and planning. In contrast to prior work, SceneDiffuser is intrinsically scene-aware, physics-based, and goal-oriented. With an iterative sampling strategy, SceneDiffuser jointly formulates the scene-aware generation, physics-based optimization, and goal-oriented planning via a diffusion-based denoising process in a fully differentiable fashion. Such a design alleviates the discrepancies among different modules and the posterior collapse of previous scene-conditioned generative models. We evaluate the SceneDiffuser on various 3D scene understanding tasks, including human pose and motion generation, dexterous grasp generation, path planning for 3D navigation, and motion planning for robot arms. The results show significant improvements compared with previous models, demonstrating the tremendous potential of the SceneDiffuser for the broad community of 3D scene understanding.

1. Introduction

The ability to generate, optimize, and plan in 3D scenes is a long-standing goal across computer vision, graphics, and robotics. Various tasks have been devised to achieve these goals, fostering downstream applications in motion generation [32, 69, 73, 90], motion planning [42, 43, 60, 75], grasp generation [25, 31, 34], navigation [1, 95], scene synthesis [24, 47, 72, 82], embodied perception and manipulation [30, 39, 61], and autonomous driving [3, 52].

Despite rich applications and great successes, existing models designed for these tasks exhibit two fundamental limitations for real-world 3D scene understanding.

* These authors contributed equally to this work.
✉ Corresponding authors: Siyuan Huang (syhuang@bigai.ai) and Wei Liang (liangwei@bit.edu.cn).

First, most prior work [8, 14, 32, 49, 60, 68–70, 73] leverages the conditional Variational Autoencoder (cVAE) for the conditional generation in 3D scenes. cVAE model utilizes an encoder-decoder structure to learn the posterior distribution and relies on the learned latent variables to sample. Although cVAE is easy to train and sample due to its simple architecture and one-step sampling procedure, it suffers from the **posterior collapse problem** [12, 17, 26, 64, 69, 84, 93]; the learned latent variable is ignored by a strong decoder, leading to limited generation diversity from these collapsed modes. Such collapse is further magnified in 3D tasks with stronger 3D decoders and more complex and noisy input conditions, *e.g*., the natural 3D scans [9].

Second, despite the close relations among generation, optimization, and planning in 3D scenes, there **lacks a unified framework** that could address existing discrepancies among these models. Previous work [15, 34, 69] applies off-the-shelf physics-based post-optimization methods over outputs of generative models and often produces inconsistent and implausible generations, especially when transferring to novel scenes. Similarly, planners are usually standalone modules over results of generative model [8, 14] for trajectory planning or learned separately with the reinforcement learning (RL) [95], leading to gaps between planning and other modules (*e.g*., generation) during inference, especially in novel scenes where explorations are limited.

To tackle the above limitations, we introduce the SceneDiffuser, a conditional generative model based on the diffusion process. SceneDiffuser eliminates the discrepancies and provides a single home for scene-conditioned generation, optimization, and planning. Specifically, with a denoising process, it learns a diffusion model for scene-conditioned generation during training. In inference, SceneDiffuser jointly solves the scene-aware generation, physics-based optimization, and goal-oriented planning through a unified iterative guided-sampling framework. Such a design equips the SceneDiffuser with the following three superiority:

1. **Generation:** Building upon the diffusion model, SceneDiffuser solves the posterior collapse problem of scene-conditioned generative models. Since the forward diffusion process can be treated as data augmentation in 3D scenes, it helps traverse sufficient scene-conditioned distribution modes.
2. **Optimization:** SceneDiffuser integrates the physics-based objective into each step of the sampling process as conditional guidance, enabling the differentiable physics-based optimization during both the learning and sampling process. This design facilitates the physically-plausible generation, which is critical for tasks in 3D scenes.
3. **Planning:** Based on the scene-conditioned trajectory-level generator, SceneDiffuser possesses a physics- and goal-aware trajectory planner, generalizing better to long-horizon trajectories and novel 3D scenes.

As illustrated in Fig. 1, we evaluate the SceneDiffuser on diverse 3D scene understanding tasks. The human pose, motion, and dexterous grasp generation results significantly improve, demonstrating plausible and diverse generations with the 3D scene and object conditions. The results on path planning for 3D navigation and motion planning for robot arms reveal the generalizable and long-horizon planning capability of the SceneDiffuser.

2. Related Work

Conditional Generation in 3D Scenes Generating diverse contents and rich interactions in 3D scenes is essential for understanding the 3D scene affordances. Several applications have emerged on conditional scene generation [24, 47, 71, 81, 88], human pose [16, 32, 87, 90, 92] and motion generation [14, 32, 49, 60, 68–70, 73] in furnished 3D indoor scenes, and object-conditioned grasp pose generation [25, 31, 34, 62, 78]. However, most previous methods [6, 14, 16, 25, 31, 62, 68, 73, 75] rely on cVAE and suffer from the posterior collapse problem [12, 17, 26, 64, 69, 84, 93], especially when the 3D scene is natural and complex. In this work, SceneDiffuser addresses the posterior collapse with the diffusion-based denoising process.

Physics-based Optimization in 3D Scenes Producing physically plausible generations compatible with 3D scenes is challenging in the scene-conditioned generation. Previous work uses physics-based post-optimization [15, 34, 69] or differentiable objective [25, 73, 90] to integrate collision and contact constraints into the generation framework. However, post-optimization approaches [15, 34, 69] cannot be learned jointly with the generative models, yielding inconsistent generation results. Similarly, differentiable approaches [25, 73, 90] impose constraints on the final objective and cannot optimize the physical interactions during sampling, resulting in implausible generations, particularly when adapting to novel scenes. In this work, SceneDiffuser avoids such inconsistency by integrating differentiable physics-based optimization into each step of the sampling process.

Planning in 3D Scenes The ability to act and plan in 3D scenes is critical for an intelligent agent and has led to the recent culmination of embodied AI research [28, 30, 33, 36, 54, 55, 79]. Among all tasks, visual navigation has been most studied in the vision and robotics community [4, 13, 21, 40, 76, 77, 94, 95]. However, existing work relies heavily on model-based planners with the single-step dynamic model [5, 11, 48, 67, 74, 80], lacking a trajectory-level optimization for long-horizon planning. Further, these planners lack explicit modeling of physical interactions, making it difficult to generalize to natural scenes with limited exploration, where rapid learning and adaptation are required. Compared with the global trajectory planner based on a trajectory-level generator, SceneDiffuser exhibits superior generalization in long-horizon plans and novel 3D scenes.

Diffusion-based Models Diffusion model [19, 22, 57, 59] has come forth into a promising class of generative model for learning and sampling data distributions with an iterative denoising process, facilitating the image [10, 58], text [83], and shape generation [35]. With flexible conditioning, it is further extended to the language-conditioned image [41, 51, 53], video [18, 56], and 3D generation [44, 63, 85]. Notably, Janner *et al*. [23] integrate the generation and planning into the same sampling framework for behavior synthesis. To our best knowledge, SceneDiffuser is the first framework that models the 3D scene-conditioned generation with a diffusion model and integrates the generation, optimization, and planning into a unified framework.

3. Background

3.1. Problem Definition

Given a 3D scene $\mathcal{S}$, we aim to generate the optimal solution for completing the tasks (*e.g*., navigation, manipulation) given the goal $\mathcal{G}$ in the scene. We denote the state and action of an agent as $(\mathbf{s}, \mathbf{a})$. The dynamic model defines the state transition as $p(\mathbf{s}_{i+1}|\mathbf{s}_i, \mathbf{a}_i)$, which is often deterministic in scene understanding (*i.e*., $f(\mathbf{s}_i, \mathbf{a}_i)$). The trajectory is defined as $\boldsymbol{\tau} = (\mathbf{s}_0, \mathbf{a}_0, \cdots, \mathbf{s}_i, \mathbf{a}_i, \cdots, \mathbf{s}_N)$, where N denotes the horizon of task solving in discrete time.

