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 [24, 47, 72, 82], 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.
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.
3. Background

3.1. Problem Definition

Given a 3D scene \( S \), we aim to generate the optimal solution for completing the tasks (e.g., navigation, manipulation) given the goal \( G \) in the scene. We denote the state and action of an agent as \( (s, a) \). The dynamic model defines the state transition as \( p(s_{t+1} | s_t, a_t) \), which is often deterministic in scene understanding (i.e., \( f(s_t, a_t) \)). The trajectory is defined as \( \tau = (s_0, a_0, \ldots, s_t, a_t, \ldots, s_N) \), where \( N \) denotes the horizon of task solving in discrete time.

3.2. Planning with Trajectory Optimization

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

\[
\tau^* = \arg \max_{\tau} J(\tau | S, G).
\]  
(1)

The dynamic model is usually known for trajectory optimization. Considering the future actions and states with predictable dynamics, the entire trajectory \( \tau \) 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 \( a_{0:N} = \arg \max_{a_{0:N}} \sum_{i=0}^{N} r(s_i, a_i | S, G) \).

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 \( q(\tau^* | \tau^{t-1}) \) gradually destroys data \( \tau^0 \sim \tau(0) \) into Gaussian noise. The parametrized reverse process \( p_\theta(\tau^{t-1} | \tau^t) \) recovers the data from noise with the learned normal distribution from a fixed timestep. The training objective for \( \theta \) 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 \( p(\tau^0 | S, G) \):

\[
p(\tau^0 | S, G) = p_\theta(\tau^0 | S)p_\phi(G | \tau^0, S) p(G | S) \propto p_\theta(\tau^0 | S)p_\phi(G | \tau^0, S).
\]  
(2)

Generation \( p_\theta(\tau^0 | S) \) 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:

\[
p_\theta(\tau^0 | S) = p(\tau^0) \prod_{t=1}^{T} p(\tau^{t-1} | \tau^t, S),
\]  
(3)

Optimization and Planning \( p_\phi(G | \tau^0, S) \) 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 \( p_\phi(O = 1 | \tau^0, S, G) \), 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 \( p_\phi(G | \tau^t, S) \) as its exponential in Eq. (5):

\[
p_\phi(G | \tau^t, S) = p_\phi(O = 1 | \tau^t, S, G) \propto \exp(J(\tau^t | S, G))
\]  
(4)

\[
= \exp(\varphi_p(\tau^t | S, G) + \varphi_o(\tau^t | S)),
\]  
(5)

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

4.1. Learning

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

\[
\mathcal{L}_\theta(\tau^0 | S) = \mathbb{E}_{t, \epsilon, \tau^0} \left[ \| \epsilon - \epsilon_\theta(\sqrt{\alpha_t} \tau^t + \sqrt{1-\alpha_t} \epsilon, t, S) \|_2^2 \right] = \mathbb{E} \left[ \| \epsilon - \epsilon_\theta(\tau^t, t, S) \|_2^2 \right],
\]  
(7)
where $\hat{t}^i$ is the pre-determined function in the forward process. With the learned $p_0(\tau^0|S)$, we sample $p(\tau^0|S, G)$ by taking the advantage of the diffusion model’s flexible conditioning \cite{10,23}. Specifically, we approximate $p_\phi(G|\tau^t, S)$ using the Taylor expansion around $\tau^t = \mu$ at timestep $t$ as

$$\log p_\phi(G|\tau^t, S) \approx (\tau^t - \mu) g + C,$$

where $C$ is a constant, $\mu = \mu_\phi(\tau^t, t, S)$ and $\Sigma = \Sigma_\phi(\tau^t, t, S)$ are the inferred parameters of original diffusion process, and

$$g = \nabla_{\tau^t} \log p_\phi(G|\tau^t, S)|_{\tau^t=\mu},$$

$$= \nabla_{\tau^t}(\phi_o(\tau^t|S) + \phi_p(\tau^t|S, G))|_{\tau^t=\mu}.$$ (9)

Therefore, we have

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

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

Of note, $\phi_o$ and $\phi_p$ 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 $p_0(\tau^0|S)$, followed by learning $\phi_o$ and $\phi_p$ for optimization and planning with the following objective:

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

Alg. 1 summarizes the training procedure.

