Spatially-Conditional 3D Furniture Generation Model for Indoor Scene Synthesis

Jeongho Park, Obin Kwon and Songhwai Oh*

Department of Electrical and Computer Engineering and ASRI, Seoul National University,
Seoul, 08826, Korea (jeongho.park@rllab.snu.ac.kr, obin.kwon@rllab.snu.ac.kr, songhwai@snu.ac.kr)

* Corresponding author

Abstract: Recent advances in generative models have significantly enhanced the capabilities of 3D indoor scene synthesis, a field that is rapidly gaining interest due to its implications for applications such as embodied AI. These applications often require diverse, large-scale indoor datasets that include realistically modeled 3D furniture. Traditional text-to-3D models, while capable of producing realistic assets, often lack precise control over spatial dimensions, which is critical for ensuring that furniture fits appropriately within the scenes. This paper introduces SC-Shap·E, a spatially conditional 3D generative model that not only enhances the realism of 3D furniture but also ensures it adheres to specified spatial dimensions of height, width, and depth. Built on the pretrained Shap·E model, SC-Shap·E incorporates an additional network that utilizes spatial information alongside textual prompts, offering improved control over furniture sizing within generated scenes. By comparing our model with the original Shap·E, we demonstrate its superior ability to reflect accurate spatial conditions. Additionally, we present a novel three-stage system for 3D indoor scene synthesis that includes floorplan creation, furniture layout, and 3D furniture mesh production, showing its effective application in creating diverse and realistic 3D indoor scenes.

Keywords: 3D generative models, 3D indoor scene synthesis, diffusion model

1. INTRODUCTION

Recent advances in generative models have brought significant attention to 3D indoor scene synthesis, prompting numerous studies in this field [1–4]. The ability to create realistic and diverse indoor scenes is particularly useful in research such as embodied AI, which requires large-scale indoor datasets. Generating a variety of indoor scenes necessitates the creation of 3D furniture that is diverse and realistic in terms of visual aspects. Recent text-to-3D generative models can produce realistic 3D assets conditioned on a variety of categories of the objects [5–8]. However, to have more control over the scene synthesis process, it is also important to ensure that the items are appropriately sized and shaped for the intended space. For instance, when placing a bed and a wardrobe in a generated bedroom, the sizes of the bed and wardrobe must not exceed the size of the bedroom itself. One approach is to forcibly scale the generated furniture to fit the space, but this impairs the quality of the generated scenes.

In this work, we introduce SC-Shap·E, a spatially conditional 3D generative model that generates plausible 3D furniture while following conditions specified by spatial information like height, width, and depth of the generated assets. Our proposed model is built on the pretrained Shap·E [6] model while including an additional network that can take spatial information as a condition as well as the text prompt. It is further trained on 3D-FRONT [9] dataset, which is a realistic large-scale 3D furniture dataset. By comparing our proposed model with the original Shap·E model, which takes spatial information through input text prompts, we demonstrate that our model better reflects the given spatial conditions in generating 3D furniture.

We also introduce a novel system that utilizes SC-Shap·E for 3D indoor scene synthesis. This process is modularized into three stages: 1. floorplan generation, 2. furniture layout generation, and 3. production of 3D furniture meshes. We provide some demonstrations of the synthesized 3D scenes and illustrate that our proposed model can be effectively utilized in 3D indoor scene synthesis.

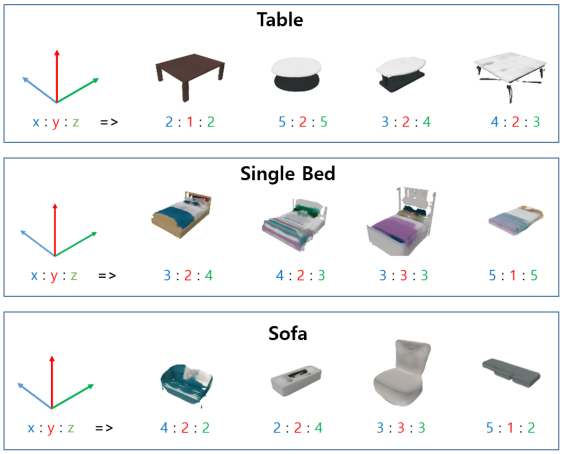

Fig. 1.: Examples of 3D furniture meshes generated by SC-Shap·E which uses furniture categories and spatial conditions (size ratios of the 3D bounding box) as inputs to produce the corresponding meshes.