3.2. Planning with Trajectory Optimization

The scene-conditional trajectory optimization is defined as maximizing the task objective:

$$\boldsymbol{\tau}^* = \arg\max_{\boldsymbol{\tau}} \mathcal{J}(\boldsymbol{\tau}|\mathcal{S}, \mathcal{G}). \tag{1}$$

The dynamic model is usually known for trajectory optimization. Considering the future actions and states with predictable dynamics, the entire trajectory $\boldsymbol{\tau}$ can be optimized jointly and non-progressively with traditional [29] or data-driven [7] planning algorithms. Trajectory-based optimization benefits from its global awareness of history and future states, thus can better model the long-horizon tasks compared with single-step models in RL, where $\mathbf{a}^*_{0:N} = \arg\max_{\mathbf{a}_{0:N}} \sum_{i=0}^{N} r(\mathbf{s}_i, \mathbf{a}_i|\mathcal{S}, \mathcal{G})$.

3.3. Diffusion Model

Diffusion model [19, 22, 57] is a class of generative models representing the data generation with an iterative denoising process from Gaussian noise. It consists of a forward and a reverse process. The forward process $q(\boldsymbol{\tau}^t|\boldsymbol{\tau}^{t-1})$ gradually destroys data $\boldsymbol{\tau}^0 \sim q(\boldsymbol{\tau}^0)$ into Gaussian noise. The parametrized reverse process $p_\theta(\boldsymbol{\tau}^{t-1}|\boldsymbol{\tau}^t)$ recovers the data from noise with the learned normal distribution from a fixed timestep. The training objective for θ is denoising score matching over multiple noise scale [22, 66]. Please refer to the Appendix A for details of diffusion model and variants.

4. SceneDiffuser

SceneDiffuser models planning as trajectory optimization and solves the aforementioned problem with the spirit of *planning as sampling*, where the trajectory optimization is achieved by sampling trajectory-level distribution learned by the model. Leveraging the diffusion model with gradient-based sampling and flexible conditioning, SceneDiffuser models the scene-conditioned goal-oriented trajectory $p(\boldsymbol{\tau}^0|\mathcal{S}, \mathcal{G})$:

$$\begin{aligned} p(\boldsymbol{\tau}^0|\mathcal{S}, \mathcal{G}) &= \frac{p_\theta(\boldsymbol{\tau}^0|\mathcal{S}) p_\phi(\mathcal{G}|\boldsymbol{\tau}^0, \mathcal{S})}{p(\mathcal{G}|\mathcal{S})} \\ &\propto p_\theta(\boldsymbol{\tau}^0|\mathcal{S}) p_\phi(\mathcal{G}|\boldsymbol{\tau}^0, \mathcal{S}). \end{aligned} \tag{2}$$

Generation $p_\theta(\boldsymbol{\tau}^0|\mathcal{S})$ characterizes the probability of generating certain trajectories with the scene condition. It can be modeled using a conditional diffusion model [19, 57] with an iterative denoising process:

$$\begin{aligned} p_\theta(\boldsymbol{\tau}^0|\mathcal{S}) &= p(\boldsymbol{\tau}^T) \prod_{t=1}^{T} p(\boldsymbol{\tau}^{t-1}|\boldsymbol{\tau}^t, \mathcal{S}), \\ p(\boldsymbol{\tau}^{t-1}|\boldsymbol{\tau}^t, \mathcal{S}) &= \mathcal{N}(\boldsymbol{\tau}^{t-1}; \boldsymbol{\mu}_\theta(\boldsymbol{\tau}^t, t, \mathcal{S}), \boldsymbol{\Sigma}_\theta(\boldsymbol{\tau}^t, t, \mathcal{S})). \end{aligned} \tag{3}$$

Optimization and Planning $p_\phi(\mathcal{G}|\boldsymbol{\tau}^0, \mathcal{S})$ represents the probability of reaching the goal with the sampled trajectory, where the goal can be flexibly defined by customized objective functions in various tasks. As shown in Eq. (4), the precise definition of this probability is $p_\phi(\mathcal{O} = 1|\boldsymbol{\tau}^0, \mathcal{S}, \mathcal{G})$, where O is an optimality indicator that represents if the goal were achieved. Intuitively, the trajectory objective in Eq. (1) can indicate such optimality. We therefore expand $p_\phi(\mathcal{G}|\boldsymbol{\tau}^t, \mathcal{S})$ as its exponential in Eq. (5):

$$p_\phi(\mathcal{G}|\boldsymbol{\tau}^t, \mathcal{S}) = p_\phi(\mathcal{O} = 1|\boldsymbol{\tau}^t, \mathcal{S}, \mathcal{G}) \tag{4}$$
$$\propto \exp(\mathcal{J}(\boldsymbol{\tau}^t|\mathcal{S}, \mathcal{G})) \tag{5}$$
$$= \exp(\varphi_p(\boldsymbol{\tau}^t|\mathcal{S}, \mathcal{G}) + \varphi_o(\boldsymbol{\tau}^t|\mathcal{S}), \tag{6}$$

where $\varphi_o(\boldsymbol{\tau}^t|\mathcal{S})$ denotes the objective for optimizing the trajectory with scene condition and is independent of task goal $\mathcal{G}$. In scene understanding, φ_o usually denotes plausible physical relationships (*e.g*., collision, contact, and intersection). $\varphi_p(\boldsymbol{\tau}^t|\mathcal{S}, \mathcal{G})$ indicates the objective for planning (*i.e*., goal-reaching) with scene condition. Both φ_o and φ_p can be explicitly defined or implicitly learned from observed trajectories with proper parametrization.

4.1. Learning

$p_\theta(\boldsymbol{\tau}^0|\mathcal{S})$ is the scene-conditioned generator, which can be learned by the conditional diffusion model with the simplified objective of estimating the noise $\boldsymbol{\epsilon}$ [10, 19, 20], where

$$\begin{aligned} \mathcal{L}_\theta(\boldsymbol{\tau}^0|\mathcal{S}) &= \mathbb{E}_{t,\boldsymbol{\epsilon},\boldsymbol{\tau}^0}\left[\left\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\hat{\alpha}^t}\boldsymbol{\tau}^0 + \sqrt{1-\hat{\alpha}^t}\boldsymbol{\epsilon}, t, \mathcal{S})\right\|^2\right] \\ &= \mathbb{E}_{t,\boldsymbol{\epsilon},\boldsymbol{\tau}^0}\left[\left\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\boldsymbol{\tau}^t, t, \mathcal{S})\right\|^2\right], \end{aligned} \tag{7}$$

where $\hat{\alpha}^t$ is the pre-determined function in the forward process. With the learned $p_\theta(\boldsymbol{\tau}^0|\mathcal{S})$, we sample $p(\boldsymbol{\tau}^0|\mathcal{S},\mathcal{G})$ by taking the advantage of the diffusion model's flexible conditioning [10, 23]. Specifically, we approximate $p_\phi(\mathcal{G}|\boldsymbol{\tau}^t,\mathcal{S})$ using the Taylor expansion around $\boldsymbol{\tau}^t=\boldsymbol{\mu}$ at timestep t as

$$\log p_\phi(\mathcal{G}|\boldsymbol{\tau}^t,\mathcal{S}) \approx (\boldsymbol{\tau}^t-\boldsymbol{\mu})\mathbf{g}+C, \tag{8}$$

where C is a constant, $\boldsymbol{\mu}=\boldsymbol{\mu}_\theta(\boldsymbol{\tau}^t,t,\mathcal{S})$ and $\boldsymbol{\Sigma}=\boldsymbol{\Sigma}_\theta(\boldsymbol{\tau}^t,t,\mathcal{S})$ are the inferred parameters of original diffusion process, and

$$\begin{aligned}\mathbf{g} &= \nabla_{\boldsymbol{\tau}^t}\log p_\phi(\mathcal{G}|\boldsymbol{\tau}^t,\mathcal{S})|_{\boldsymbol{\tau}^t=\boldsymbol{\mu}}\\ &= \nabla_{\boldsymbol{\tau}^t}(\varphi_o(\boldsymbol{\tau}^t|\mathcal{S})+\varphi_p(\boldsymbol{\tau}^t|\mathcal{S},\mathcal{G}))|_{\boldsymbol{\tau}^t=\boldsymbol{\mu}}.\end{aligned} \tag{9}$$

