JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

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

Project Page

https://scenediffuser.github.io/

SceneDiffuser

- human pose generation
- human motion generation
- dexterous grasp generation
- path planing for 3D navigation with goals
- motion planning for robot arms

***SceneDiffuser** is a conditional generative model for 3D scene understanding.*

It is applicable to various scene-conditioned 3D tasks.

Long-standing Goals for 3D Scene Understanding

Scene-aware ***Generation*** Physics-based ***Optimization*** Goal-oriented ***Planning***

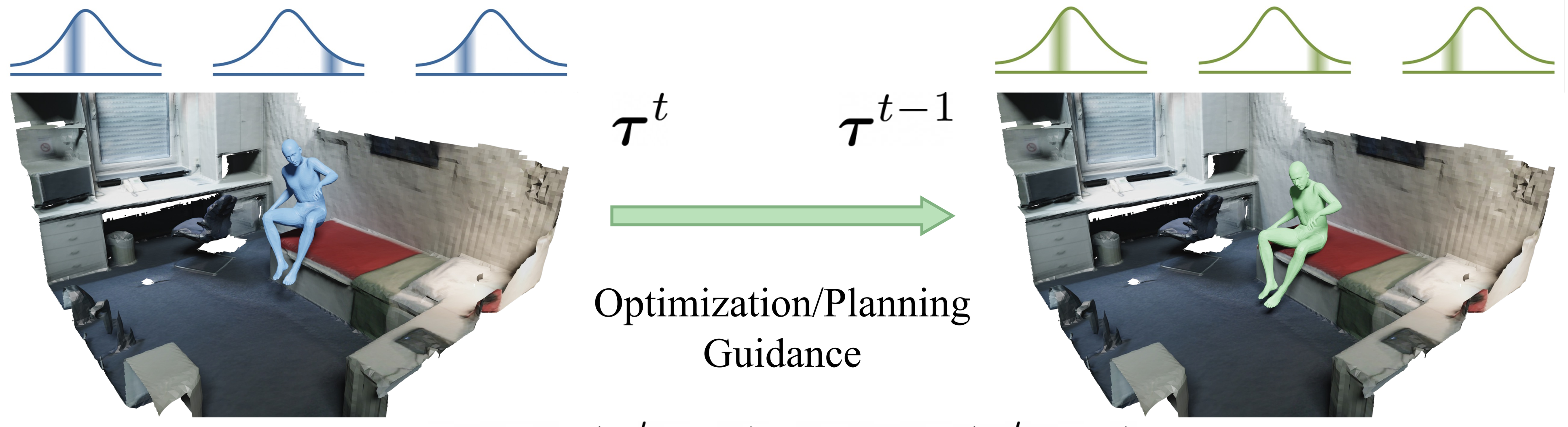
Two Fundamental Limitations

- Lack of ***powerful*** generative model
- Lack of ***unified*** framework

Contribution

- ✓ We propose the **SceneDiffuser** as a general conditional generative model for *generation*, *optimization*, and *planning* in 3D scenes.
- ✓ **SceneDiffuser** is intrinsically *scene-aware*, *physics-based*, and *goal-oriented*, applicable to various scene-conditioned 3D tasks.
- ✓ We demonstrate that the **SceneDiffuser** outperforms previous models by a *large margin* on **five** scene understanding tasks, establishing its **efficacy** and **flexibility**.

Sampling with Optimization/Planning Guidance



$$\boldsymbol{\mu} = \boldsymbol{\mu}_\theta(\boldsymbol{\tau}^t, t, \mathcal{S}), \boldsymbol{\Sigma} = \boldsymbol{\Sigma}_\theta(\boldsymbol{\tau}^t, t, \mathcal{S})$$

$$\boldsymbol{\tau}^{t-1} = \mathcal{N}(\boldsymbol{\tau}^{t-1}; \boldsymbol{\mu} + \lambda \boldsymbol{\Sigma} \nabla_{\boldsymbol{\tau}^t} (\mathcal{J}(\boldsymbol{\tau}^t | \mathcal{S}, \mathcal{G}))|_{\boldsymbol{\tau}^t = \boldsymbol{\mu}}, \boldsymbol{\Sigma})$$