2. RELATED WORK

2.1 3D Generative Model

Generation of 3D objects is a difficult task due to the need to consider not only the visual aspects of textures

but also the 3D geometric structures of the generated objects. GET3D [5] represents a high-quality textured 3D mesh as a function that maps coordinates to colors, signed distances, and vertex offsets, and can be conditioned on input text or images. Point·E [7] employs transformer-based diffusion model to generate 3D point clouds with colors on them. Following this model's network architecture, Shap·E [6] further generates neural radiance field (NeRF) and signed distance field (SDF) which can be rendered into images and converted into 3D meshes with textures. DMV3D [8] also employs transformer-based diffusion model in order to generate full 3D NeRF of the partial information obtained from the input such as a single image or text prompt.

2.2 3D Indoor Scene Synthesis

3D indoor scene synthesis is also a complex task that requires not only the creation of furniture but also the realistic generation of structures such as floors and walls, as well as their spatial relationships. ProcTHOR [10] is a framework for the procedural generation of indoor scenes, specifically designed for training Embodied AI. However, it constructs scene furniture using a limited set of hand-crafted 3D asset furniture. Recently, end-to-end 3D scene synthesis methods [1–3, 11] based on 3D GAN techniques have been proposed that aim to handle the process of synthesizing scenes through a single module. CommonSenses [4] generates controllable 3D scenes from input scene graphs through the diffusion process.

3. BACKGROUND

3.1 Shap·E

Shap·E is a conditional generative model for 3D assets, which is capable of producing diverse and recognizable samples conditioned on text prompts or images. In this paper, we only consider Shap·E based on text prompts. Shap·E consists of a 3D encoder, a decoder, and a latent diffusion model. The 3D encoder, which is used only during the training process, is based on the structure of the transformer model [12]. It processes the 3D point cloud and multi-view images of the input asset through attention layers to obtain latent representations as a sequence of vectors. These latent representations pass through projection layers to become the parameters of multi-layer perceptrons (MLP) that represent the asset as an implicit function.

The decoder then utilizes the MLPs to decode the latent representations back to images. Specifically, a decoder has three separate output heads that generate opacity values, RGB color values, and signed distance function (SDF) values, respectively. Images can be rendered based on the NeRF formulation [13], or meshes can be generated by using Marching Cubes method [14].

During training, the encoder and decoder are trained using a large dataset of 3D assets. The loss is calculated by comparing the images of the encoded 3D asset with the images decoded using the NeRF method and also, the images projected from the mesh generated using the Marching Cubes 33 method.

The generative module of Shap·E is a diffusion model [15, 16] that generate the latent representations from Gaussian noise. The architecture of the diffusion model is also based on a transformer model. The latents are sequences of vectors, each treated as a token, and are fed into the transformer-based diffusion model. While training, every 3D asset in the dataset is encoded into a latent representation by the pretrained 3D encoder, and the diffusion model is trained to generate these latents from noise. For text-conditional generation, a single token containing the CLIP text embedding is prepended to the sequence of tokens. To support classifier-free guidance [17], a zero vector is randomly inserted into the condition with a probability of 0.1.

4. METHOD

Our proposed model, SC-Shap·E takes text prompts representing the furniture category and spatial conditions defined by the ratios of depth, height, and width as inputs, and outputs NeRF and SDF for 3D furniture assets. These NeRF and SDF can be rendered into images and converted into meshes, respectively, and further used for 3D indoor scene synthesis. We present and describe a system capable of creating 3D indoor scenes utilizing SC-Shap·E model.

4.1 Spatially-Conditional 3D Furniture Generation

In this paper, the spatial condition $\mathbf{c_s}$ is defined as the ratio of an object's 3D bounding box's depth, height, and width, which can be expressed as

$$\mathbf{c_s} = (\mathbf{x}, \mathbf{y}, \mathbf{z}), \quad \text{where} \quad \mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^+, \tag{1}$$

where the variables x, y, and z are defined to represent the proportions of depth, height, and width of the 3D bounding box, respectively. However, to utilize these ratios as conditions in the diffusion model, they are normalized so that their sums equal one. We opt for size ratios rather than exact values because the original Shap·E model is trained on objects of consistent size, independent of their actual scales. This approach helps prevent significant deviations from the established model formulation.

The original Shap·E model represents the 3D information of an object as a sequence of vectors generated by a diffusion model. Each vector in the sequence is considered a token, with the final token corresponding to a CLIP text embedding.To reflect our objectives in the application of the Shap·E model, we introduce an additional token that encapsulates spatial condition information. The spatial condition, $\mathbf{c_s}$, is normalized and encoded into an embedding using a straightforward MLP-based encoder. This new token embedding is designed to be the same length as the embeddings of existing tokens, allowing it to be processed seamlessly through the transformer-based network of the Shap·E.