Therefore, we have

$$p(\boldsymbol{\tau}^{t-1}|\boldsymbol{\tau}^t,\mathcal{S},\mathcal{G})=\mathcal{N}(\boldsymbol{\tau}^{t-1};\boldsymbol{\mu}+\lambda\boldsymbol{\Sigma}\mathbf{g},\boldsymbol{\Sigma}), \tag{10}$$

where λ is the scaling factor for the guidance. With Eq. (10), we can sample $\boldsymbol{\tau}^t$ with the guidance of optimizing and planning objectives.

Of note, φ_p and φ_o serve as the pre-defined guidance for tilting the original trajectory with physical and goal constraints. However, they can also be learned from the observed trajectories. During training, we first fix the learned base model of $p_\theta(\boldsymbol{\tau}^0|\mathcal{S})$, followed by learning ϕ_o and ϕ_p for optimization and planning with the following objective:

$$\mathcal{L}_\phi(\boldsymbol{\tau}^0|\mathcal{S},\mathcal{G})=\mathbb{E}_{t,\boldsymbol{\epsilon},\boldsymbol{\tau}^0}\left[\|\boldsymbol{\epsilon}-\boldsymbol{\epsilon}_\theta(\boldsymbol{\tau}^t,t,\mathcal{S})-\boldsymbol{\Sigma}\mathbf{g}\|^2\right]. \tag{11}$$

Alg. 1 summarizes the training procedure.

Algorithm 1: Training SceneDiffuser

`// train base generation model`
Input: Trajectory in 3D scene $(\boldsymbol{\tau}^0,\mathcal{S})$
repeat
  $\boldsymbol{\tau}^0 \sim p(\boldsymbol{\tau}^0|\mathcal{S})$
  $\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I})$, $t\sim\mathcal{U}(\{1,\cdots,T\})$
  $\boldsymbol{\tau}^t=\sqrt{\hat{\alpha}_t}\boldsymbol{\tau}_0+\sqrt{1-\hat{\alpha}_t}\boldsymbol{\epsilon}$
  $\theta=\theta-\eta\nabla_\theta\|\boldsymbol{\epsilon}-\boldsymbol{\epsilon}_\theta(\boldsymbol{\tau}^t,t,\mathcal{S})\|_2^2$
until *converged*;
`// (optional) train optimization and planning model`
Input: Trajectory in 3D scene with goal $(\boldsymbol{\tau}^0,\mathcal{S},\mathcal{G})$, learned θ for $p_\theta(\boldsymbol{\tau}^0|\mathcal{S})$
repeat
  $\boldsymbol{\tau}^0\sim p(\boldsymbol{\tau}^0|\mathcal{S},\mathcal{G})$
  $\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I})$, $t\sim\mathcal{U}(\{1,\cdots,T\})$
  $\boldsymbol{\mu}=\boldsymbol{\mu}_\theta(\boldsymbol{\tau}^t,t,\mathcal{S})$, $\boldsymbol{\Sigma}=\boldsymbol{\Sigma}_\theta(\boldsymbol{\tau}^t,t,\mathcal{S})$
  $\mathbf{g}=\nabla_{\boldsymbol{\tau}^t}\log p_\phi(\mathcal{G}|\boldsymbol{\tau}^t,\mathcal{S})|_{\boldsymbol{\tau}^t=\boldsymbol{\mu}}$
  $\boldsymbol{\tau}^t=\sqrt{\hat{\alpha}_t}\boldsymbol{\tau}_0+\sqrt{1-\hat{\alpha}_t}\boldsymbol{\epsilon}$
  $\phi=\phi-\eta\nabla_\phi\|\boldsymbol{\epsilon}-\boldsymbol{\epsilon}_\theta(\boldsymbol{\tau}^t,t,\mathcal{S})-\lambda\boldsymbol{\Sigma}\mathbf{g}\|_2^2$
until *converged*;

4.2. Sampling

With different sampling strategies, SceneDiffuser can generate, optimize, and plan the trajectory in 3D scenes under a unified framework of guided sampling. Alg. 2 summarizes the detailed sampling algorithm.

Algorithm 2: Sampling SceneDiffuser for generation, optimization, and planning

Modules: Model $p_\theta(\cdot|\mathcal{S})$, optimization objective $\varphi_o(\cdot|\mathcal{S})$, and planner objective $\varphi_p(\cdot|\mathcal{S},\mathcal{G})$
`// one-step guided sampling`
function sample $(\boldsymbol{\tau}^t,\mathcal{J})$:
  $\boldsymbol{\mu}=\boldsymbol{\mu}_\theta(\boldsymbol{\tau}^t,t,\mathcal{S})$, $\boldsymbol{\Sigma}=\boldsymbol{\Sigma}_\theta(\boldsymbol{\tau}^t,t,\mathcal{S})$
  $\boldsymbol{\tau}^{t-1}=\mathcal{N}(\boldsymbol{\tau}^{t-1};\boldsymbol{\mu}+\lambda\boldsymbol{\Sigma}\nabla_{\boldsymbol{\tau}^t}(\mathcal{J}(\boldsymbol{\tau}^t|\mathcal{S},\mathcal{G}))|_{\boldsymbol{\tau}^t=\boldsymbol{\mu}},\boldsymbol{\Sigma})$
  return $\boldsymbol{\tau}^{t-1}$
`// physics-based generation`
Input: initial trajectory $\boldsymbol{\tau}^T\sim\mathcal{N}(\mathbf{0},\mathbf{I})$
for $t=T,\cdots,1$ **do**
  `// sampling with optimization`
  $\boldsymbol{\tau}^{t-1}=$ **sample**$(\boldsymbol{\tau}^t,\varphi_o(\cdot|\mathcal{S}))$
return $\boldsymbol{\tau}^0$
`// goal-oriented planning`
Input: planning steps N, starting state $\hat{\mathbf{s}}_0$, initial plan $\boldsymbol{\tau}_0^T\sim\mathcal{N}(\mathbf{0},\mathbf{I})$
$i=1$
while *not done and planning step* $i<N$ **do**
  for $t=T,\cdots,1$ **do**
    $\boldsymbol{\tau}_i^{t-1}=$ **sample**$(\boldsymbol{\tau}_i^t,\varphi_o(\cdot|\mathcal{S})+\varphi_p(\cdot|\mathcal{S},\mathcal{G}))$
    `// planning as inpainting`
    $\boldsymbol{\tau}_i^{t-1}[0:i]=\hat{\mathbf{s}}_{0:i}$
  Act $\hat{a}_{i-1}$ to reach $\hat{\mathbf{s}}_i=\boldsymbol{\tau}_i^0[i]$, $\hat{\mathbf{s}}_{0:i}=\hat{\mathbf{s}}_{0:i-1}\cup\hat{\mathbf{s}}_i$
  Increment planning step $i=i+1$

Scene-aware Generation Sampling $\boldsymbol{\tau}^0$ from the distribution $p_\theta(\boldsymbol{\tau}^0|\mathcal{S})$ in Eq. (3) directly solves the conditional generation tasks. The sampled trajectories represent diverse modes and possible interactions with the 3D scenes.

