DiffuScene: Denoising Diffusion Models for Generative Indoor Scene Synthesis

Jiapeng Tang[1] Yinyu Nie[1] Lev Markhasin[2] Angela Dai[1] Justus Thies[3] Matthias Nießner[1]

[1] Technical University of Munich [2] Sony Semiconductor Solutions Europe
[3] Technical University of Darmstadt
https://tangjiapeng.github.io/projects/DiffuScene

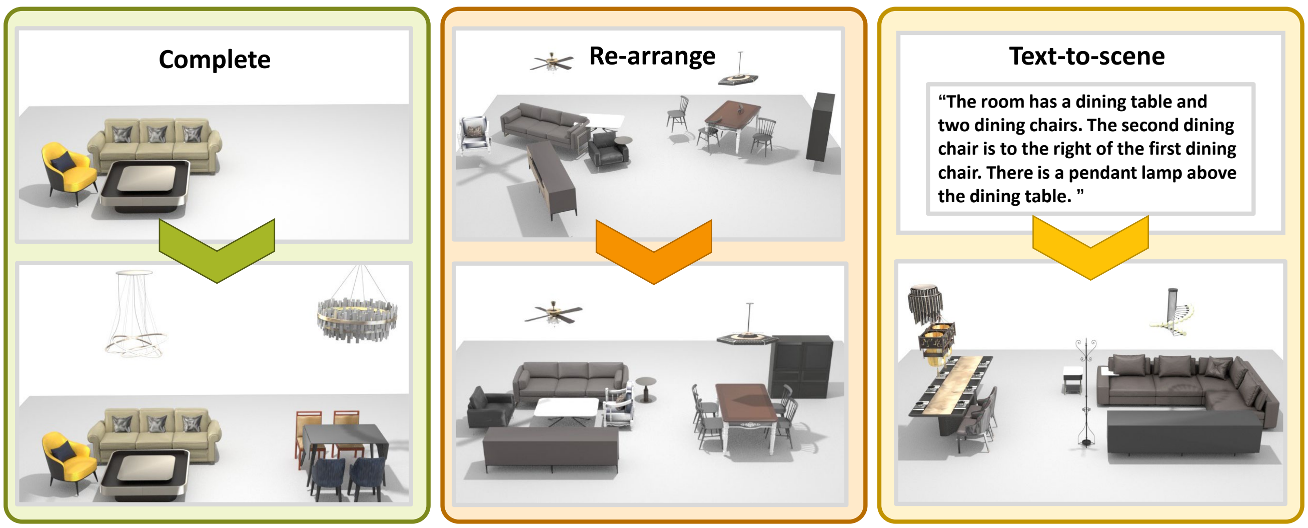


Figure 1. We present *DiffuScene*, a diffusion model for diverse and realistic indoor scene synthesis. It facilitates various downstream applications: scene completion from partial scenes (left); scene arrangements of given objects (middle); and scene generation from a text prompt describing partial scene configurations. (right).

Abstract

We present DiffuScene for indoor 3D scene synthesis based on a novel scene configuration denoising diffusion model. It generates 3D instance properties stored in an unordered object set and retrieves the most similar geometry for each object configuration, which is characterized as a concatenation of different attributes, including location, size, orientation, semantics, and geometry features. We introduce a diffusion network to synthesize a collection of 3D indoor objects by denoising a set of unordered object attributes. Unordered parametrization simplifies and eases the joint distribution approximation. The shape feature diffusion facilitates natural object placements, including symmetries. Our method enables many downstream applications, including scene completion, scene arrangement, and text-conditioned scene synthesis. Experiments on the 3D-FRONT dataset show that our method can synthesize more physically plausible and diverse indoor scenes than state-of-the-art methods. Extensive ablation studies verify the effectiveness of our design choice in scene diffusion models.

1. Introduction

Synthesizing 3D indoor scenes that are realistic, semantically meaningful, and diverse is a long-standing problem in computer graphics. It can significantly reduce costs in game development, CGI for films, and virtual reality. Furthermore, scene synthesis has practical applications in virtual interior design, enabling virtual rearrangement based on existing furniture or textual descriptions. It also serves as a fundamental component in data-driven approaches for 3D scene understanding and reconstruction, necessitating large-scale 3D datasets with ground-truth labels.

Traditional scene modeling and synthesis formulate this as an optimization problem. With pre-defined scene prior constraints defined by room design rules such as layout guidelines [38, 78], object category frequency distributions [4, 5, 14], affordance maps from human-object interactions [16, 19, 29], or scene arrangement examples [15, 19], they initially sample an initial scene and subsequently refine scene configurations through iterative optimization. However, defining precise rules is time-consuming and demands significant artistic expertise. The scene optimization stage is often laborious and computationally inefficient. Additionally, predefined design rules may limit the expression of complex and diverse scene compositions.

To automate the scene synthesis, some approaches [33, 42, 44, 46, 51, 67–69, 75, 76, 85] resort to deep generative models to learn scene priors from large-scale datasets. GAN-based methods [76] implicitly fit the scene distribution via adversarial training, yielding favorable results. However, they often lack diversity due to limited mode coverage and are prone to mode collapse. VAE-based methods [46, 75] explicitly approximate the scene distribution, offering better generative diversity but with lower-fidelity results. Recent auto-regressive models [42, 44, 69] progressively predict object properties sequentially. However, the sequential process may not accurately capture inter-object relationships and can accumulate prediction errors.

To capture more complicated scene configuration patterns for diverse scene synthesis, we strive to design a diffusion model for 3D scene synthesis. Diffusion models offer a compelling balance between diversity and realism and are relatively easier to train compared to other generative models [6, 13, 20, 21, 31, 49, 50, 64, 65]. In this work, we represent a scene as a set of unordered objects, with each element comprising a concatenation of various attributes, including location, size, orientation, semantics, and geometry features. Compared to other scene representations like multi-view images [10, 22, 32], voxel grids [8, 71], and neural fields [7, 39, 40, 43, 61], our representation is more compact and lightweight, making it suitable for learning through diffusion models. Rather than representing a scene as an ordered object sequence and diffusing them sequentially [44, 69], unordered set diffusion simplifies and eases the approximation of the joint distribution of object instances. To this end, we design a denoising diffusion model [24, 25, 59] to estimate object attributes to determine the placements and types of 3D instances and then perform shape retrieval to obtain final surface geometries. The scene diffusion priors are learned through iterative transitions between noisy and clean object sets, allowing for generating a diverse range of physically plausible scenes. During denoising, we simultaneously refine the properties of all objects within a scene, explicitly leveraging spatial relationships through an attention mechanism [66]. Different from previous works [44, 69, 75] that only predict object bounding boxes, we diffuse semantics, oriented bounding boxes, and geometry features together to promote a holistic understanding of composition structure and surface geometries. The synthesized shape codes for geometry retrieval can produce more natural object arrangements, such as symmetric relations commonly seen in the real world. We show compelling results in the unconditional and conditional settings against state-of-the-art scene generation models and provide extensive ablation studies to verify the design choices of our method.

Our contributions can be summarized as follows.