In the training phase, our approach employs classifier-

free guidance following the original Shap·E, where we randomly decide whether to condition or uncondition the model based on a specific probability during each iteration. However, unlike typical implementations that may consider a single type of condition, our model handles two distinct conditions: text prompts and spatial conditions. We treat these two conditions independently, randomizing each one to ensure that our model experiences and learns from all possible combinations of conditions during training.

We leverage the large-scale 3D furniture dataset from 3D-FRONT [9] for the training, which includes a vast array of 3D photorealistic synthetic furniture meshes complete with textures. Each piece of furniture in the dataset is tagged with a textual category label, which we utilize as a text prompt during training. Additionally, we pre-calculate the depth, height, and width of all furniture items to use as spatial condition information. This integration of detailed physical dimensions with textual descriptions allows our model to develop a deeper understanding of how textual and spatial conditions interact and influence the generation process.

4.2 3D Indoor Scene Synthesis System

We introduce a novel 3D indoor scene synthesis system utilizing our proposed model, which is structured into three modular stages. 1. floorplan generation, 2. furniture layout generation, and 3. 3D furniture generation. While existing methods [18, 19] are employed in the first and the second stage, we utilize our proposed model in the third stage to generate the 3D furniture meshes following the requested categories and size conditions from the previous stages.

Floorplan generation. We employ the HouseDiffusion model [19], which takes as input conditions the types and number of rooms, their connections, and the connectivity information between the corners of the floor, and generates the floorplan of the indoor scene. The generated floorplan includes the 2D locations of the corners of the floors of the rooms. Structural parts of the indoor scene, such as floors and walls, are first generated as 3D meshes by employing a simple rule-based method.

Furniture layout generation. A furniture layout includes the semantic categories, sizes, locations, and rotation angles of the furniture assets. Given the floorplan generated in the first stage, ATISS [18] generates diverse furniture layouts considering the semantic relationships between the furniture.

3D furniture generation. Based on the categories and sizes of the furniture produced in the layout, we create text prompts and spatial conditions. Then SC-Shap·E generates 3D furniture following the conditions. To construct a 3D scene, the NeRF and SDF of the produced 3D furniture sets are converted into 3D meshes using the Marching Cubes method [14]. The generated furniture meshes are moved to the positions proposed in the previous stage and rotated by the suggested angles.

Table 1.: **3D furniture generation result.** This evaluation measures how closely the 3D furniture meshes generated by each model match the input spatial conditions, specifically in terms of the depth, height, and width of the 3D bounding box. 'MSE', 'L1', and 'CS' denote mean squared error, L1 error, and cosine similarity values, respectively.

	SC-Shap·E (ours)	Shap·E-mod	Shap·E
MSE (↓)	**0.003**	0.018	0.017
L1 (↓)	**0.041**	0.097	0.100
CS (↑)	**0.990**	0.864	0.833

5. EXPERIMENTS

We compare our spatially-conditional 3D furniture generation model against the original pretrained Shap·E model, which is publically available and evaluate how well they reflect the spatial conditions in their outputs. Also, We demonstrate the effectiveness of our spatially-conditional 3D furniture generation model for 3D indoor scene synthesis by showcasing the demonstrations of our proposed 3D indoor scene synthesis system.

5.1 Spatially-Conditional 3D Furniture Generation

First, we provide some qualitative results of our qualitative results in Fig 1. It shows examples of 3D furniture meshes generated by the SC-Shap·E model, which receives text prompts based on furniture categories and spatial conditions defined by size ratios. From these results, it is evident that the SC-Shap·E model is capable of generating reasonable 3D shapes that conform to the input conditions.

We also quantitatively evaluate how well our proposed model reflects the given spatial conditions by comparing it with several baseline models. While the original Shap·E model does not have any explicit way to constrain the size-ratio of the output, it can indirectly incorporate the desired spatial condition through text prompts. As a comparison baseline, we employ a model that generates furniture by adding the following phrase to the text prompt: "*with proportions of depth, height, and width in a x:y:z ratio*", which we denote as "Shap·E-mod". x,y, and z are filled with the depth, height, and width values in the input spatial condition. We also employ the original Shap·E model, which does not receive any spatial condition as input, to serve as a benchmark for evaluating the performance of other models.