Physics-based Optimization In a differentiable manner, the physical relations between each state and the environment are defined by φ_o in Eq. (4). For general optimization without the planning objective, the task goal $\mathcal{G}$ is to sample a plausible trajectory in 3D scenes. Therefore, we can draw physically plausible trajectories in 3D scenes by sampling from $p(\boldsymbol{\tau}^0|\mathcal{S},\mathcal{G})$ with Eq. (10).

Goal-oriented Planning This can be formulated as motion inpainting under the sampling framework. Given the start state $\hat{\mathbf{s}}_s$ and the goal state $\hat{\mathbf{s}}_g$, the planning module returns trajectory $\hat{\boldsymbol{\tau}}=(\hat{\mathbf{s}}_0,\hat{\mathbf{a}}_0,\cdots,\hat{\mathbf{s}}_i,\hat{\mathbf{a}}_i,\cdots,\hat{\mathbf{s}}_g)$ that can reach the goal state. We set the first state as $\hat{\mathbf{s}}_0=\hat{\mathbf{s}}_s$ and define the goal state and reward of goal-reaching in φ_p. For each step i, we first keep the previous states and inpaint the remaining trajectory by sampling the goal-oriented SceneDiffuser with an iterative denoising process. Next, we take the action that can reach the next sampled state with $(\hat{\mathbf{a}}_{i-1},\hat{\mathbf{s}}_i)$. As illustrated in Alg. 2, we repeat the planning steps until the goal or maximal planning step is reached. Our planner leverages the trajectory-level generator, thus more generalizable to long-horizon trajectories and novel scenes.

4.3. Implementation

Model Architecture The design of SceneDiffuser follows the practices of conditional diffusion model [20, 51, 53]. Specifically, we augment the time-conditional diffusion

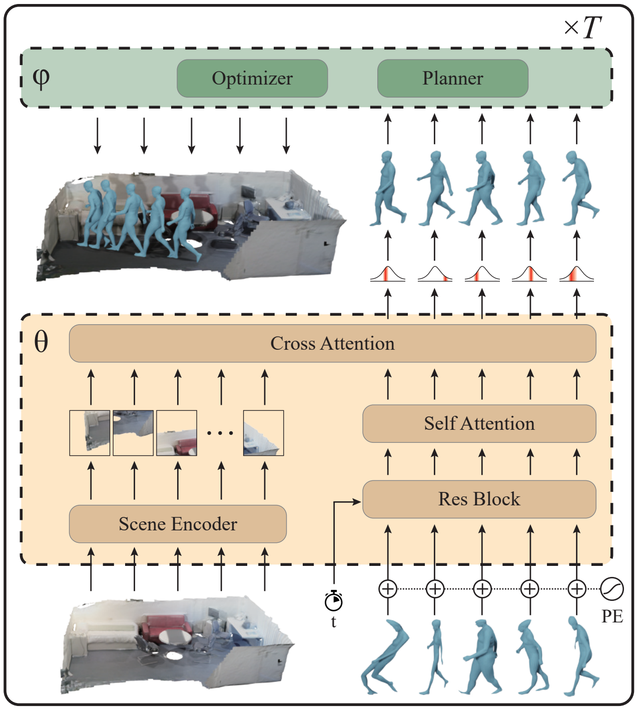


Figure 2. **Model architecture of the SceneDiffuser.** We use cross-attention to learn the relation between the input trajectory and scene condition. The optimizer and planner serve as the guidance for physically-plausible and goal-oriented trajectories.

model with cross-attention [65] for flexible conditioning. As shown in Fig. 2, for each sampling step, the model computes the cross-attention between the 3D scene condition and input trajectory, wherein the key and value are learned from the condition, and the query is learned from the input trajectory. The computed vector is fed into a feed-forward layer to estimate the noise $\boldsymbol{\epsilon}$. The 3D scene is processed by a scene encoder [45, 91]); please refer to Appendix B.

Objective Design We consider two types of trajectory objectives for optimization and planning: (i) trajectory-level objective, and (ii) the accumulation of step-wise objective. For optimization, we consider step-wise collision, contact objective, and trajectory level smoothness objective, $\{\varphi_o^{\text{collision}}, \varphi_o^{\text{contact}}, \varphi_o^{\text{smoothness}}\}$. For planning, we consider the accumulation of simple step-wise distance, $\varphi_p^{L_2}$. Please refer to the Appendix C for details. Empirically, we observe that parameterizing the objectives with timestep t and increasing the guidance during the last several diffusion steps will enhance the effect of guidance.

5. Experiments

To demonstrate SceneDiffuser is general and applicable to various scenarios, we evaluate the SceneDiffuser on five scene understanding tasks. For generation, we evaluate the scene-conditioned human pose and motion generation and object-conditioned dexterous grasp generation. For planning, we evaluate the path planning for 3D navigation and motion planning for robot arms. Below we first introduce baseline methods, followed by detailed settings, results analyses, and ablative studies for each task. We provide detailed implementation and experimental settings in Appendix D, additional ablative studies in Appendix E, additional experiments in Appendix F, and additional qualitative results in Appendix G.

Baseline Methods For conditional generation tasks, we primarily compare SceneDiffuser with the widely-adopted cVAE model [25, 31, 32, 73, 90] and its variants. We also compare with strategies for optimizing the physics of the trajectory in the cVAE, including integrating into training as loss and plugging upon as the post-optimization. For planning, we compare with a stochastic planner learned by imitation learning using Behavior Cloning (BC) and a simple heuristic-based deterministic planner guided by L_2 distance.

5.1. Task 1: Human Pose Generation

Setup Scene-conditioned human pose generation aims to generate semantically plausible and physically feasible single-frame human bodies within the given 3D scenes. We evaluate the task on the 12 indoor scenes provided by PROX [15] and the refined version of PROX'S per-frame SMPL-X parameters from LEMO [86]. The input is the colored point cloud extracted by randomly downsampling the scene meshes provided in PROX. Training/testing splits are created following the literature [69, 90], resulting in $\sim$ 53k frames in 8 scenes for training and others for testing.

Metrics We evaluate the physical plausibility of generated poses with both direct human evaluations and indirect collision and contact scores. We randomly selected 1000 frames in the four test scenes for the direct measure and instructed 26 Turkers to decide whether the generated human pose was plausible. We compute the mean percentage of plausible generation and term this metric as the plausible rate. For indirect measures, we report (i) the non-collision score of the generated human bodies by calculating the proportion of the scene vertices with positive SDF to the human body and (ii) the contact score by checking if the body contact with the scene in a distance [15] below a pre-defined threshold. Following the literature [84, 90], we evaluate the diversities of global translation, generated SMPL-X parameters, and the marker-based body-mesh representation [89]. Specifically, we calculate the diversity of generated pose with the Average Pairwise Distance (APD) and standard deviation (std).