- We introduce 3D scene denoising diffusion models for diverse indoor scene synthesis, which learn holistic scene configurations of object semantics, placements, and geometries.
- We introduce shape latent feature diffusion for geometry retrieval, which exploits accurate inter-object relationships for symmetry formation.
- Based on this proposed model we facilitate completion from partial scenes, object re-arrangement in an existing scene, as well as text-conditioned scene synthesis.

2. Related work

Traditional Scene Modeling and Synthesis Traditional methods usually formulate this problem into a data-driven optimization task. To synthesize plausible 3D scenes, prior knowledge of reasonable configurations is required to drive scene optimization. Scene priors were often defined by following guidelines of interior design [38, 78], object frequency distributions (e.g., co-occurrence map of object categories) [4, 5, 14], affordance maps from human motions [16, 19, 29, 36, 47], or scene arrangement examples [15, 19]. Constrained by scene priors, a new scene can be sampled from the formulation using different optimization methods, e.g., iterative methods [16, 19], non-linear optimization [4, 15, 47, 74, 78, 80], or manual interaction [5, 38, 54]. Unlike them, we learn complicated scene composition patterns from datasets, avoiding human-defined constraints and iterative optimization processes.

Learning-based Generative Scene Synthesis 3D deep learning reforms this task by learning scene priors in a fully automatic, end-to-end, and differentiable manner. The capacity to process large-scale datasets dramatically increases the inference ability in synthesizing diverse object arrangements. Existing generative models for 3D scene synthesis are usually based on feed-forward networks [72, 85], VAEs [46, 75] , GANs [76], or Autoregressive models [42, 44, 69]. GAN methods generate high-quality results rapidly but often lack mode coverage and diversity. VAEs offer better mode coverage but face challenges in generating faithful samples [73]. Recurrent networks [33, 42, 44, 51, 67–69] including autoregressive models predict each new object conditioned on the previously generated objects. In con-

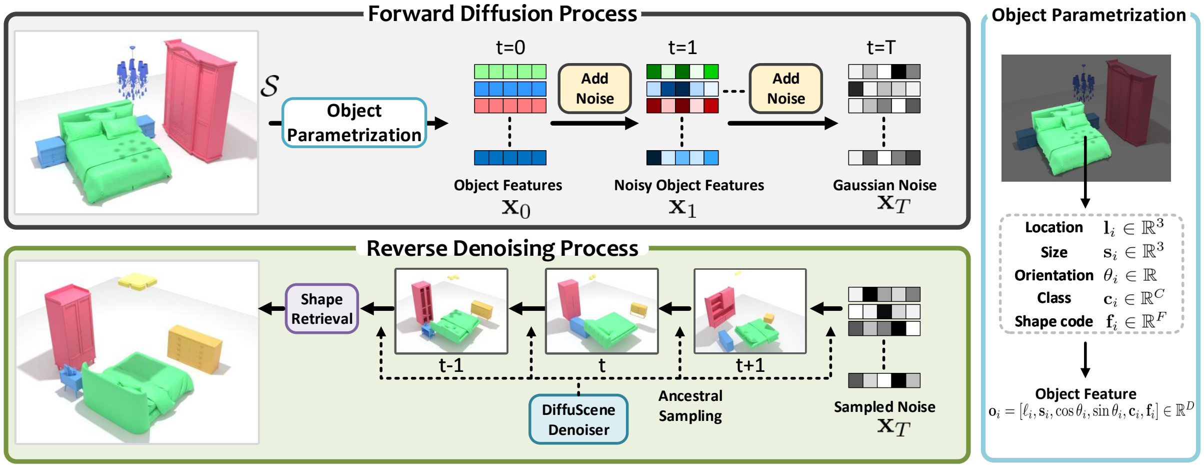


Figure 2. **Overview.** Given a 3D scene $\mathcal{S}$ of N objects, we represent it as an unordered set $\mathbf{x}_0 = \{\mathbf{o}_i\}_{i=1}^N$, by parametrizing each object $\mathbf{o}_i$ as a vector storing all object attributes *i.e.*, location $\mathbf{l}_i$, size $\mathbf{s}_i$, orientation θ_i, class label $\mathbf{c}_i$, and latent shape code $\mathbf{f}_i$. Based on a set of all possible $\mathbf{x}_0$, we propose *DiffuScene*, a denoising diffusion probabilistic model for 3D scene generation. In the forward process, we gradually add noise to $\mathbf{x}_0$ until we obtain a standard Gaussian noise $\mathbf{x}_T$. In the reverse process i.e. generative process, a denoising network iteratively cleans the noisy scene using ancestral sampling. Finally, we use the denoised class labels and shape latent codes to perform shape retrieval, and place object geometries through denoised locations, sizes, and orientations.

trast, we approach scene generation as an unordered object-set diffusion process where we explicitly model the joint distribution of object compositions. Multiple object properties are denoised synchronously, enhancing inter-object relationships and object composition plausibility.

3D Diffusion Models Recently, diffusion models [25, 55–58] have shown impressive visual quality in generative tasks, especially in various applications of 2D image synthesis [1, 9, 12, 25–27, 30, 34, 37, 41, 52, 53] and 3D shape generation [3, 28, 35, 45, 60, 62, 63, 81, 83, 84, 86]. However, diffusion models in the 3D scene receive much less attention. A concurrent work of LEGO-Net [70] aims to predict 2D object locations and orientations, taking the input of a floor plane, object semantics, and geometries. Meanwhile, CommonScene [82] generates 3D indoor scenes conditioned on scene graphs. In contrast, DiffuScene is a scene-generative model that predicts 3D instance properties from random noise, including 3D locations and orientations, semantics, and geometries. Our method is more generic and versatile, which can benefit scene completion and conditioned scene synthesis from multi-modal signals like texts. In terms of implementation, our approach is based on a denoising diffusion model [25], while LEGO-Net uses a Langevin Dynamics scheme based on a score-based method [57]. We use a UNet-1D with attention as a denoiser rather than a transformer in LEGO-Net. These implementation differences contribute to our model's ability to acquire more natural scene arrangements, as evidenced by the discovery of more symmetric pairs in our method.

3. DiffuScene

We introduce DiffuScene, a scene denoising diffusion model aiming at learning the distribution of 3D indoor scenes which includes semantic classes, surface geometries, and placements of multiple objects. Specifically, we assume indoor scenes to be located in a world coordinate system with the origin at the floor center, and each scene $\mathcal{S}$ is a composition of at most N objects $\{\mathbf{o}\}_{i=1}^N$. We represent each scene as an unordered set with N objects, each object in a scene set is defined by its class category $\mathbf{c} \in \mathbb{R}^C$, object size $\mathbf{s} \in \mathbb{R}^3$, location $\ell \in \mathbb{R}^3$, rotation angle around the vertical axis $\theta \in \mathbb{R}$, and shape code $\mathbf{f} \in \mathbb{R}^F$ extracted from object surfaces in the canonical system through a pre-trained shape auto-encoder [77]. Since the number of objects varies across different scenes, we define an additional 'empty' object and pad it into scenes to have a fixed number of objects across scenes. As proposed in [79], we represent the object rotation angle by parametrizing a 2-D vector of cosine and sine values. In summary, each object $\mathbf{o}_i$ is characterized by the concatenation of all attributes, *i.e.* $\mathbf{o}_i = [\ell_i, \mathbf{s}_i, \cos\theta_i, \sin\theta_i, \mathbf{c}_i, \mathbf{f}_i] \in \mathbb{R}^D$, where D is the dimension of concatenated attributes. Based on this representation, we design our denoising diffusion model in Sec. 3.1, which supports many different downstream applications like scene completion, scene re-arrangement, and text-conditioned scene synthesis in Sec. 3.2.

