arXiv:2110.03675v1 [cs.CV] 7 Oct 2021

ATISS: Autoregressive Transformers for Indoor Scene Synthesis

Despoina Paschalidou[1,3,4] **Amlan Kar**[4,5,6] **Maria Shugrina**[4] **Karsten Kreis**[4]

Andreas Geiger[1,2,3] **Sanja Fidler**[4,5,6]

[1]Max Planck Institute for Intelligent Systems Tübingen [2]University of Tübingen
[3]Max Planck ETH Center for Learning Systems
[4]NVIDIA [5]University of Toronto [6]Vector Institute
{firstname.lastname}@tue.mpg.de {amlank, mshugrina, kkreis, sfidler}@nvidia.com

Abstract

The ability to synthesize realistic and diverse indoor furniture layouts automatically or based on partial input, unlocks many applications, from better interactive 3D tools to data synthesis for training and simulation. In this paper, we present ATISS, a novel autoregressive transformer architecture for creating diverse and plausible synthetic indoor environments, given only the room type and its floor plan. In contrast to prior work, which poses scene synthesis as sequence generation, our model generates rooms as unordered sets of objects. We argue that this formulation is more natural, as it makes ATISS generally useful beyond fully automatic room layout synthesis. For example, the same trained model can be used in interactive applications for general scene completion, partial room rearrangement with any objects specified by the user, as well as object suggestions for any partial room. To enable this, our model leverages the permutation equivariance of the transformer when conditioning on the partial scene, and is trained to be permutation-invariant across object orderings. Our model is trained end-to-end as an autoregressive generative model using only labeled 3D bounding boxes as supervision. Evaluations on four room types in the 3D-FRONT dataset demonstrate that our model consistently generates plausible room layouts that are more realistic than existing methods. In addition, it has fewer parameters, is simpler to implement and train and runs up to 8x faster than existing methods.

1 Introduction

Generating synthetic 3D content that is both realistic and diverse is a long-standing problem in computer vision and graphics. In the last decade, there has been increased demand for tools that automate the creation of 3D artificial environments for applications like video games and AR/VR, as well as general 3D content creation [76, 19, 44, 5, 77]. These tools can also synthesize data to train computer vision models, avoiding expensive and laborious annotations. Generative models [34, 24, 15, 35, 71] have demonstrated impressive results on synthesizing photorealistic images [8, 1, 29, 9, 30] and intelligible text [60, 2], and are beginning to be adopted for the generation of 3D environments.

Recent works proposed to solve the scene synthesis task by incorporating procedural modeling techniques [59, 57, 28, 11] or by generating scene graphs with generative models [41, 72, 80, 42, 58, 79, 78, 32, 14]. Procedural modeling requires specifying a set of rules for the scene formation process, but acquiring these rules is a time-consuming task, requiring skills of experienced artists. Similarly, graph-based approaches require scene graph annotations, which may be laborious to obtain.

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

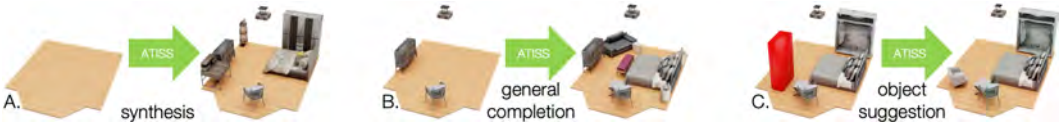


Figure 1: Motivation In addition to fully automatic layout synthesis (A), our formulation in terms of unordered sets of objects allows our model to be used for novel interactive applications with versatile user control: scene completion given any number of existing furniture pieces of any class pinned to a specific location by the user (B), and object suggestions with user-provided constraints (object centroid constraint shown in red) (C).

Another line of research utilizes CNN-based [73, 61] and transformer-based [74] architectures to generate rooms by autoregressively selecting and placing objects in a scene, i.e. one after the other. These approaches represent scenes as ordered sequences of objects. Typically, the ordering is defined using the spatial arrangement of objects in a room (e.g. left-to-right) [27] or the object class frequency (e.g. most to least probable) [61, 74]. Such orderings impose unnatural constraints on the scene generation process, inhibiting practical applications. For example, in [61, 74], which order objects by class frequency, the probability of a bed (more common) appearing after an ottoman (less common) in the training set is zero. As a result, these methods cannot generate more common objects after less common objects, which makes them impractical for interactive tasks like general room completion and partial room re-arrangement, where input is unconstrained (e.g. Fig.1B).

To address these limitations, we pose scene synthesis as an unordered set generation problem and introduce ATISS, a novel autoregressive transformer architecture to model this process. Given a room type (e.g. bedroom, living room) and its shape, our model generates meaningful furniture arrangements by sequentially placing objects in a permutation-invariant fashion. We train ATISS to maximize the log-likelihood of all possible permutations of object arrangements in a collection of training scenes, labeled only with object classes and $3D$ bounding boxes, which are easier to obtain, than costly support relationship [72] or scene graph annotations [41]. Unlike existing works [73, 61, 74], we propose the first model to perform scene synthesis as an autoregressive *set generation* task. ATISS is significantly simpler to implement and train, requires fewer parameters and is up to $8\times$ faster at run-time than the fastest available baseline [74]. Furthermore, we demonstrate that our model generates more plausible object arrangments without any post-processing on the predicted layout. Our formulation allows applying a single trained model to automatic layout synthesis and to a number of interactive scenarios with versatile user input (Fig.1), such as automatic placement of user-provided objects, object suggestion with user-provided constraints, and room completion. Code and data are publicaly available at https://nv-tlabs.github.io/ATISS.

2 Related Work

In this section, we discuss the most relevant literature on interior scene synthesis, as well as transformer architectures [71] in the context of generative modeling.