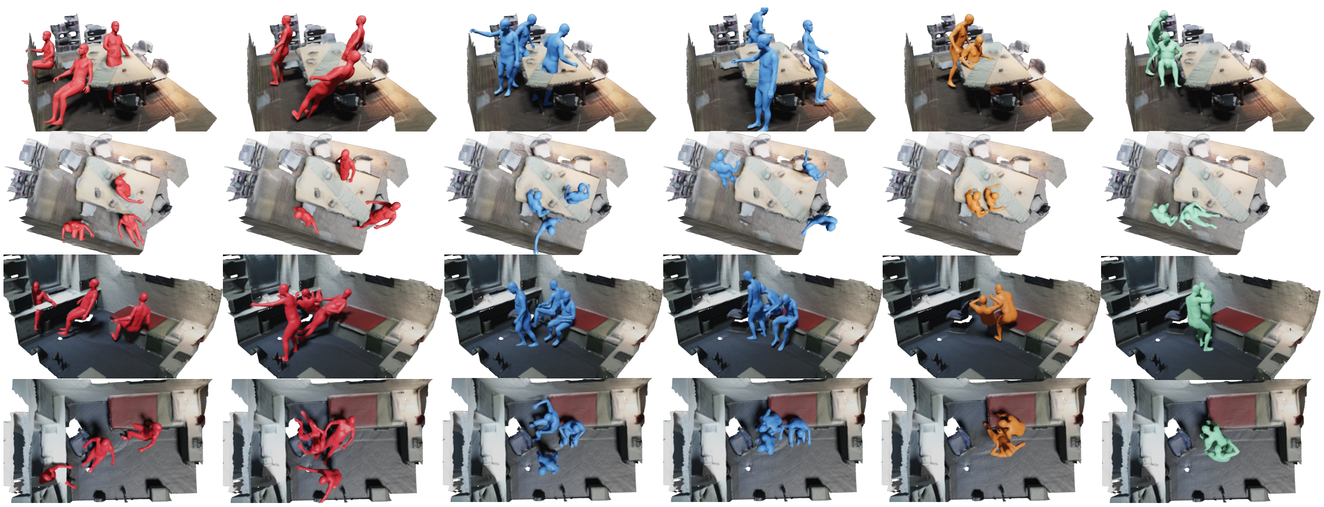
Results Tab. 1 quantitatively demonstrates that SceneDiffuser generates significantly better poses while maintaining generation diversity. We further provide qualitative comparisons between baseline models and SceneDiffuser in Fig. 3. While achieving a comparable diversity, collision, and contact performance, our model generates results that contain considerably more physically plausible poses (*e.g.*, floating, severe collision). This is reflected by the significant superiority (*i.e.*, over 28%) over

Table 1. **Quantitative results of human pose generation in 3D scenes.** We report metrics for physical plausibility and diversity.

model	plausible rate↑	non-collision score↑	contact score↑	APD (trans.)↑	std (trans.)↑	APD (param)↑	std (param)↑	APD (marker)↑	std (marker)↑
cVAE (w/o. L_{HS}) [90]	14.88	99.78	96.42	**1.218**	**0.494**	2.878	0.166	3.638	0.172
cVAE (w/ L_{HS}) [90]	16.19	99.75	99.25	1.013	0.416	2.994	0.170	3.614	0.169
our (w/o opt.)	21.04	99.74	**99.43**	0.776	0.331	3.204	0.195	3.483	0.167
our (w/ opt.)	**44.77**	**99.93**	98.05	1.009	0.413	**3.297**	**0.197**	**3.679**	**0.177**

Table 2. **Quantitative results of human motion generation in 3D scenes.** We report model variants with and without the start pose.

model	plausible rate↑	non-collision score↑	contact score↑	APD (trans.)↑	std (trans.)↑	APD (param)↑	std (param)↑	APD (marker)↑	std (marker)↑
cVAE (w/o start) [73]	7.72	99.86	86.26	**1.628**	**0.613**	**2.766**	**0.155**	3.275	0.150
ours (w/o start)	21.67	99.71	97.92	0.568	0.237	2.339	0.126	3.299	0.151
ours (w/o start & w/ opt.)	**25.83**	**99.93**	**98.91**	0.473	0.196	2.405	0.127	**3.385**	**0.154**
cVAE (w/ start) [73]	17.65	99.88	95.44	**0.478**	**0.188**	**1.747**	**0.091**	**2.308**	**0.105**
ours (w/ start)	36.56	99.85	**98.47**	0.193	0.081	1.372	0.065	1.568	0.072
ours (w/ start & w/ opt.)	**38.44**	**99.94**	98.43	0.168	0.070	1.389	0.065	1.575	0.072



(a) cVAE (b) SceneDiffuser (c) Without Opt. (d) With Opt.

Figure 3. **Qualitative results of human pose generation in 3D scenes.** From left to right: (a) cVAE generation, (b) SceneDiffuser generation without optimization, and poses generated (c) with and (d) without applying our optimization-guided sampling.

cVAE-based baselines on plausible rates. We observe this large improvement both quantitatively from the plausible rate and non-collision score and qualitatively in Fig. 3. Notably, our optimization-guided sampling improves the generator with 23% on the plausible rate, showing the efficacy of the proposed optimization-guided sampling strategy and its potential for a broader range of 3D tasks with physic-based constraints or objectives.

5.2. Task 2: Human Motion Generation

Setup We consider generating human motion sequences under two different settings: (1) condition solely on the 3D scene, and (2) condition on both the starting pose and the 3D scene. We use the same human and scene representation as in Sec. 5.1 and clip the original LEMO motion sequence into segments with a fixed duration (60 frames). In total, we obtain 28k motion segments, with the distance between each start and end pose being longer than 0.2 meters. We follow the same split in Sec. 5.1 for training/testing and the same evaluation metrics for the pose generation. We report the average values of pose metrics over motion sequence as our performance measure.

Results As quantitatively shown in Tab. 2, SceneDiffuser consistently generates high-quality motion sequences compared to cVAE baselines. Specifically, our generated motion outperforms baseline models on plausible rate, non-collision score, and contact score. This performance gain indicates better coverage of motion that involves rich interaction with the scene while remaining physically plausible. It also causes lower diversity in metrics (*e.g.*, translation variance) since the plausible space for the motion is limited compared with cVAE. Empirically, we observe that providing the start position of motion as a condition constrains possible future motion sequences and leads to a drop in generation diversity for all models. We also note only a marginal performance improvement after applying optimization-guided sampling. One potential reason is that the generated motions are already plausible

Figure 4. **Human motions generated by SceneDiffuser.** Each row shows sampled human motions from the same start pose.

and receive small guidance from the optimization. As qualitatively shown in Fig. 4, SceneDiffuser generates diverse motions (*e.g.*, "sit," "walk") from the same start position in unseen 3D scenes.

5.3. Task 3: Dexterous Grasp Generation

Setup Dexterous grasp generation aims to generate diverse and stable grasping poses for the given object with a human-like dexterous hand. We use the Shadowhand subset of the MultiDex [31] dataset, which contains diverse dexterous grasping poses for 58 daily objects. We represent the pose of Shadowhand as $q := (t, R, \theta) \in \mathbb{R}^{33}$, where $t \in \mathbb{R}^3$ and $R \in \mathbb{R}^6$ denote the global translation and orientation respectively, and $\theta \in \mathbb{R}^{24}$ describes the rotation angles of the revolute joints. An object is represented by its point cloud $\mathcal{O} \in \mathbb{R}^{2048\times 3}$. We split the dataset into 48 seen objects and 10 unseen objects for training and testing, respectively.

Metrics We evaluate models in terms of success rate, diversity, and collision depth. We test if a grasp is successful in IsaacGym [37] by applying external forces to the object and measuring the movement of the object. To measure how learned models capture the diversity of successful grasping pose in the training data, we report the success rate of generated poses that lies at different variance levels from the mean pose of training data. We measure the collision depth as the maximum depth that the hand penetrates the object in each successful grasp for testing models' performance on physically correct grasps. In all cases, we ignore the root transformation of the hand as it does not contribute to the diversity of grasping types.

Results Tab. 3 quantitatively demonstrates that SceneDiffuser generates significantly better grasp poses in terms of success rate while correctly balancing the diversity of generation and grasp success. This result indicates that the SceneDiffuser achieves a consistently high success rate without much performance drop when the generated pose diverges from the mean pose in the training data. We also show that, by applying optimizer upon SceneDiffuser, the guided sampling process can reduce the violation of physically implausible grasping poses, outperforming the state-of-the-art baseline [25] without additional training or intermediate representation (*i.e.*, contact maps). We provide qualitative results in Fig. 5 for visualization.