3.1. Object Set Diffusion

An overview of our approach is shown in Fig. 2. We design a denoising diffusion model that employs Gaussian noise corruptions and removals on object attributes to transition between noisy and clean scene distributions.

Diffusion process. The (forward) diffusion process is a pre-defined discrete-time Markov chain in the data space $\mathcal{X}$ spanning all possible scene configurations represented as 2D tensors of fixed size $\mathbf{x} \in \mathbb{R}^{N\times D}$, which are the concatenations of N object properties $\{\mathbf{o}_i\}_{i=1}^{N}$ within a scene $\mathcal{S}$. Given a clean scene configuration $\mathbf{x}_0$ from the underlying distribution $q(\mathbf{x}_0)$, we gradually add Gaussian noise to $\mathbf{x}_0$, obtaining a series of intermediate scene variables $\mathbf{x}_1, ..., \mathbf{x}_T$ with the same dimensionality as $\mathbf{x}_0$, according to a pre-defined, linearly increased noise variance schedule $\beta_1, ..., \beta_T$ (where $\beta_1 < ... < \beta_T$). The joint distribution $q(\mathbf{x}_{1:T}|\mathbf{x}_0)$ of the diffusion process can be expressed as:

$$q(\mathbf{x}_{1:T}|\mathbf{x}_0) := \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1}), \tag{1}$$

where the diffusion step at time t is defined as:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) := \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}). \tag{2}$$

A helpful property of diffusion processes is that we can directly sample $\mathbf{x}_t$ from $\mathbf{x}_0$ via the conditional distribution:

$$q(\mathbf{x}_t|\mathbf{x}_0) := \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I}), \tag{3}$$

where $\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$ where $\alpha_t := 1-\beta_t$, $\bar{\alpha}_t := \prod_{r=1}^{t}\alpha_s$, and ϵ is the noise used to corrupt $\mathbf{x}_t$.

Generative process. The generative (*i.e.* denoising) process is parameterized as a Markov chain of learnable reverse Gaussian transitions. Given a noisy scene from a standard multivariate Gaussian distribution $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ as the initial state, it corrects $\mathbf{x}_t$ to obtain a cleaner version $\mathbf{x}_{t-1}$ at each time step by using a learned Gaussian transition $p_\phi(\mathbf{x}_{t-1}|\mathbf{x}_t)$ which is parameterized by a learnable network ϕ. By repeating this reverse process until the maximum number of steps T, we can reach the final state $\mathbf{x}_0$, the clean scene configuration we aim to obtain. Specifically, the joint distribution of the generative process $p_\phi(\mathbf{x}_{0:T})$ is formulated as:

$$p_\phi(\mathbf{x}_{0:T}) := p(\mathbf{X}_T)\prod_{t=1}^{T} p_\phi(\mathbf{x}_{t-1}|\mathbf{x}_t). \tag{4}$$

$$p_\phi(\mathbf{x}_{t-1}|\mathbf{x}_t) := \mathcal{N}(\mathbf{x}_{t-1}; \mu_\phi(\mathbf{x}_t, t), \mathbf{\Sigma}_\phi(\mathbf{x}_t, t)), \tag{5}$$

where $\mu_\phi(\mathbf{x}_t)$ and $\mathbf{\Sigma}_\phi(\mathbf{x}_t)$ are the predicted mean and covariance of the Gaussian $\mathbf{x}_{t-1}$ by feeding $\mathbf{x}_t$ into the denoising network ϕ. For simplicity, we pre-define the constants of $\Sigma_\phi(\mathbf{x}_t) := \sigma_t := \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$, although Song et al. have shown that learnable covariances can increase generation quality in DDIM [58]. Ho et al. empirically found in DDPM [25] that rather than directly predicting $\mu_\phi(\mathbf{x}_t, t)$, we can synthesize more high-frequent details by estimating the noise $\epsilon_\phi(\mathbf{x}_t, t)$ applied to perturb $\mathbf{x}_t$. Then $\mu_\phi(\mathbf{x}_t)$ can be re-parametrized by subtracting the predicted noise according to Bayes's theorem:

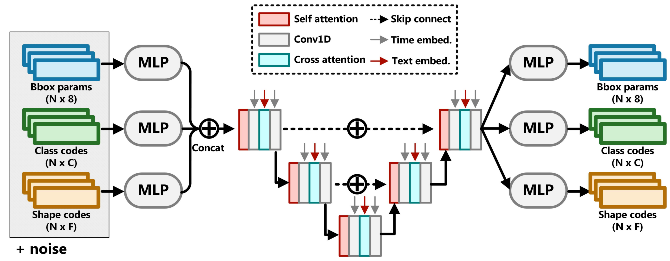


Figure 3. The denoising network architecture takes the attributes of multiple objects (bounding box, object class, geometry code) as input and denoises them using 1D convolutions with skip connections and attention blocks.

$$\mu_\phi(\mathbf{x}_t, t) := \frac{1}{\sqrt{\alpha_t}}(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\phi(\mathbf{x}_t, t)). \tag{6}$$

Denoising network. As shown in Fig. 3, the denoiser in our method is based on 1D convolution with skip connections, where convolution blocks are interleaved with attention blocks [66] to aggregate the features of different objects, exploiting the inter-object relationships and capturing the global scene context.

Training objective. The goal of training the reverse diffusion process is to find optimal denoising network parameters ϕ that can generate natural and plausible scenes. Our training objective is composed of two parts: i) A loss L_{sce} to constrain that the generated object set can approximate the underlying data distribution, and ii) a regularization term L_{iou} to penalize the object intersections. The L_{sce} is derived by maximizing the negative log-likelihood of the last denoised scene $\mathbb{E}[-\log p_\phi(\mathbf{x}_0)]$, which is yet not intractable to optimize directly. Thus, we can instead choose to maximize its variational upper bound:

$$L_{\text{sce}} := \mathbb{E}_q[-\log\frac{p_\phi(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}] \geq \mathbb{E}[-\log p_\phi(\mathbf{x}_0)]. \tag{7}$$

By surrogating variables, we can further simplify L_{sce} as the sum of KL divergence between posterior $p_\phi(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)$ and conditional distribution $q(\mathbf{x}_t|\mathbf{x}_{t-1})$ at each t :

$$L_{\text{sce}} := \mathbb{E}_q[-\log p(\mathbf{x}_T) - \sum_{t=1}^{T}\log\frac{p_\phi(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)}{q(\mathbf{x}_t|\mathbf{x}_{t-1})}], \tag{8}$$

where $-\log p(\mathbf{x}_T)$ is a fixed constant since $\mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I})$. Here, we refer to DDPM [25] for the details of the derivation process. Moreover, we can re-write L_{sce} into a simple and intuitive version that constrains the correct prediction of the corrupted noise on $\mathbf{x}_t$:

$$\begin{aligned} L_{\text{sce}} &:= \mathbb{E}_{\mathbf{x}_0, \epsilon, t}[\|\epsilon - \epsilon_\phi(\mathbf{x}_t, t)\|^2] \\ &:= \mathbb{E}_\phi[\|\epsilon - \epsilon_\phi(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t)\|^2]. \end{aligned} \tag{9}$$

