Language-driven Scene Synthesis using Multi-conditional Diffusion Model

An Vuong [1] Minh Nhat Vu [2, 3] Toan Nguyen [1, 4] Baoru Huang [5] Dzung Nguyen [1] Thieu Vo [6] Anh Nguyen [7]

[1]FPT Software AI Center [2]ACIN - TU Wien [3]Austrian Institute of Technology

[4]VNUHCM-University of Science [5]Imperial College London [6]Ton Duc Thang University [7]University of Liverpool

October 14, 2023

Contents

Introduction

Introduction

(a) Input scene

(b) Synthesis result

(c) Editing result

Imagine you are a VR character entering an apartment room. Initially, the room is empty, and you want to decorate the room with some furniture. With our proposed method, LSDM (**L**anguage-driven **S**cene Synthesis using Multi-conditional **D**iffusion **M**odel), you can synthesize objects such as a desk just by saying "*Place a desk in front of me*."

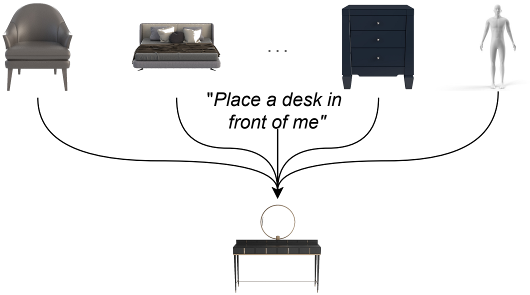
Key Contributions

- We present the language-driven scene synthesis task, a new challenge that generates objects based on human motions and given objects while following user linguistic commands.
- We propose a new multi-conditional diffusion model to tackle the language-driven scene synthesis task from multiple conditions.
- We validate our method empirically and theoretically, and introduce several scene-editing applications. The results show remarkable improvements over state-of-the-art approaches.

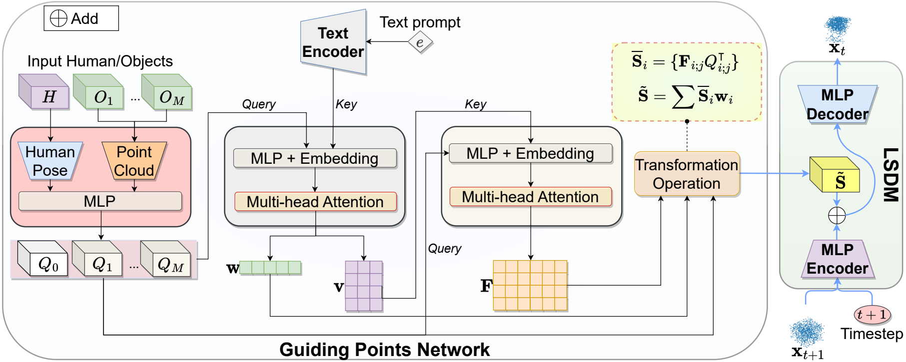
Methodology

Problem Statement

Given M objects represented as 3D point clouds in a room, and a natural language command e given by the user H, for instance, "*Place me an office chair under me*"; our goal is to synthesize the $M + 1$ object semantically aligned with the existing M objects, the human pose H, and the command e.



Method Overview



Theoretical Findings

Remark 1.2. We demonstrate that the guiding points $\tilde{\mathrm{S}}$ serves as the estimation of the original datapoint x_0, *explicitly* contributing to the denoising process q as follows

$$\hat{q}(\mathrm{x}_t|\mathrm{x}_{t+1}, \mathrm{y}) \approx \frac{q(\mathrm{x}_t|\mathrm{x}_{t+1})\hat{q}(\mathrm{y}|\mathrm{x}_t)}{\hat{q}(\mathrm{y}, \mathrm{x}_{t+1})} \frac{1}{|\widehat{\mathrm{S}}|} \sum_{\mathrm{x}_0 \in \hat{\mathrm{S}}} q(\mathrm{x}_{t+1}|\mathrm{x}_0) q(\mathrm{x}_0) \tag{1}$$

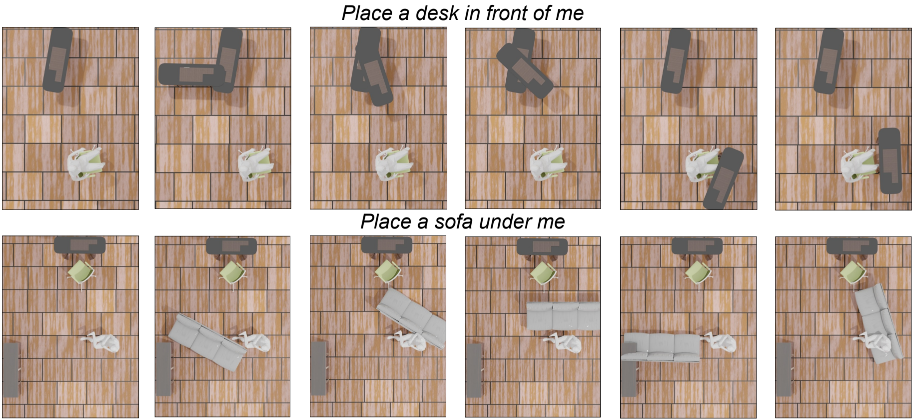
Results

Scene Synthesis

We compare our approach **L**anguage-driven **S**cene Synthesis using Multi-conditional **D**iffusion **M**odel (LSDM) with state-of-the-art scene synthesis literature.