Table 3. **Quantitative results of dexterous grasp generation on MultiDex [31] dataset.** We measure the success rates under different diversities and depth collisions. TTA. denotes test-time optimization with physics and contact.

model	**succ. rate** (%)↑			**depth coll.** (mm)↓
	σ	2σ	all	
cVAE [25]	0.00	10.09	14.06	22.98
cVAE (w/ TTA.) [25]	0.00	21.91	17.97	15.19
ours (w/o opt.)	70.65	**71.25**	**71.25**	17.34
ours (w/ opt.)	**71.27**	69.84	69.84	**14.61**

5.4. Task 4: Path Planning for Navigation

Setup We selected 61 indoor scenes from ScanNet [9] to construct room-level scenarios for navigational path planning and annotated these scenes with navigation graphs. As shown in Fig. 6b, these annotations are more spatially dense and physically plausible compared to previous methods [1]. We represent the physical robot with a cylinder to simulate physically plausible trajectories; see Fig. 6a. In total, we collected around 6k trajectories by searching the shortest paths between the randomly selected start and target nodes on the graph. We use trajectories in 46 scenes for training and the rest 15 scenes for evaluation. Models take the input as the scene point cloud $\mathcal{S} \in \mathbb{R}^{32768\times 3}$, a given start position $\hat{\mathbf{s}}_0 \in \mathbb{R}^2$, and a target position $\mathcal{G} \in \mathbb{R}^2$ on the floor plane.

Metrics We evaluate the planned results by checking if the "robot" can move from the start to the target without collision along the planned trajectory. We report the average success rate and planning steps over all test cases.

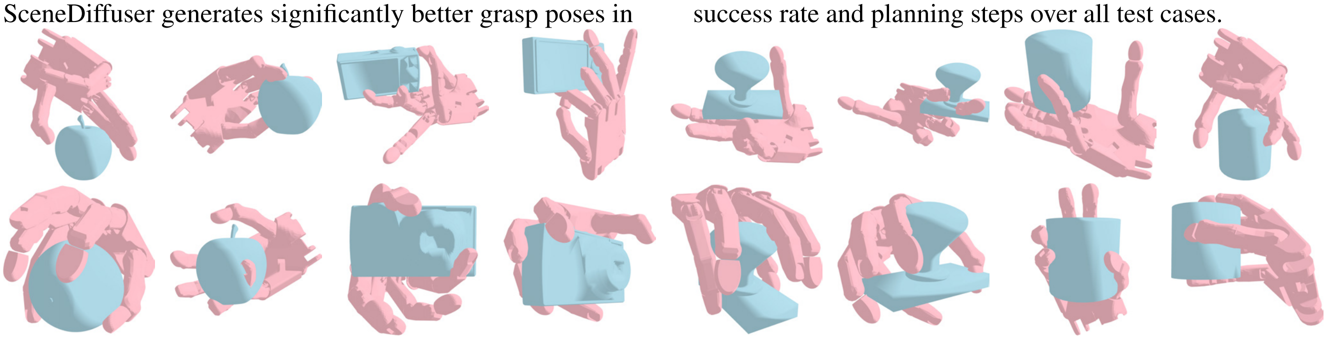
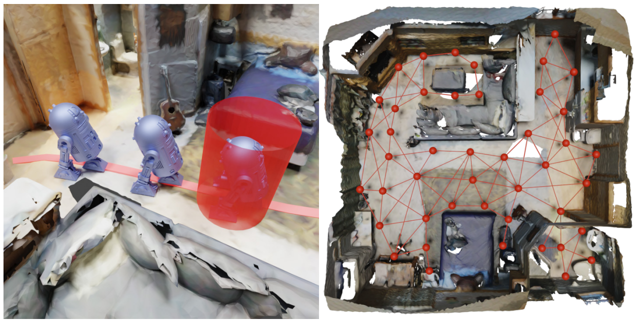


Figure 5. **Qualitative results of dexterous grasp generation.** Compared to grasps generated by cVAE (first row), SceneDiffuser (second row) generates fewer colliding or floating poses, which helps to achieve a higher success rate.



(a) path planning (b) navigation graph

Figure 6. **Path planning for 3D scene navigation.** SceneDiffuser generates trajectories in long-horizon tasks.

Results As shown in Tab. 4, SceneDiffuser outperforms both the BC and the deterministic planner baseline. These results indicate the efficacy of guided sampling with the planning objective, especially given that all test scenes are unseen during training. Crucially, as simple heuristics (like L_2) oftentimes lead to dead-ends in path planning, SceneDiffuser can correctly combine past knowledge on the scene-conditioned trajectory distribution and planning objective under specific unseen scenes to redirect planning direction, which helps to avoid obstacles and dead-ends to reach the goal successfully. Compared with the baseline models, our model also requires fewer planning steps while maintaining a higher success rate. This suggests that SceneDiffuser successfully navigates to the target without diverging even in long-horizon tasks, where classic RL-based stochastic planners suffer (*i.e.*, the low performance of BC).

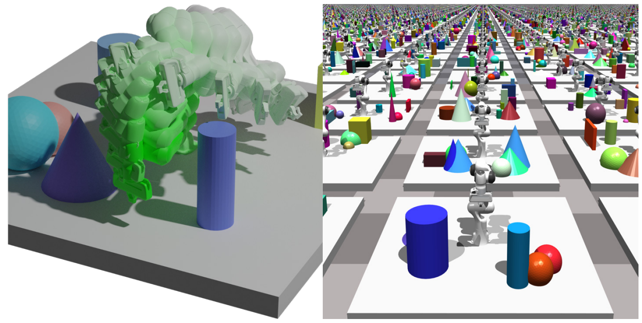
Table 4. **Quantitative results of path planning in 3D navigation and motion planning for robot arms.**

task	model	succ. rate(%)↑	planning steps↓
path plan	BC	0	150
	deterministic(L_2)	13.50	137.98
	ours	**73.75**	**90.38**
arm motion	BC	0.31	299.08
	deterministic(L_2)	72.87	**141.28**
	ours	**78.59**	147.60

5.5. Task 5: Motion Planning for Robot Arms

Setup Aiming to generate valid robot arm motion trajectories in cluttered scenes, we used the Franka Emika arm with seven revolute joints and collected 19,800 trajectories over 200 randomly generated cluttered scenes using the MoveIt 2.0 [50], as shown in Fig. 7. We represent the scene with point clouds $\mathcal{S} \in \mathbb{R}^{4096\times 3}$ and the robot arm trajectory with a sequence of joint angles $\mathcal{R} \in [-\pi, \pi]$. We train our model on 160 scenes and test on 40 unseen scenes.

Metrics Similar to Sec. 5.4, we evaluate the generated trajectories by success rate on unseen scenes and the average number of planning steps. We consider a trajectory successful if the robot arm reaches the goal pose by a certain distance threshold within a limited number of steps.



(a) motion planning (b) cluttered scenes

Figure 7. **Motion planning for robot arms.** SceneDiffuser generates arm motions in tabletop scenes with obstacles.

Results We observe similar overall performance as in Sec. 5.4. Tab. 4 shows that SceneDiffuser consistently outperforms both the RL-based BC and the deterministic planner baseline. SceneDiffuser's planning steps for successful trials are also comparable with the deterministic planner, showing the efficacy of the planner in long-horizon scenarios.

5.6. Ablation Analyses

We explore how the scaling coefficient λ influences the human pose generation results. We report the diversity and physics metrics of sampling results under different λs, ranging from 0.1 to 100. As shown in Tab. 5, λ balances generation collision/contact and diversity in human pose generation. Specifically, $\lambda = 1.0$ leads to the best physical plausibility, while larger λ values lead to diverse generation results. We attribute this effect to the optimization as with bigger λs; the optimization will draw poses away from the scene.