Table 1 compares the results of how well the 3D furniture meshes generated by each model match the input spatial condition which is the ratio of depth, height, and width of the 3D bounding box. Three different metrics are used to evaluate the results. Mean-squared error (MSE) and L1 error directly measure the discrepancies between the generated meshes and the input spatial con-

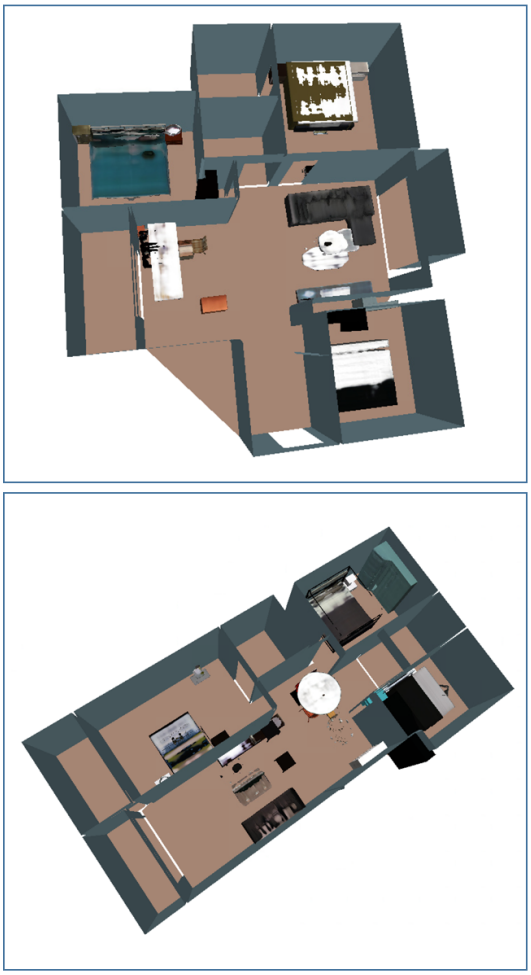

Fig. 2.: Examples of 3D indoor scenes synthesized by our proposed system. Despite being synthesized under the same input conditions—1 living room and 3 bedrooms—these two scenes showcase the system's capability to produce diverse configurations.

ditions. Cosine Similarity (CS) is defined as follows:

$$\mathrm{CS}(\mathbf{c_s}, \hat{\mathbf{c}}_\mathbf{s}) = \frac{\mathbf{c_s} \cdot \hat{\mathbf{c}}_\mathbf{s}}{\|\mathbf{c_s}\|\|\hat{\mathbf{c}}_\mathbf{s}\|}, \tag{2}$$

where $\hat{\mathbf{c}}_\mathbf{s}$ refers to the size ratio of the generated furniture. It indicates how closely the directional alignment of the size ratio within the 3D bounding box of the generated mesh aligns with the input spatial conditions. We conducted experiments using 6 furniture categories: 'armchair', 'double bed', 'wardrobe', 'single bed', 'bookshelf', and 'desk'. For each category, we tested 4 different spatial conditions, and for all scenarios, we ran the experiments with two different seeds.

Results in Table 1 show that SC-Shap·E significantly better reflects the input spatial conditions compared to other baselines across all three metrics. Note that Shap·E-mod does not show better performance compared to the original Shap·E model that works without input spatial conditions, demonstrating the limitations of specifying precise spatial conditions through text prompts.

5.2 3D Indoor Scene Synthesis

Here, we present demonstrations of the synthetic 3D indoor scenes synthesized by our proposed system, as shown in Fig. 2. The synthesized scenes include 3D furniture meshes, such as beds, sofas, and tables, which are placed in reasonable spots without any unnatural visual features. All scene elements are rendered as 3D meshes, making them suitable for use in physics-based simulators for embodied AI. Note that although these indoor scenes are synthesized from the same input condition—1 living room and 3 bedrooms—they display totally different layouts and furniture shapes. This diversity is due to the stochasticity embedded in each stage of our system, allowing for the generation of an infinite variety of 3D scenes.