Based on Eq. 6, we can obtain the approximation of clean scene $\tilde{\mathbf{x}}_0^t$. Thus, we can compute L_{iou} as the IoU summation of arbitrary two bounding boxes:

$$L_{\text{iou}} := \sum_{t=1}^{T} 0.1 * \bar{\alpha}_t * \sum_{\mathbf{o}_i, \mathbf{o}_j \in \tilde{\mathbf{x}}_0^t} \text{IoU}(\mathbf{o}_i, \mathbf{o}_j). \tag{10}$$

3.2. Applications

Based on our diffusion model above, we can support various downstream tasks (see Fig. 1) with few modifications.

Scene completion. Assuming a partial scene with $M(\leq N)$ objects, *i.e.* $\mathbf{y} \in \mathbb{R}^{M \times D}$, we utilize the learned scene priors from diffusion models to complement novel $\hat{\mathbf{x}}_0$) into $\mathbf{y}_0$ to obtain a complete object set $\mathbf{x}_0 = (\mathbf{y}, \hat{\mathbf{x}}_0)$. We keep the already known elements and only hallucinate the missing ones through learnable reverse Gaussian transitions q_ϕ conditioning on $\mathbf{y}$. The complemented scene $\hat{\mathbf{x}}_t$ at time step t is generated by:

$$p_\phi(\hat{\mathbf{x}}_{t-1}|\hat{\mathbf{x}}_t) := \mathcal{N}(\mu_\phi(\mathbf{x}_t, t, \mathbf{y}), \sigma_t^2 \mathbf{I}). \tag{11}$$

Scene re-arrangement. Given a set of objects with random spatial positions, we can leverage the priors of our diffusion model to rearrange reasonable object placements by estimating their locations and orientations. We denote the noisy scene initialization as $\hat{\mathbf{x}}_0 = [\hat{\mathbf{u}}_0, \mathbf{v}]$, where $\hat{\mathbf{u}}_0 = \{[\mathbf{l}_i, \cos\theta_i, \sin\theta_i]\}_{i=1}^N$ is the concatenation of N objects' locations and orientations, and $\mathbf{v} = \{[\mathbf{s}_i, \mathbf{c}_i, \mathbf{f}]\}_{i=1}^N$ is the concatenation of N objects' sizes, category classes, and shape codes. The intermediate scenes during the arrangement diffusion process can be expressed as:

$$p_\phi(\hat{\mathbf{u}}_{t-1}|\hat{\mathbf{u}}_t) := \mathcal{N}(\mu_\phi(\hat{\mathbf{u}}_t, t, \mathbf{v}), \sigma_t^2 \mathbf{I}), \tag{12}$$

where we iteratively update the object locations and orientations $\mathbf{u}_t$ via p_ϕ conditioned on $\mathbf{v}$.

Text-conditioned scene synthesis. Given a list of sentences describing the desired object classes and inter-object spatial relationship as conditional inputs, we can employ a pre-trained BERT encoder [11] to extract word embeddings $\mathbf{z} \in \mathbb{R}^{48 \times 768}$, then we utilize cross attention layers to inject the language guidance into the denoising network that predicts out noise via $\epsilon_\phi(\mathbf{x}_t, t, \mathbf{z})$, as depicted in Fig. 3.

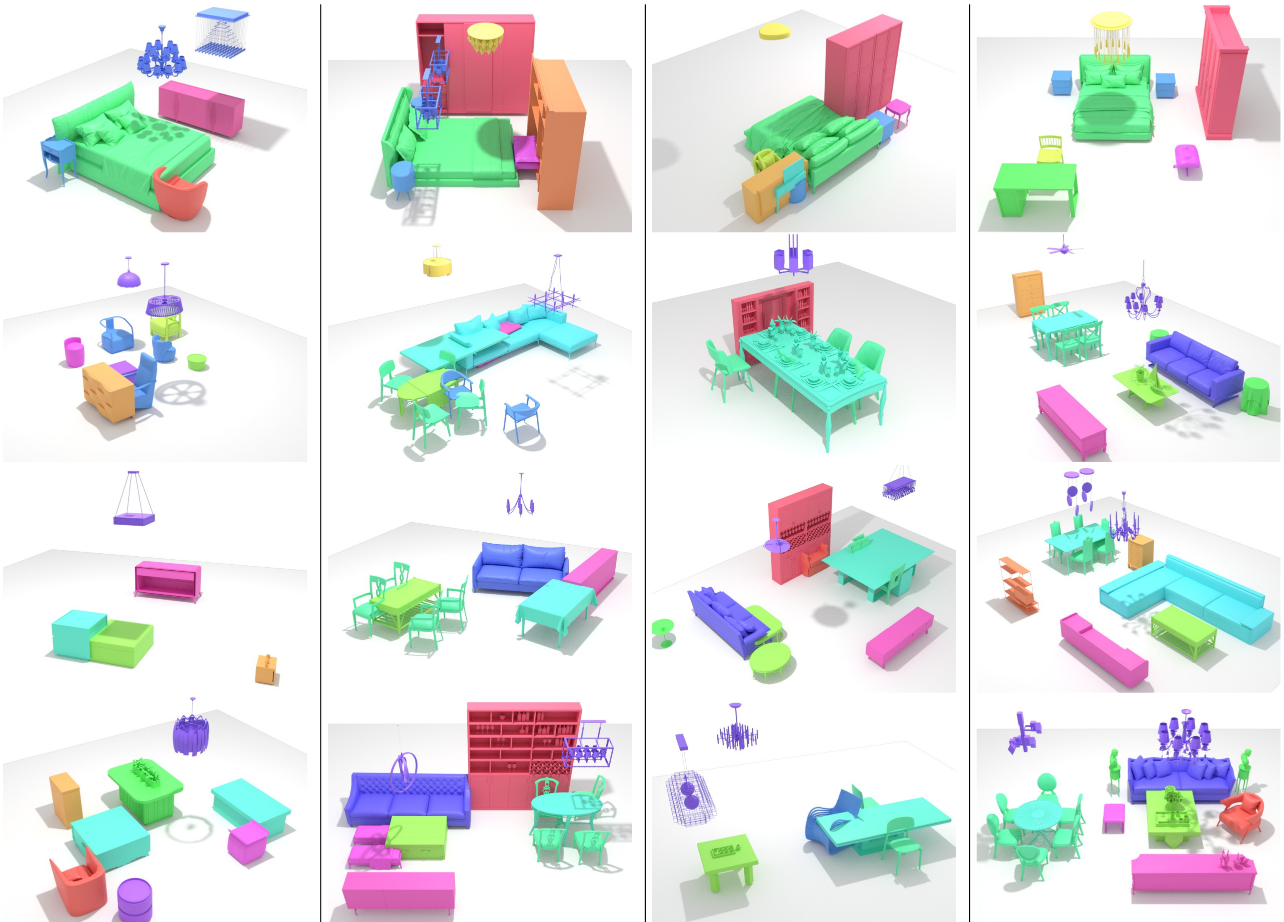
4. Experiments

Datasets For experimental comparisons, we use the large-scale 3D indoor scene dataset 3D-FRONT [17] as the benchmark. 3D-FRONT is a synthetic dataset composed of 6,813 houses with 14,629 rooms, where each room is arranged by a collection of high-quality 3D furniture objects from the 3D-FUTURE dataset [18]. Following ATISS [44], we use three types of indoor rooms for training and evaluation, including 4,041 bedrooms, 900 dining rooms, and 813 living rooms. For each room type, we use 80% of rooms as the training sets, while the remaining are for testing.