Table 5. **Ablation of the scale coefficient for optimization.**

metric	$\lambda = 0.1$	$\lambda = 1.0$	$\lambda = 10.0$	$\lambda = 100.0$
plausible rate ↑	28.75	**52.5**	21.25	0
APD (trans.) ↑	0.764	0.886	1.564	**23.96**
APD (param) ↑	3.206	3.243	9.040	**573.6**
non-collison score ↑	99.76	**99.87**	99.85	74.9
contact score ↑	**99.70**	99.65	81.75	0.0

6. Conclusion

We propose the SceneDiffuser as a general conditional generative model for generation, optimization, and planning in 3D scenes. SceneDiffuser is designed with appealing properties, including scene-aware, physics-based, and goal-oriented. We demonstrate that the SceneDiffuser outperforms previous models by a large margin on various tasks, establishing its efficacy and flexibility.

Acknowledgments and disclosure of funding We thank Ruiqi Gao and Ying Nian Wu for their helpful discussions and suggestions. This work is supported in part by the National Key R&D Program of China (2021ZD0150200), the Beijing Nova Program, and the National Natural Science Foundation of China (NSFC) (62172043).

References

[1] 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 *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 7

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. A5

[3] Sergio Casas, Abbas Sadat, and Raquel Urtasun. Mp3: A unified model to map, perceive, predict and plan. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, A5

[4] Devendra Singh Chaplot, Dhiraj Gandhi, Saurabh Gupta, Abhinav Gupta, and Ruslan Salakhutdinov. Learning to explore using active neural slam. In *International Conference on Learning Representations (ICLR)*, 2020. 2

[5] Devendra Singh Chaplot, Dhiraj Prakashchand Gandhi, Abhinav Gupta, and Russ R Salakhutdinov. Object goal navigation using goal-oriented semantic exploration. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2

[6] Yuxiao Chen, Boris Ivanovic, and Marco Pavone. Scept: Scene-consistent, policy-based trajectory predictions for planning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[7] Sanjiban Choudhury, Mohak Bhardwaj, Sankalp Arora, Ashish Kapoor, Gireeja Ranade, Sebastian Scherer, and Debadeepta Dey. Data-driven planning via imitation learning. *International Journal of Robotics Research (IJRR)*, 37(13-14):1632–1672, 2018. 3

[8] Alexander Cui, Sergio Casas, Abbas Sadat, Renjie Liao, and Raquel Urtasun. Lookout: Diverse multi-future prediction and planning for self-driving. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[9] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2, 7, A3

[10] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 3, 4

[11] Heming Du, Xin Yu, and Liang Zheng. Learning object relation graph and tentative policy for visual navigation. In *European Conference on Computer Vision (ECCV)*, 2020. 2

[12] Hao Fu, Chunyuan Li, Xiaodong Liu, Jianfeng Gao, Asli Celikyilmaz, and Lawrence Carin. Cyclical annealing schedule: A simple approach to mitigating kl vanishing. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2019. 2

[13] Weituo Hao, Chunyuan Li, Xiujun Li, Lawrence Carin, and Jianfeng Gao. Towards learning a generic agent for vision-and-language navigation via pre-training. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[14] Mohamed Hassan, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, and Michael J Black. Stochastic scene-aware motion prediction. In *International Conference on Computer Vision (ICCV)*, 2021. 2

[15] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *International Conference on Computer Vision (ICCV)*, 2019. 2, 5, A1

[16] Mohamed Hassan, Partha Ghosh, Joachim Tesch, Dimitrios Tzionas, and Michael J Black. Populating 3d scenes by learning human-scene interaction. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[17] Junxian He, Daniel Spokoyny, Graham Neubig, and Taylor Berg-Kirkpatrick. Lagging inference networks and posterior collapse in variational autoencoders. In *International Conference on Learning Representations (ICLR)*, 2018. 2

[18] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 3

[19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 3, A1, A3

[20] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *Journal of Machine Learning Research (JMLR)*, 23(47):1–33, 2022. 3, 4

[21] Yicong Hong, Qi Wu, Yuankai Qi, Cristian Rodriguez-Opazo, and Stephen Gould. Vln bert: A recurrent vision-and-language bert for navigation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[22] Aapo Hyvärinen and Peter Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research (JMLR)*, 6(4), 2005. 3

[23] Michael Janner, Yilun Du, Joshua B Tenenbaum, and Sergey Levine. Planning with diffusion for flexible behavior synthesis. In *International Conference on Machine Learning (ICML)*, 2022. 3, 4

[24] 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, 2

[25] Hanwen Jiang, Shaowei Liu, Jiashun Wang, and Xiaolong Wang. Hand-object contact consistency reasoning for human grasps generation. In *International Conference on Computer Vision (ICCV)*, 2021. 1, 2, 5, 7

[26] Yoon Kim, Sam Wiseman, Andrew Miller, David Sontag, and Alexander Rush. Semi-amortized variational autoencoders. In *International Conference on Machine Learning (ICML)*, 2018. 2

[27] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. A2

[28] Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Daniel Gordon, Yuke Zhu, Abhinav Gupta, and Ali Farhadi. Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474*, 2017. 2

[29] Steven M LaValle. *Planning algorithms*. Cambridge university press, 2006. 3

[30] 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. In *Conference on Robot Learning (CoRL)*, 2021. 1, 2

[31] Puhao Li, Tengyu Liu, Yuyang Li, Yixin Zhu, Yaodong Yang, and Siyuan Huang. Gendexgrasp: Generalizable dexterous grasping. In *International Conference on Robotics and Automation (ICRA)*, 2023. 1, 2, 5, 7, A2

[32] Xueting Li, Sifei Liu, Kihwan Kim, Xiaolong Wang, Ming-Hsuan Yang, and Jan Kautz. Putting humans in a scene: Learning affordance in 3d indoor environments. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2, 5

[33] Jenny Lin, Xingwen Guo, Jingyu Shao, Chenfanfu Jiang, Yixin Zhu, and Song-Chun Zhu. A virtual reality platform for dynamic human-scene interaction. In *SIGGRAPH ASIA 2016 virtual reality meets physical reality: Modelling and simulating virtual humans and environments*, 2016. 2

[34] Tengyu Liu, Zeyu Liu, Ziyuan Jiao, Yixin Zhu, and Song-Chun Zhu. Synthesizing diverse and physically stable grasps with arbitrary hand structures using differentiable force closure estimator. *IEEE Robotics and Automation Letters (RA-L)*, 7(1):470–477, 2021. 1, 2

[35] Shitong Luo and Wei Hu. Diffusion probabilistic models for 3d point cloud generation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[36] Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. Sqa3d: Situated question answering in 3d scenes. In *International Conference on Learning Representations (ICLR)*, 2023. 2

[37] Viktor Makoviychuk, Lukasz Wawrzyniak, Yunrong Guo, Michelle Lu, Kier Storey, Miles Macklin, David Hoeller, Nikita Rudin, Arthur Allshire, Ankur Handa, et al. Isaac gym: High performance gpu based physics simulation for robot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 7, A3

[38] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. A5

[39] 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. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 1

[40] Medhini Narasimhan, Erik Wijmans, Xinlei Chen, Trevor Darrell, Dhruv Batra, Devi Parikh, and Amanpreet Singh. Seeing the un-scene: Learning amodal semantic maps for room navigation. In *European Conference on Computer Vision (ECCV)*, 2020. 2

[41] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning (ICML)*, 2022. 3, A1

[42] Xue Bin Peng, Glen Berseth, and Michiel Van de Panne. Terrain-adaptive locomotion skills using deep reinforcement learning. *ACM Transactions on Graphics (TOG)*, 35(4):1–12, 2016. 1

[43] Xue Bin Peng, Ze Ma, Pieter Abbeel, Sergey Levine, and Angjoo Kanazawa. Amp: Adversarial motion priors for stylized physics-based character control. *ACM Transactions on Graphics (TOG)*, 40(4):1–20, 2021. 1