**Algorithm 1**: Training SceneDiffuser

1. \(/\!\!/ \text{train base generation model}\)
   
   **Input**: Trajectory in 3D scene ($\tau^0, S$)

   2. repeat
   3. \(\tau^0 \sim p(\tau^0|S)\)
   4. \(e \sim \mathcal{N}(0, I), t \sim U(\{1, \ldots, T\})\)
   5. \(\tau^t = \sqrt{\alpha_t} \tau_0 + \sqrt{1 - \alpha_t} e\)
   6. \(\theta = -\nabla_\theta \mathcal{E}_o(e - \epsilon_\theta(\tau^t, t, S))^2\)
   7. \text{until converged;}
   8. \(/\!\!/ \text{(optional) train optimization and planning model}\)

   **Input**: Trajectory in 3D scene with goal ($\tau^0, S, G$), learned $\theta$ for $p_0(\tau^0|S)$

   9. repeat
   10. \(\tau^0 \sim p(\tau^0|S)\)
   11. \(e \sim \mathcal{N}(0, I), t \sim U(\{1, \ldots, T\})\)
   12. \(\mu = \mu_\phi(\tau^t, t, S), \Sigma = \Sigma_\phi(\tau^t, t, S)\)
   13. \(g = \nabla_{\tau^t} \log p_\phi(G|\tau^t, S)|_{\tau^t=\mu}\)
   14. \(\tau^t = \sqrt{\alpha_t} \tau_0 + \sqrt{1 - \alpha_t} e\)
   15. \(\phi = -\nabla_\phi \mathcal{E}_p(e - \epsilon_\theta(\tau^t, t, S) - \lambda \Sigma g)^2\)
   16. \text{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

1. \(/\!\!/ \text{one-step guided sampling}\)

   **function** sample ($\tau^0, G$):
   2. \(\mu = \mu_\phi(\tau^t, t, S), \Sigma = \Sigma_\phi(\tau^t, t, S)\)
   3. \(\tau^t+1 = \mathcal{N}(\tau^t+1 - \mu + \lambda \Sigma \nabla_p \phi(G(\tau^t, S), G)|_{\tau^t=\mu}, \Sigma)\)
   4. \text{return } \tau^t+1
   5. \(/\!\!/ \text{physics-based generation}\)

   **Input**: initial trajectory $\tau^0 \sim \mathcal{N}(0, I)$

   6. for $t = T, \ldots, 1$ do
   7. \(/\!\!/ \text{sampling with optimization}\)
   8. \(\tau^t+1 = \text{sample}(\tau^t, \phi_\phi(\cdot|S))\)
   9. \text{return } \tau^0
   10. \(/\!\!/ \text{goal-oriented planning}\)

   **Input**: planning steps $N$, starting state $s_0$, initial plan $\tau_0^0 \sim \mathcal{N}(0, I)$

   11. while not done and planning step $i < N$ do
   12. \(\text{for } t = T, \ldots, 1 \text { do}\)
   13. \(/\!\!/ \text{planning as inpainting}\)
   14. \(\tau^{t+1} = \text{sample}(\tau^t, \phi_\phi(\cdot|S) + \phi_\phi(\cdot|S, G))\)
   15. \text{Act } \hat{s}_{i+1} \text{ to reach } \hat{s}_{i+1}
   16. \(\text{Increment planning step } i = i + 1\)

**Scene-aware Generation** Sampling $\tau^0$ from the distribution $p_\phi(\tau^0|S)$ 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 $\phi_p$ in Eq. (4). For general optimization without the planning objective, the task goal $G$ is to sample a plausible trajectory in 3D scenes. Therefore, we can draw physically plausible trajectories in 3D scenes by sampling from $p(\tau^0|S, G)$ with Eq. (10).