Baselines We compare against state-of-the-art scene synthesis approaches using various generative models, including: 1) DepthGAN [76], learning a volumetric generative adversarial network from multi-view semantic-segmented depth maps; 2) Sync2Gen [75], learning a latent space through a variational auto-encoder of scene object arrangements represented by a sequence of 3D object attributes; A Bayesian optimization stage based on the relative attributes prior model further regularized and refined the results. 3) ATISS [44], an autoregressive model to sequentially predict the 3D object bounding box attributes.

Implementation We train our scene diffusion models on different types of indoor rooms respectively. They are trained on a single RTX 3090 with a batch size of 128 for $T = 100,000$ epochs. The learning rate is initialized to $lr = 2\text{e}{-4}$ and then gradually decreases with the decay rate of 0.5 in every 15,000 epochs. For the diffusion processes, we use the default settings from the denoising diffusion probabilistic models (DDPM) [25], where the noise intensity is linearly increased from 0.0001 to 0.02 with 1,000-time steps. During inference, we first use the ancestral sampling strategy to obtain the object properties and then retrieve the most similar CAD model in the 3D-FUTURE [18] for each object based on generated shape codes.

Evaluation Metrics Following previous works [44, 69, 75], we use Fréchet inception distance (FID) [23], Kernel inception distance [2] (KID $\times$ 0.001), scene classification accuracy (SCA), and Category KL divergence (CKL $\times$ 0.01) to measure the plausibility and diversity of 1,000 synthesized scenes. For FID, KID, and SCA, we render the generated and ground-truth scenes into 256$\times$256 semantic maps through top-down orthographic projections, where the texture of each object is uniquely determined by the associate color of its semantic class. We use a unified camera and rendering setting for all methods to ensure fair comparisons. For CKL, we calculate the KL divergence between the semantic class distributions of synthesized scenes and ground-truth scenes. For FID, KID, and CKL, the lower number denotes a better approximation of the data distribution. FID and KID can also manifest the result diversity. For the SCA, a score close to 50% represents that the generated scenes are indistinguishable from real scenes. Additionally, we delve into scene complexity, symmetry, and object interactions using the following metrics: Number of objects (Obj): This metric quantifies the average object count per scene. Number of symmetric object pairs (Sym): It measures the average number of symmetric object pairs in each scene. Pair-wise object bounding box intersection over union (PIoU $\times$ 0.01) assesses the intersection over union between pairwise object bounding boxes. This metric provides insights into object interactions and intersections. The



(a) DepthGAN [76] (b) Sync2Gen [75] (c) ATISS [44] (d) Ours

Figure 4. **Unconditional scene synthesis**. We compare our method with the state-of-the-art by generating from random noises, where our results present higher diversity and better plausibility with fewer penetration issues and more symmetric pairs.

Method	Bedroom				Dining room				Living room			
	FID ↓	KID ↓	SCA %	CKL ↓	FID ↓	KID ↓	SCA %	CKL ↓	FID ↓	KID ↓	SCA %	CKL ↓
DepthGAN [76]	40.15	18.54	96.04	5.04	81.13	50.63	98.59	9.72	88.10	63.81	97.85	7.95
Sync2Gen*	33.59	13.78	87.11	2.67	48.79	12.01	91.43	5.03	47.14	11.42	86.71	1.60
Sync2Gen [75]	31.07	11.21	82.97	2.24	46.05	8.74	88.02	4.96	48.45	12.31	84.57	7.52
ATISS [44]	18.60	1.72	61.71	0.78	38.66	5.62	71.34	0.64	40.83	5.18	72.66	0.69
Ours	**17.21**	**0.70**	**52.15**	**0.35**	**32.60**	**0.72**	**55.50**	**0.22**	**36.18**	**0.88**	**57.81**	**0.21**

Table 1. Quantitative comparisons on the task of **unconditional scene synthesis**. The Sync2Gen* is a variant of Sync2Gen [75] without Bayesian optimization. Note that for the Scene Classification Accuracy (SCA), the score closer to 50% is better.

proximity of Obj, Sym, and PIoU to the ground truth statistics indicates closeness in scene configuration patterns.

4.1. Unconditional Scene Synthesis

Fig. 4 visualizes the qualitative comparisons of different scene synthesis methods. We observe that both DepthGAN [76] and Sync2Gen [75] are vulnerable to object intersections. While ATISS [44] can alleviate the penetration issue by autoregressive scene priors, it cannot always generate reasonable scene results. However, our scene diffusion can synthesize natural and diverse scene arrangements. Tab. 1 presents the quantitative comparisons under various evaluation metrics. Our method consistently outperforms others in all metrics, which clearly demonstrates that our method can generate more diverse and plausible scenes.

Method	Bedroom			Dining			Living		
	Obj	Sym	PIoU	Obj	Sym	PIoU	Obj	Sym	PIoU
DepthGAN	5.12	0.03	0.35	9.64	0.19	0.17	6.70	0.01	0.14
Sync2Gen	6.25	0.85	0.51	8.65	2.85	**0.55**	9.03	2.27	**0.39**
ATISS	5.47	0.33	0.50	11.96	2.75	1.61	10.81	1.42	1.10
Ours	**4.99**	**0.72**	**0.43**	**10.95**	**4.47**	0.65	**11.85**	**3.47**	**0.39**
GT	5.00	0.71	0.43	10.80	4.22	0.48	11.70	3.59	0.30

Table 2. The average of object numbers (Obj.), symmetric object pairs (Sym.), and pairwise box IoU (PIoU) in unconditionally generated scenes. The closer to the statistics of GT, the better.

4.2. Ablation Studies

Method	FID ↓	KID ↓	SCA %	CKL ↓	Obj	Sym	PIoU
C1	29.08	4.59	73.63	0.76	5.10	0.70	0.46
C2	19.78	2.07	54.53	0.69	5.03	0.63	0.38
C3	17.93	1.29	55.14	0.46	5.02	0.64	0.47
C4	18.40	1.55	55.42	0.66	4.97	0.50	0.52
C5	**17.21**	**0.70**	**52.15**	**0.35**	**4.99**	**0.72**	**0.43**

Table 3. Quantitative ablation studies on the task of unconditional scene synthesis on the 3D-FRONT bedrooms.