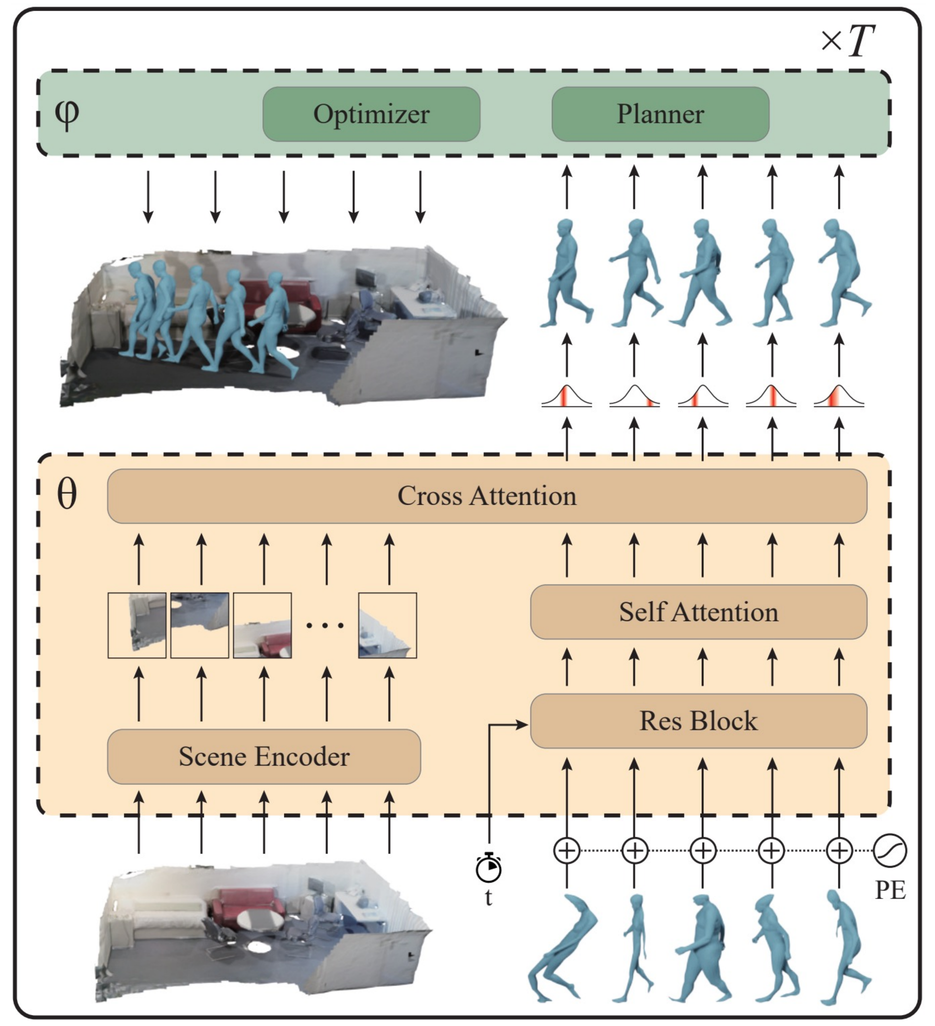
Denoising Process with Guidance

Step 0 · · · Step 30 · · · Step 60 · · · Step 100

Sampling Algorithm and Model Architecture

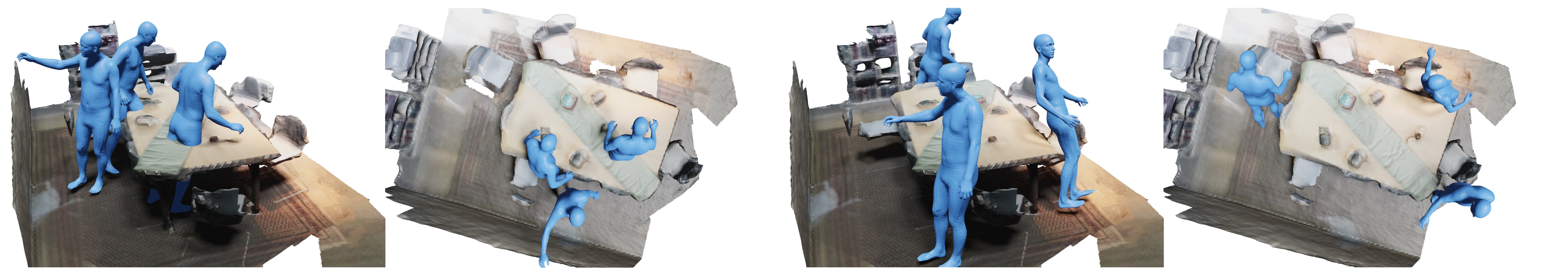
Algorithm 2: Sampling SceneDiffuser for generation, optimization, and planning

```
Modules: Model p_θ(·|S), optimization objective φ_o(·|S), and
         planner objective φ_p(·|S, G)
// one-step guided sampling
function sample (τ^t, J):
    μ = μ_θ(τ^t, t, S), Σ = Σ_θ(τ^t, t, S)
    τ^{t-1} = N(τ^{t-1}; μ + λΣ∇_{τ^t}(J(τ^t|S, G))|_{τ^t=μ}, Σ)
    return τ^{t-1}
// physics-based generation
Input: initial trajectory τ^T ~ N(0, I)
for t = T, ··· , 1 do
    // sampling with optimization
    τ^{t-1} = sample(τ^t, φ_o(·|S))
return τ^0
// goal-oriented planning
Input: planning steps N, starting state ŝ_0, initial plan
       τ_0^T ~ N(0, I)
i = 1
while not done and planning step i < N do
    for t = T, ··· , 1 do
        τ_i^{t-1} = sample(τ_i^t, φ_o(·|S) + φ_p(·|S, G))
        // planning as inpainting
        τ_i^{t-1}[0 : i] = ŝ_{0:i}
    Act â_{i-1} to reach ŝ_i = τ_i^0[i], ŝ_{0:i} = ŝ_{0:i-1} ∪ ŝ_i
    Increment planning step i = i + 1
```