**Goal-oriented Planning** This can be formulated as motion inpainting under the planning framework. Given the start state $\hat{s}_s$ and the goal state $\hat{s}_g$, the planning module returns trajectory $\tau = (\hat{s}_s, a_0, \ldots, \hat{s}_i, a_{i-1}, \ldots, \hat{s}_g)$ that can reach the goal state. We set the first state as $\hat{s}_0 = \hat{s}_s$ and define the goal state and reward of goal-reaching in $\phi_p$. 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 $\hat{a}_{i-1}, \hat{s}_i$. 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 \cite{20,51,53}. Specifically, we augment the time-conditional diffusion
Scene Encoder
Cross Attention
Self Attention
Res Block
PE
θ
Optimzie
Planner
×T

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.

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 ~ 53k frames in 8 scenes for training and others for testing.

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. 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.

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, $\{r_{\text{collision}}, r_{\text{contact}}, r_{\text{smoothness}}\}$. For planning, we consider the accumulation of simple step-wise distance, $r_{P_2}$. 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.

5.1. 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 ~ 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.

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

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

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

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.
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.

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 $s_0 \in \mathbb{R}^3$, and a target position $g \in \mathbb{R}^2$ 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 4.** Human motions generated by SceneDiffuser. Each row shows sampled human motions from the same start pose.

**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.

**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.

<table>
<thead>
<tr>
<th>model</th>
<th>succ. rate (%)</th>
<th>depth coll. (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>cVAE [25]</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>cVAE (w/ TTA.) [25]</td>
<td>0.00</td>
<td>21.91</td>
</tr>
<tr>
<td>ours (w/o opt.)</td>
<td>70.65</td>
<td>71.25</td>
</tr>
<tr>
<td>ours (w/ opt.)</td>
<td>71.27</td>
<td>69.84</td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th>task</th>
<th>model</th>
<th>succ. rate(%)</th>
<th>planning steps</th>
</tr>
</thead>
<tbody>
<tr>
<td>path plan</td>
<td>BC</td>
<td>0</td>
<td>150</td>
</tr>
<tr>
<td></td>
<td>deterministic($L_2$)</td>
<td>13.50</td>
<td>137.98</td>
</tr>
<tr>
<td></td>
<td>ours</td>
<td>73.75</td>
<td>90.38</td>
</tr>
<tr>
<td>arm motion</td>
<td>BC</td>
<td>0.31</td>
<td>299.08</td>
</tr>
<tr>
<td></td>
<td>deterministic($L_2$)</td>
<td>72.87</td>
<td>141.28</td>
</tr>
<tr>
<td></td>
<td>ours</td>
<td>78.59</td>
<td>147.60</td>
</tr>
</tbody>
</table>

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 $S \in \mathbb{R}^{4096 \times 3}$ and the robot arm trajectory with a sequence of joint angles $R \in [-\pi, \pi]$. 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.

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

Table 5. Ablation of the scale coefficient for optimization.

<table>
<thead>
<tr>
<th>metric</th>
<th>$\lambda = 0.1$</th>
<th>$\lambda = 1.0$</th>
<th>$\lambda = 10.0$</th>
<th>$\lambda = 100.0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>plausible rate ↑</td>
<td>28.75</td>
<td>52.5</td>
<td>21.25</td>
<td>0</td>
</tr>
<tr>
<td>APD (trans.) ↑</td>
<td>0.764</td>
<td>0.886</td>
<td>1.564</td>
<td>23.96</td>
</tr>
<tr>
<td>APD (param) ↑</td>
<td>3.206</td>
<td>3.243</td>
<td>9.040</td>
<td>573.6</td>
</tr>
<tr>
<td>non-collision score ↑</td>
<td>99.76</td>
<td>99.87</td>
<td>99.85</td>
<td>74.9</td>
</tr>
<tr>
<td>contact score ↑</td>
<td>99.70</td>
<td>99.65</td>
<td>81.75</td>
<td>0</td>
</tr>
</tbody>
</table>

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


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

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


Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. Sq3d: Situated question answering in 3d scenes. In International Conference on Learning Representations (ICLR), 2023. 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