6. CONCLUSIONS

In this paper, we introduced SC-Shap·E, a spatially-conditional 3D Furniture Generation Model designed to precisely generate 3D furniture based on the specified ratios of depth, height, and width. Unlike conventional methods that implicitly incorporate spatial conditions through text prompts, SC-Shap·E explicitly utilizes these conditions, demonstrating superior performance in accurately reflecting the specified spatial conditions. Furthermore, we have developed a novel 3D indoor scene synthesis system that integrates SC-Shap·E, enabling the production of 3D furniture meshes that precisely conform to and integrate with the layouts created by the system's floorplan and object layout modules, due to our model's capabilities. Our demonstrations confirm that SC-Shap·E can be effectively implemented within the 3D indoor scene synthesis process. Looking ahead, future developments will aim to expand the model's capabilities by incorporating additional conditions, such as visual style and other innovative features, thereby enhancing its ability to create even more diverse and customized 3D furniture for indoor scenes.

ACKNOWLEDGEMENT

This work was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (NRF-2022R1A2C2008239, General-Purpose Deep Reinforcement Learning Using Metaverse for Real World Applications).

REFERENCES

[1] S. Bahmani, J. J. Park, D. Paschalidou, X. Yan, G. Wetzstein, L. Guibas, and A. Tagliasacchi, "Cc3d: Layout-conditioned generation of compositional 3d scenes," in *Proceedings of the*

IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 7171–7181.

[2] Y. Xu, M. Chai, Z. Shi, S. Peng, I. Skorokhodov, A. Siarohin, C. Yang, Y. Shen, H.-Y. Lee, B. Zhou *et al.*, "Discoscene: Spatially disentangled generative radiance fields for controllable 3d-aware scene synthesis," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2023, pp. 4402–4412.

[3] M. Son, J. J. Park, L. Guibas, and G. Wetzstein, "Singraf: Learning a 3d generative radiance field for a single scene," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2023, pp. 8507–8517.

[4] G. Zhai, E. P. Örnek, S.-C. Wu, Y. Di, F. Tombari, N. Navab, and B. Busam, "Commonscenes: Generating commonsense 3d indoor scenes with scene graphs," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 36, 2024.

[5] J. Gao, T. Shen, Z. Wang, W. Chen, K. Yin, D. Li, O. Litany, Z. Gojcic, and S. Fidler, "Get3d: A generative model of high quality 3d textured shapes learned from images," *Advances In Neural Information Processing Systems (NeurIPS)*, vol. 35, pp. 31 841–31 854, 2022.

[6] H. Jun and A. Nichol, "Shap-e: Generating conditional 3d implicit functions," *arXiv preprint arXiv:2305.02463*, 2023.

[7] A. Nichol, H. Jun, P. Dhariwal, P. Mishkin, and M. Chen, "Point-e: A system for generating 3d point clouds from complex prompts," *arXiv preprint arXiv:2212.08751*, 2022.

[8] Y. Xu, H. Tan, F. Luan, S. Bi, P. Wang, J. Li, Z. Shi, K. Sunkavalli, G. Wetzstein, Z. Xu *et al.*, "Dmv3d: Denoising multi-view diffusion using 3d large reconstruction model," in *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

[9] H. Fu, B. Cai, L. Gao, L.-X. Zhang, J. Wang, C. Li, Q. Zeng, C. Sun, R. Jia, B. Zhao *et al.*, "3d-front: 3d furnished rooms with layouts and semantics," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 10 933–10 942.

[10] M. Deitke, E. VanderBilt, A. Herrasti, L. Weihs, K. Ehsani, J. Salvador, W. Han, E. Kolve, A. Kembhavi, and R. Mottaghi, "Procthor: Large-scale embodied ai using procedural generation," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35, pp. 5982–5994, 2022.

[11] T. DeVries, M. A. Bautista, N. Srivastava, G. W. Taylor, and J. M. Susskind, "Unconstrained scene generation with locally conditioned radiance fields," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 14 304–14 313.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.

[13] B. Mildenhall, "Nerf: Representing scenes as neural radiance fields for view synthesis," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[14] W. LORENSEN, "Marching cubes: A high resolution 3d surface construction alghrithm," in *Proceedings of ACM SIGGRAPH'87*. ACM Press, 1987, pp. 71–78.

[15] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems (NeurIPS)*, vol. 33, pp. 6840–6851, 2020.

[16] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2022, pp. 10 684–10 695.

[17] J. Ho and T. Salimans, "Classifier-free diffusion guidance," in *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

[18] D. Paschalidou, A. Kar, M. Shugrina, K. Kreis, A. Geiger, and S. Fidler, "Atiss: Autoregressive transformers for indoor scene synthesis," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, pp. 12 013–12 026, 2021.

[19] M. A. Shabani, S. Hosseini, and Y. Furukawa, "Housediffusion: Vector floorplan generation via a diffusion model with discrete and continuous denoising," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 5466–5475.