We conduct detailed ablation studies to verify the effectiveness of each design in our scene diffusion models. The quantitative results are provided in Tab. 3. We refer to the supplementary material for more detailed explanations.

What is the effect of UNet-1D+Attention as the denoiser? (C1 vs. C5) We investigate the different choices of denoising networks. The performances degrade when we use the transformer in DALLE-2 [48].

What is the effect of multiple prediction heads in the denoiser? (C2 vs. C5) In the denoiser, we use three different encoding and prediction heads for respective object properties, *e.g*. bounding box parameter, semantic class labels, and geometry codes. Multiple diffusion heads with individual losses for attributes can prevent biasing towards one attribute in a single encoding and prediction head.

What is the effect of the IoU loss? (C3 vs. C5) The IoU loss can penalize object intersections, promote more reasonable placements, and preserve symmetries. This is reflected by consistent improvement in each metric.

What is the effect of geometry feature diffusion? (C4 vs. C5) The geometry feature enables better capture of symmetric placements and semantically coherent arrangements. Fig. 5 shows that our model can find symmetric nightstands by beds due to the geometry awareness of the diffusion process and shape retrieval. This is supported by Sym: 0.72 (w/ shape diffusion) vs. 0.50 (w/o shape diffusion) in Tab. 3. More plausible synthesis results improve FID, KID, and SCA. Besides, the decrease in CKL can manifest that the

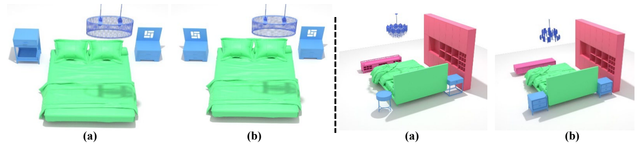


Figure 5. (b) w/ shape diffusion captures symmetries vs. (a) w/o. The shape latent diffusion promotes symmetry discovery.

joint diffusion of geometry code and object layout can learn more similar object class distribution.

Can DiffuScene generate novel scenes? In Fig. 6, We retrieve the three most similar training scenes for a generated scene using the Chamfer distance. Our result reveals unique object compositions, highlighting our method's ability to generate novel scenes rather than reproducing training data.

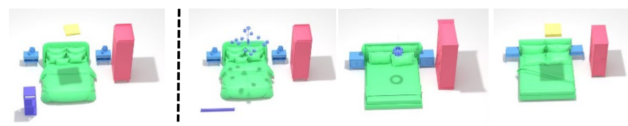


Figure 6. **Left:** Ours. **Right:** top-3 nearest scenes in the train set.

4.3. Applications

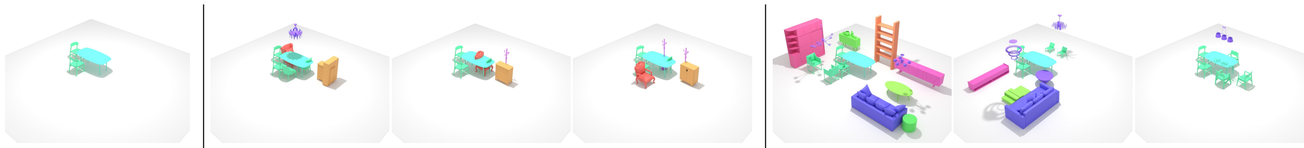
Scene Completion We compare against ATISS [44] on the task of scene completion. As shown in Fig. 7, our method can produce more diverse completion results with high fidelity, fewer intersections, and more symmetries.

Room	Method	FID ↓	KID ↓	#Sym.	PIoU
Bedroom	ATISS	27.14	1.56	0.01	0.84
	LEGO	23.73	4.70	0.45	0.89
	Ours	**22.16**	**1.02**	**0.70**	**0.61**
Living room	ATISS	44.94	5.41	1.42	1.73
	LEGO	45.40	9.57	2.50	1.63
	Ours	**41.15**	**2.24**	**3.69**	**0.95**

Table 4. Quantitative comparisons on the task of **scene arrangement** on the 3D-FRONT bedrooms and dining rooms. Given a collection of objects as inputs, we predict their locations and orientations to obtain object placements.

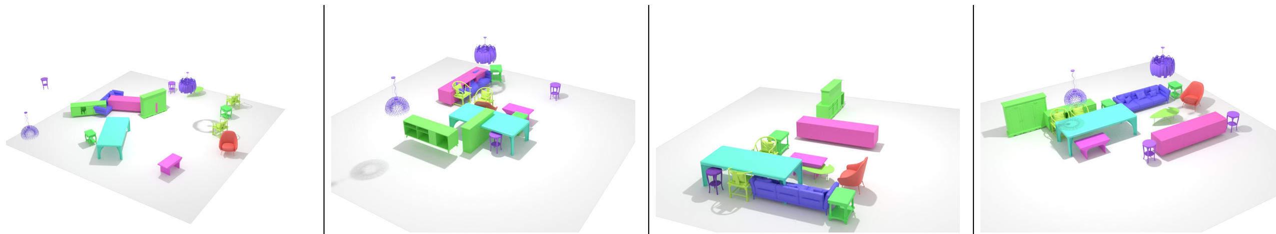
Scene Re-arrangement We also conduct comparisons with ATISS [44] on the application of scene re-arrangement. As depicted in Fig. 8, our method generates more favorable object placements and more symmetric relations compared to ATISS [44] and LEGO [70].

Text-conditioned Scene Synthesis Given a text prompt describing a partial scene configuration, we aim to synthesize a whole scene satisfying the input. We conduct a perceptual user study for the text-conditioned scene synthesis. Given a text prompt and a ground-truth scene as a reference, we ask the attendance two questions for each pair



(a) Partial Scenes (b) ATISS [44] (c) Ours

Figure 7. Scene completion from partial scenes with only 3 objects given as inputs. Compared to ATISS, our diffusion-based method produces more diverse completion results with higher fidelity, fewer intersections, and more symmetries.



(a) Noisy Scenes (b) ATISS [44] (c) LEGO [70] (d) Ours

Figure 8. Scene re-arrangements of collections of random objects. Compared to ATISS and LEGO, our method generates more favorable object placements with more symmetric pairs.

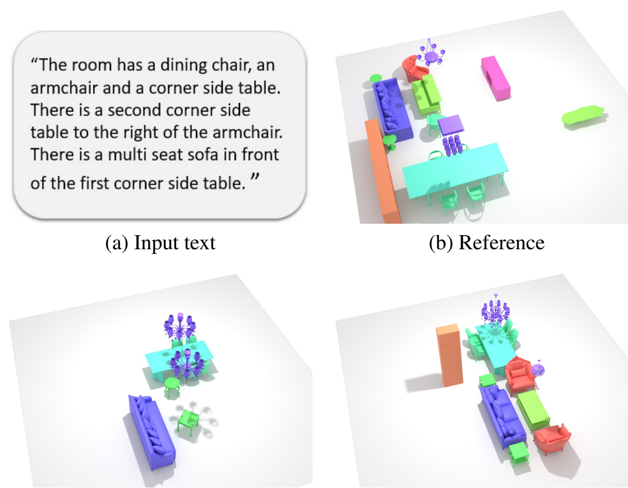
of results from ATISS and ours: which of the synthesized scenes is closely matched with the input text, and which one is more realistic and reasonable. We collect the answers of 225 scenes from 45 users. 62% of users prefer our method to ATISS in realism. 55% of users are in favor of us in the matching score. This illustrates that our text-conditioned model generates more realistic scenes while capturing more accurate object relationships described in the text prompt. Please refer to the supplementary material for more details.