Tasks and Results

Task 1: Human Pose Generation



cVAE **SceneDiffuser *with* Guidance**

Task 2: Human Motion Generation

SceneDiffuser generates diverse motions (e.g., "sit," "walk") from the same start position in unseen 3D scenes.

Task 3: Dexterous Grasp Generation

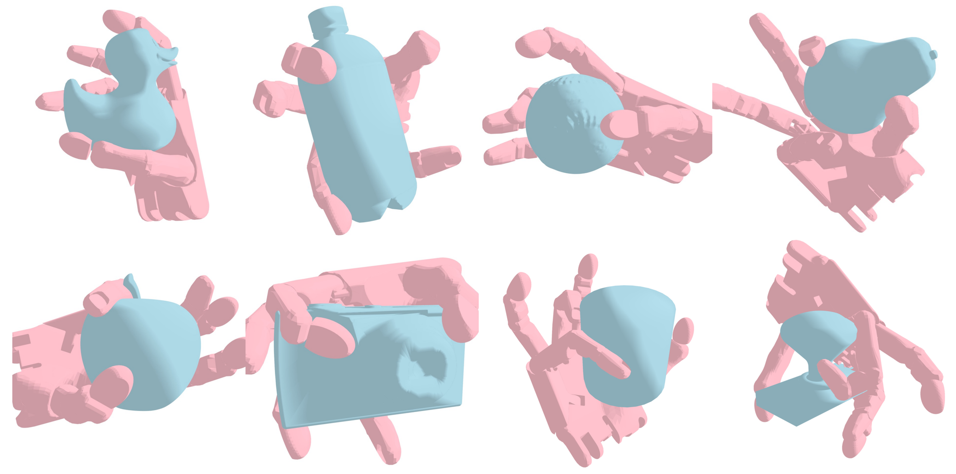


Table 3. **Quantitative results of dexterous grasp generation on MultiDex [31] dataset.** We measure the success rates under different diversities and depth collisions. TTA. denotes test-time optimization with physics and contact.

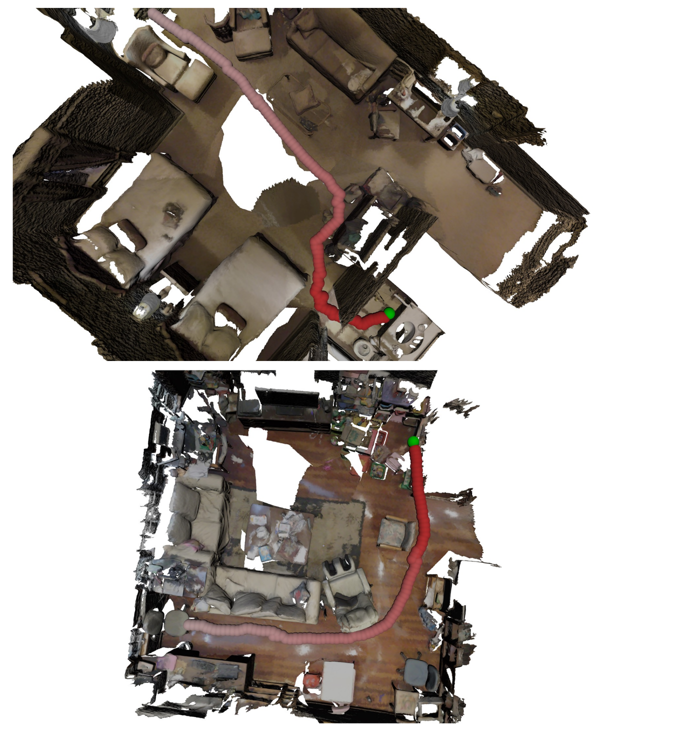
model	succ. rate (%)↑			depth coll. (mm)↓
	σ	2σ	all	
cVAE [25]	0.00	10.09	14.06	22.98
cVAE (w/ TTA.) [25]	0.00	21.91	17.97	15.19
ours (w/o opt.)	70.65	**71.25**	**71.25**	17.34
ours (w/ opt.)	**71.27**	69.84	69.84	**14.61**

Task 4: Path Planning for Navigation

Task 5: Motion Planning for Robot Arms

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



Please refer to our paper for more quantitative and qualitative results.