[44] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *International Conference on Learning Representations (ICLR)*, 2023. 3

[45] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 5, A1, A3

[46] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. A3

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

[48] Santhosh K Ramakrishnan, Dinesh Jayaraman, and Kristen Grauman. An exploration of embodied visual exploration. *International Journal of Computer Vision (IJCV)*, 129(5):1616–1649, 2021. 2

[49] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J Guibas. Humor: 3d human motion model for robust pose estimation. In *International Conference on Computer Vision (ICCV)*, 2021. 2

[50] PickNik Robotics. Moveit motion planning framework for ros 2. https://github.com/ros-planning/moveit2, 2020. 8, A3

[51] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3, 4

[52] Abbas Sadat, Mengye Ren, Andrei Pokrovsky, Yen-Chen Lin, Ersin Yumer, and Raquel Urtasun. Jointly learnable behavior and trajectory planning for self-driving vehicles. In *International Conference on Intelligent Robots and Systems (IROS)*, 2019. 1

[53] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 3, 4

[54] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *International Conference on Computer Vision (ICCV)*, 2019. 2

[55] Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[56] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. In *International Conference on Learning Representations (ICLR)*, 2023. 3

[57] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning (ICML)*, 2015. 3, A1

[58] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*, 2020. 3

[59] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 3, A1

[60] Sebastian Starke, He Zhang, Taku Komura, and Jun Saito. Neural state machine for character-scene interactions. *ACM Transactions on Graphics (TOG)*, 38(6):209–1, 2019. 1, 2

[61] 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. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 1

[62] Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. Grab: A dataset of whole-body human grasping of objects. In *European Conference on Computer Vision (ECCV)*, 2020. 2

[63] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022. 3

[64] I Tolstikhin, O Bousquet, S Gelly, and B Schölkopf. Wasserstein auto-encoders. In *International Conference on Learning Representations (ICLR)*, 2018. 2

[65] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 5

[66] Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011. 3

[67] Ayzaan Wahid, Austin Stone, Kevin Chen, Brian Ichter, and Alexander Toshev. Learning object-conditioned exploration using distributed soft actor critic. In *Conference on Robot Learning (CoRL)*, 2021. 2

[68] Jingbo Wang, Yu Rong, Jingyuan Liu, Sijie Yan, Dahua Lin, and Bo Dai. Towards diverse and natural scene-aware 3d human motion synthesis. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[69] Jiashun Wang, Huazhe Xu, Jingwei Xu, Sifei Liu, and Xiaolong Wang. Synthesizing long-term 3d human motion and interaction in 3d scenes. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 2, 5

[70] Jingbo Wang, Sijie Yan, Bo Dai, and Dahua Lin. Scene-aware generative network for human motion synthesis. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[71] Kai Wang, Manolis Savva, Angel X Chang, and Daniel Ritchie. Deep convolutional priors for indoor scene synthesis. *ACM Transactions on Graphics (TOG)*, 37(4):1–14, 2018. 2

[72] Weiqi Wang, Zihang Zhao, Ziyuan Jiao, Yixin Zhu, Song-Chun Zhu, and Hangxin Liu. Rearrange indoor scenes for human-robot co-activity. In *International Conference on Robotics and Automation (ICRA)*, 2023. 1

[73] Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. Humanise: Language-conditioned human motion generation in 3d scenes. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 1, 2, 5, 6

[74] Erik Wijmans, Abhishek Kadian, Ari Morcos, Stefan Lee, Irfan Essa, Devi Parikh, Manolis Savva, and Dhruv Batra. Dd-ppo: Learning near-perfect pointgoal navigators from 2.5 billion frames. In *International Conference on Learning Representations (ICLR)*, 2020. 2

[75] Jungdam Won, Deepak Gopinath, and Jessica Hodgins. Physics-based character controllers using conditional vaes. *ACM Transactions on Graphics (TOG)*, 41(4):1–12, 2022. 1, 2

[76] Mitchell Wortsman, Kiana Ehsani, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Learning to learn how to learn: Self-adaptive visual navigation using meta-learning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[77] Qi Wu, Cheng-Ju Wu, Yixin Zhu, and Jungseock Joo. Communicative learning with natural gestures for embodied navigation agents with human-in-the-scene. In *International Conference on Intelligent Robots and Systems (IROS)*, 2021. 2

[78] Yan Wu, Jiahao Wang, Yan Zhang, Siwei Zhang, Otmar Hilliges, Fisher Yu, and Siyu Tang. Saga: Stochastic whole-body grasping with contact. In *European Conference on Computer Vision (ECCV)*, 2022. 2

[79] Xu Xie, Hangxin Liu, Zhenliang Zhang, Yuxing Qiu, Feng Gao, Siyuan Qi, Yixin Zhu, and Song-Chun Zhu. Vrgym: A virtual testbed for physical and interactive ai. In *Proceedings of the ACM Turing celebration conference-China*, 2019. 2

[80] Wei Yang, Xiaolong Wang, Ali Farhadi, Abhinav Gupta, and Roozbeh Mottaghi. Visual semantic navigation using scene priors. In *International Conference on Learning Representations (ICLR)*, 2019. 2

[81] Hongwei Yi, Chun-Hao P. Huang, Shashank Tripathi, Lea Hering, Justus Thies, and Michael J. Black. MIME: Human-aware 3D scene generation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[82] Lap Fai Yu, Sai Kit Yeung, Chi Keung Tang, Demetri Terzopoulos, Tony F Chan, and Stanley J Osher. Make it home: automatic optimization of furniture arrangement. *ACM Transactions on Graphics (TOG)*, 30(4), 2011. 1

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

[84] Ye Yuan and Kris Kitani. Dlow: Diversifying latent flows for diverse human motion prediction. In *European Conference on Computer Vision (ECCV)*, 2020. 2, 5

[85] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*, 2022. 3

[86] Siwei Zhang, Yan Zhang, Federica Bogo, Pollefeys Marc, and Siyu Tang. Learning motion priors for 4d human body capture in 3d scenes. In *International Conference on Computer Vision (ICCV)*, 2021. 5

[87] Siwei Zhang, Yan Zhang, Qianli Ma, Michael J Black, and Siyu Tang. Generating person-scene interactions in 3d scenes. In *International Conference on 3D Vision (3DV)*, 2020. 2

[88] Song-Hai Zhang, Shao-Kui Zhang, Yuan Liang, and Peter Hall. A survey of 3d indoor scene synthesis. *Journal of Computer Science and Technology*, 34(3):594–608, 2019. 2

[89] Yan Zhang, Michael J Black, and Siyu Tang. We are more than our joints: Predicting how 3d bodies move. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 5, A2

[90] Yan Zhang, Mohamed Hassan, Heiko Neumann, Michael J Black, and Siyu Tang. Generating 3d people in scenes without people. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 2, 5, 6, A1, A2

[91] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *International Conference on Computer Vision (ICCV)*, 2021. 5, A1

[92] Kaifeng Zhao, Shaofei Wang, Yan Zhang, Thabo Beeler, and Siyu Tang. Compositional human-scene interaction synthesis with semantic control. In *European Conference on Computer Vision (ECCV)*, 2022. 2

[93] Shengjia Zhao, Jiaming Song, and Stefano Ermon. Infovae: Information maximizing variational autoencoders. *arXiv preprint arXiv:1706.02262*, 2017. 2

[94] Fengda Zhu, Yi Zhu, Xiaojun Chang, and Xiaodan Liang. Vision-language navigation with self-supervised auxiliary reasoning tasks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[95] Yuke Zhu, Roozbeh Mottaghi, Eric Kolve, Joseph J Lim, Abhinav Gupta, Li Fei-Fei, and Ali Farhadi. Target-driven visual navigation in indoor scenes using deep reinforcement learning. In *International Conference on Robotics and Automation (ICRA)*, 2017. 1, 2