	PRO-teXt			**HUMANISE**		
Baseline	CD ↓	EMD ↓	F1 ↑	CD ↓	EMD ↓	F1 ↑
ATISS [1]	2.0756	1.4140	0.0663	5.3595	2.0843	0.0308
SUMMON [2]	2.1437	1.3994	0.0673	5.3260	2.0827	0.0305
MIME [3]	2.0493	1.3832	<u>0.0990</u>	5.4259	2.0837	<u>0.0628</u>
MIME [3] + text embedding	1.8424	1.2865	0.1032	4.7035	1.8201	0.0849
MCDM	<u>0.6301</u>	<u>0.7269</u>	<u>0.3574</u>	<u>0.8586</u>	<u>0.8757</u>	<u>0.2515</u>
LSDM w.o. text (Ours)	0.9134	1.0156	0.0506	1.1740	1.1128	0.0412
LSDM (Ours)	**0.5365**	**0.5906**	**0.5160**	**0.7379**	**0.7505**	**0.4395**

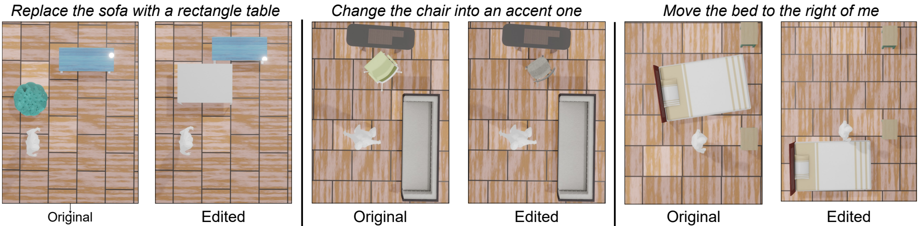
Qualitative Results



(a) Input (b) ATISS (c) SUMMON (d) MIME (e) Ours w.o. text (f) Ours

Scene Editing Applications

We introduce three editing applications: *i)* Object Replacement, *ii)* Shape Alternation, *iii)* Object Displacement. The results showcase meaningful scene editing demonstrations.



(a) Object Replacement **(b) Shape Alternation** **(c) Object Displacement**

Ablation Study (1/2)

How does each modality contribute to the performance? We analyze modality contributions to overall performance, confirming a significant enhancement in performance with the guiding point network.

Baseline	Input used	$\tilde{\mathrm{S}}$	CD ↓	EMD ↓	F1 ↑
LSDM w.o. predicting v	$\emptyset$	none	4.6172	2.1086	0.0391
LSDM w.o. predicting F	text, human, objects	partial	1.8933	1.1350	0.2400
LSDM predicting $\tilde{\mathrm{S}}$ from only objects	text, objects	partial	1.5050	1.0653	0.3185
LSDM predicting $\tilde{\mathrm{S}}$ from only human	text, human	partial	1.0119	0.8419	0.3855
LSDM (ours)	text, human, objects	full	**0.5365**	**0.5906**	**0.5160**

Ablation Study (2/2)

Can guiding points represent the target object? We provide both quantitative and qualitative assessments of predicted guiding points. In summary, the guiding points output from LSDM is meaningful, fulfilling the architecture design.

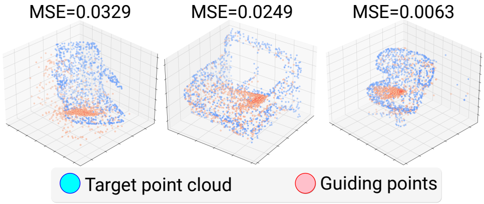


Figure: **Guiding points visualization.**

Baseline	MSE ↓
LSDM w.o. predicting F	0.5992
LSDM predicting $\tilde{\mathrm{S}}$ from only objects	0.4618
LSDM predicting $\tilde{\mathrm{S}}$ from only human	0.3388
LSDM (ours)	<u>0.2091</u>
Minimal squared distance d_0^2	**0.0914**

Table: **Guiding points evaluation.**

Conclusion

Conclusion

- We propse LSDM, a multi-conditional diffusion model based on the guiding point technique that can be further applied in other areas of Machine Learning.
- Theoretical findings and empirical evidence indicate our method demonstrate semantically plausible scene synthesis given room objects and linguistic instruction.
- The introduced language-driven scene synthesis and its editing operations have potential for applying into metaverse, animation, and design.

Reference

[1] Despoina Paschalidou, Amlan Kar, Maria Shugrina, Karsten Kreis, Andreas Geiger, and Sanja Fidler. Atiss: Autoregressive transformers for indoor scene synthesis. *Advances in Neural Information Processing Systems*, 34:12013–12026, 2021.

[2] Sifan Ye, Yixing Wang, Jiaman Li, Dennis Park, C Karen Liu, Huazhe Xu, and Jiajun Wu. Scene synthesis from human motion. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–9, 2022.

[3] Hongwei Yi, Chun-Hao P Huang, Shashank Tripathi, Lea Hering, Justus Thies, and Michael J Black. Mime: Human-aware 3d scene generation. *CVPR*, 2023.

Thank you!