4.4. Limitations

Although we have shown impressive scene synthesis results, our method still has some limitations. First, the shape retrieval searches the closest shape with the same semantics within defined classes of CAD models. Thus, the retrieved model could fail to match the style of the desired scene. Second, the object textures are from the provided 3D CAD model dataset via shape retrieval. An interesting direction is to integrate texture diffusion into our model. Third, we only consider single-room generation and train our model on a specific room type. Thus, our method cannot synthesize large-scale scenes with multiple rooms. Finally, we rely on 3D labeled scenes to drive the learning of scene diffusion. Leveraging scene datasets with only 2D labels to learn scene diffusion priors is also a promising direction. We leave these mentioned limitations as our future efforts.

5. Conclusion

In this work, we introduced DiffuScene, a novel method for generative indoor scene synthesis based on a denoising diffusion probabilistic model that learns holistic scene configuration priors in the full set diffusion process of object semantics, bounding boxes, and geometry features.We applied our method to several downstream applications, namely scene completion, scene re-arrangement, and text-conditioned scene synthesis. Compared to prior state-of-the-art methods. Our approach can synthesize more plausible and diverse indoor scenes as has been measured by different metrics and confirmed in a user study. Our method is an important piece in the puzzle of 3D generative modeling and we hope that it will inspire research in denoising diffusion-based 3D synthesis.



(c) ATISS [44] (d) Ours

Figure 9. Text-conditioned scene synthesis. The input text only describes a partial scene configuration. Our method generates a more plausible scene matching the input text.

Acknowledgement. This work is supported by a TUM-IAS Rudolf Mößbauer Fellowship, the ERC Starting Grant Scan2CAD (804724), and Sony Semiconductor Solutions.

References

[1] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18208–18218, 2022. 3

[2] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018. 5

[3] Wei Cao, Chang Luo, Biao Zhang, Matthias Nießner, and Jiapeng Tang. Motion2vecsets: 4d latent vector set diffusion for non-rigid shape reconstruction and tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 3

[4] Angel Chang, Manolis Savva, and Christopher D Manning. Learning spatial knowledge for text to 3d scene generation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 2028–2038, 2014. 2

[5] Angel X Chang, Mihail Eric, Manolis Savva, and Christopher D Manning. Sceneseer: 3d scene design with natural language. *arXiv preprint arXiv:1703.00050*, 2017. 2

[6] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018. 2

[7] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *CVPR*, 2019. 2

[8] Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, pages 628–644. Springer, 2016. 2

[9] Yuren Cong, Mengmeng Xu, christian simon, Shoufa Chen, Jiawei Ren, Yanping Xie, Juan-Manuel Perez-Rua, Bodo Rosenhahn, Tao Xiang, and Sen He. FLATTEN: optical FLow-guided ATTENtion for consistent text-to-video editing. In *The Twelfth International Conference on Learning Representations*, 2024. 3

[10] Angela Dai and Matthias Nießner. 3dmv: Joint 3d-multi-view prediction for 3d semantic scene segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 452–468, 2018. 2

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 5

[12] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 3

[13] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 12873–12883, 2021. 2

[14] Matthew Fisher and Pat Hanrahan. Context-based search for 3d models. In *ACM SIGGRAPH Asia 2010 papers*, pages 1–10. 2010. 2

[15] Matthew Fisher, Daniel Ritchie, Manolis Savva, Thomas Funkhouser, and Pat Hanrahan. Example-based synthesis of 3d object arrangements. *ACM Transactions on Graphics (TOG)*, 31(6):1–11, 2012. 2

[16] Matthew Fisher, Manolis Savva, Yangyan Li, Pat Hanrahan, and Matthias Nießner. Activity-centric scene synthesis for functional 3d scene modeling. *ACM Transactions on Graphics (TOG)*, 34(6):1–13, 2015. 2

[17] Huan Fu, Bowen Cai, Lin Gao, Ling-Xiao Zhang, Jiaming Wang, Cao Li, Qixun Zeng, Chengyue Sun, Rongfei Jia, Binqiang Zhao, et al. 3d-front: 3d furnished rooms with layouts and semantics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10933–10942, 2021. 5

[18] Huan Fu, Rongfei Jia, Lin Gao, Mingming Gong, Binqiang Zhao, Steve Maybank, and Dacheng Tao. 3d-future: 3d furniture shape with texture. *International Journal of Computer Vision*, 129:3313–3337, 2021. 5

[19] Qiang Fu, Xiaowu Chen, Xiaotian Wang, Sijia Wen, Bin Zhou, and Hongbo Fu. Adaptive synthesis of indoor scenes via activity-associated object relation graphs. *ACM Transactions on Graphics (TOG)*, 36(6):1–13, 2017. 2

[20] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2

[21] Alex Graves. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*, 2013. 2

[22] Xiaoguang Han, Zhaoxuan Zhang, Dong Du, Mingdai Yang, Jingming Yu, Pan Pan, Xin Yang, Ligang Liu, Zixiang Xiong, and Shuguang Cui. Deep reinforcement learning of volume-guided progressive view inpainting for 3d point scene completion from a single depth image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 234–243, 2019. 2

[23] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 5

[24] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 2

[25] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 2, 3, 4, 5

[26] 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.

[27] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *J. Mach. Learn. Res.*, 23:47–1, 2022. 3

[28] Ka-Hei Hui, Ruihui Li, Jingyu Hu, and Chi-Wing Fu. Neural wavelet-domain diffusion for 3d shape generation. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–9, 2022. 3

[29] Yun Jiang, Marcus Lim, and Ashutosh Saxena. Learning object arrangements in 3d scenes using human context. *arXiv preprint arXiv:1206.6462*, 2012. 2

[30] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2426–2435, 2022. 3

[31] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2

[32] Jiabao Lei, Jiapeng Tang, and Kui Jia. Rgbd2: Generative scene synthesis via incremental view inpainting using rgbd diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8422–8434, 2023. 2

[33] Manyi Li, Akshay Gadi Patil, Kai Xu, Siddhartha Chaudhuri, Owais Khan, Ariel Shamir, Changhe Tu, Baoquan Chen, Daniel Cohen-Or, and Hao Zhang. Grains: Generative recursive autoencoders for indoor scenes. *ACM Transactions on Graphics (TOG)*, 38(2):1–16, 2019. 2

[34] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11461–11471, 2022. 3

[35] Shitong Luo and Wei Hu. Diffusion probabilistic models for 3d point cloud generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2837–2845, 2021. 3

[36] Rui Ma, Honghua Li, Changqing Zou, Zicheng Liao, Xin Tong, and Hao Zhang. Action-driven 3d indoor scene evolution. *ACM Trans. Graph.*, 35(6):173–1, 2016. 2

[37] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2021. 3

[38] Paul Merrell, Eric Schkufza, Zeyang Li, Maneesh Agrawala, and Vladlen Koltun. Interactive furniture layout using interior design guidelines. *ACM transactions on graphics (TOG)*, 30(4):1–10, 2011. 2

[39] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *CVPR*, 2019. 2

[40] 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. 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. *arXiv preprint arXiv:2112.10741*, 2021. 3

[42] Yinyu Nie, Angela Dai, Xiaoguang Han, and Matthias Nießner. Learning 3d scene priors with 2d supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 792–802, 2023. 2

[43] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *CVPR*, 2019. 2

[44] 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, 5, 6, 7, 8

[45] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 3

[46] Pulak Purkait, Christopher Zach, and Ian Reid. Sg-vae: Scene grammar variational autoencoder to generate new indoor scenes. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16*, pages 155–171. Springer, 2020. 2

[47] Siyuan Qi, Yixin Zhu, Siyuan Huang, Chenfanfu Jiang, and Song-Chun Zhu. Human-centric indoor scene synthesis using stochastic grammar. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5899–5908, 2018. 2

[48] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 7

[49] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019. 2

[50] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR, 2015. 2

[51] Daniel Ritchie, Kai Wang, and Yu-an Lin. Fast and flexible indoor scene synthesis via deep convolutional generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6182–6190, 2019. 2

[52] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 3

[53] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 3

[54] Manolis Savva, Angel X Chang, and Maneesh Agrawala. Scenesuggest: Context-driven 3d scene design. *arXiv preprint arXiv:1703.00061*, 2017. 2

[55] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015. 3

[56] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[57] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019. 3

[58] Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. *Advances in neural information processing systems*, 33:12438–12448, 2020. 3, 4

[59] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 2

[60] Jiapeng Tang, Xiaoguang Han, Junyi Pan, Kui Jia, and Xin Tong. A skeleton-bridged deep learning approach for generating meshes of complex topologies from single rgb images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4541–4550, 2019. 3

[61] Jiapeng Tang, Jiabao Lei, Dan Xu, Feiying Ma, Kui Jia, and Lei Zhang. Sa-convonet: Sign-agnostic optimization of convolutional occupancy networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6504–6513, 2021. 2

[62] Jiapeng Tang, Lev Markhasin, Bi Wang, Justus Thies, and Matthias Nießner. Neural shape deformation priors. In *Advances in Neural Information Processing Systems*, 2022. 3

[63] Jiapeng Tang, Angela Dai, Yinyu Nie, Lev Markhasin, Justus Thies, and Matthias Niessner. Dphms: Diffusion parametric head models for depth-based tracking. 2024. 3

[64] Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with pixelcnn decoders. *Advances in neural information processing systems*, 29, 2016. 2

[65] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 2

[66] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2, 4

[67] 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

[68] Kai Wang, Yu-An Lin, Ben Weissmann, Manolis Savva, Angel X Chang, and Daniel Ritchie. Planit: Planning and instantiating indoor scenes with relation graph and spatial prior networks. *ACM Transactions on Graphics (TOG)*, 38(4):1–15, 2019.

[69] Xinpeng Wang, Chandan Yeshwanth, and Matthias Nießner. Sceneformer: Indoor scene generation with transformers. In *2021 International Conference on 3D Vision (3DV)*, pages 106–115. IEEE, 2021. 2, 5

[70] Qiuhong Anna Wei, Sijie Ding, Jeong Joon Park, Rahul Sajnani, Adrien Poulenard, Srinath Sridhar, and Leonidas Guibas. Lego-net: Learning regular rearrangements of objects in rooms. *arXiv preprint arXiv:2301.09629*, 2023. 3, 7, 8

[71] Jiajun Wu, Chengkai Zhang, Tianfan Xue, Bill Freeman, and Josh Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. *Advances in neural information processing systems*, 29, 2016. 2

[72] Mingdong Wu, Fangwei Zhong, Yulong Xia, and Hao Dong. Targf: Learning target gradient field for object rearrangement. *arXiv preprint arXiv:2209.00853*, 2022. 2

[73] Zhisheng Xiao, Karsten Kreis, and Arash Vahdat. Tackling the generative learning trilemma with denoising diffusion gans. *arXiv preprint arXiv:2112.07804*, 2021. 2

[74] Kun Xu, Kang Chen, Hongbo Fu, Wei-Lun Sun, and Shi-Min Hu. Sketch2scene: Sketch-based co-retrieval and co-placement of 3d models. *ACM Transactions on Graphics (TOG)*, 32(4):1–15, 2013. 2

[75] Haitao Yang, Zaiwei Zhang, Siming Yan, Haibin Huang, Chongyang Ma, Yi Zheng, Chandrajit Bajaj, and Qixing Huang. Scene synthesis via uncertainty-driven attribute synchronization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5630–5640, 2021. 2, 5, 6

[76] Ming-Jia Yang, Yu-Xiao Guo, Bin Zhou, and Xin Tong. Indoor scene generation from a collection of semantic-segmented depth images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15203–15212, 2021. 2, 5, 6

[77] Yaoqing Yang, Chen Feng, Yiru Shen, and Dong Tian. Foldingnet: Point cloud auto-encoder via deep grid deformation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 206–215, 2018. 3

[78] Yi-Ting Yeh, Lingfeng Yang, Matthew Watson, Noah D Goodman, and Pat Hanrahan. Synthesizing open worlds with constraints using locally annealed reversible jump mcmc. *ACM Transactions on Graphics (TOG)*, 31(4):1–11, 2012. 2

[79] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3d object detection and tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11784–11793, 2021. 3

[80] 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)-Proceedings of ACM SIGGRAPH 2011, v. 30,(4), July 2011, article no. 86*, 30(4), 2011. 2

[81] Xiaohui Zeng, Arash Vahdat, Francis Williams, Zan Gojcic, Or Litany, Sanja Fidler, and Karsten Kreis. Lion: Latent point diffusion models for 3d shape generation. *arXiv preprint arXiv:2210.06978*, 2022. 3

[82] Guangyao Zhai, Evin Pınar Örnek, Shun-Cheng Wu, Yan Di, Federico Tombari, Nassir Navab, and Benjamin Busam. Commonscenes: Generating commonsense 3d indoor scenes with scene graphs. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[83] Biao Zhang and Peter Wonka. Functional diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 3

[84] Biao Zhang, Jiapeng Tang, Matthias Niessner, and Peter Wonka. 3dshape2vecset: A 3d shape representation for neural fields and generative diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4):1–16, 2023. 3

[85] Zaiwei Zhang, Zhenpei Yang, Chongyang Ma, Linjie Luo, Alexander Huth, Etienne Vouga, and Qixing Huang. Deep generative modeling for scene synthesis via hybrid representations. *ACM Transactions on Graphics (TOG)*, 39(2):1–21, 2020. 2

[86] Linqi Zhou, Yilun Du, and Jiajun Wu. 3d shape generation and completion through point-voxel diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5826–5835, 2021. 3