Procedural Modeling with Grammars: Procedural modeling describes methods that recursively apply a set of functions for content synthesis. Grammars are a formal instantiation of this idea and have been used for modeling 3D structures such as plants [66], buildings and cities [47, 50], indoor [59] and outdoor [57] scenes. [66] employed reversible-jump MCMC to control the output of stochastic context-free grammars. Meta-Sim [28] learned a model that modifies attributes of scene graphs sampled from a known probabilistic context-free grammar to match visual statistics between generated and real data. [11] extended this model to also learn to sample from the grammar, allowing context dependent relationships to be learnt. Concurrently, [58] employed Grammar-VAE [38] to generate scenes using a scene grammar generated from annotated data. In contrast, our model implicitly encapsulates inter-object relationships, without having to impose hand-crafted constraints.
Graph-based Scene Synthesis: Representing scenes as graphs has been extensively studied in literature [41, 72, 80, 42, 58, 79, 78, 32, 14]. Zhou et al. [80] introduced a neural message passing algorithm for scene graphs that predicts the category of the next object to be placed at a specific location. Similarly, [41, 79, 58, 42] utilized a VAE [34] to synthesize 3D scenes as parse trees [58], adjacency matrices [79], scene graphs [42] and scene hierarchies [41]. Concurrently, [72, 78] adopted a two-stage generation process that disentangles planning the scene layout from instantiating the scene based on this plan. Note that graph-based models require supervision either in the form of relation graphs [72, 78, 42] or scene hierarchies [41]. In contrast, ATISS infers functional and spatial relations between objects directly from data labeled only with object classes and 3D bounding boxes.

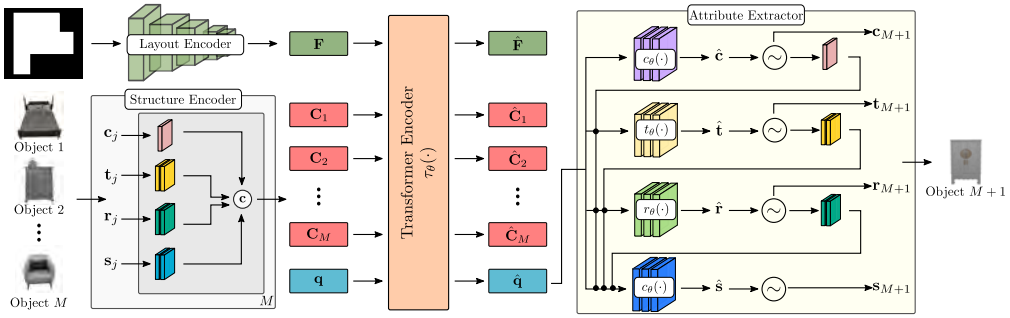


Figure 2: Method Overview. Starting from a scene with M objects and a floor layout, the **layout encoder** maps the floor into a feature representation $\mathbf{F}$ and the **structure encoder** maps the objects into a context embedding $\mathbf{C} = \{\mathbf{C}_j\}_{j=1}^{M}$. The floor layout feature $\mathbf{F}$, the context embedding $\mathbf{C}$ and a learnable query vector $\mathbf{q}$ are then passed to the **transformer encoder** that predicts $\hat{\mathbf{q}}$. Using $\hat{\mathbf{q}}$ the **attribute extractor** autoregressively predicts the attribute distributions that are used to sample the attributes for the next object to be generated.

Autoregressive Scene Synthesis: Closely related to our work are autoregressive indoor scene generation models [73, 61, 74]. Ritchie et al. [61] introduced a CNN-based architecture that operates on a top-down image-based representation of a scene and inserts objects in it sequentially by predicting their category, location, orientation and size with separate network modules. [61] requires supervision in the form of 2D bounding boxes as well as auxiliary supervision such as depth maps and object segmentation masks. In concurrent work, Wang et al. [74] introduced SceneFormer, a series of transformers that autoregressively add objects in a scene similar to [61]. Both [61, 74] use separate models to generate object attributes (e.g. category, location) that are trained independently and represent scenes as ordered sequences of objects, ordered by the category frequency. In contrast, we propose a simpler architecture that consists of a single model trained end-to-end to predict all attributes. We provide experimental evidence that our model generates more realistic object arrangements while being significantly faster. While [61, 74] assume a fixed ordering of the objects in each scene, our model does not impose any constraint on the ordering of objects. Instead, during training, we enforce that our model generates objects with all orderings, in a permutation invariant fashion. This allows us to represent scenes as unordered sets of objects and perform various interactive tasks such as rearranging any object in a room or suggesting new objects given any room.

Transformers for Set Generation: Transformer models [71] demonstrated impressive results on various tasks such as machine translation [64, 49], language-modeling [2, 12], object detection [40, 3, 81], image recognition [16, 67], semantic segmentation [75] as well as on image [52, 31, 6, 17, 68] and music [13] generation tasks. While there are works [39, 36] that utilize the permutation equivariance property of transformers for unordered set processing and prediction, existing generative models with transformers assume ordered sequences [2, 6, 7] even when there exists no natural order such as for pointclouds [48] and objects in a scene [74]. Instead, we introduce an autoregressive transformer for unordered set generation that enforces that the probability of adding a new element in the set is invariant to the order of the elements already in the set. We show that for the scene synthesis task, our model outperforms transformers that consider ordered sets of elements in every metric.

3 Method

Given an empty or a partially complete room of a specific type (e.g. bedroom) together with its shape, as a top-down orthographic projection of its floor, we want to learn a generative model that populates the room with objects, whose functional composition and spatial arrangement is plausible. To this end, we propose an autoregressive model that represents scenes as *unordered sets of objects* (Sec. 3.1) and describe our implementation using a transformer network (Sec. 3.2). Finally, we analyse the training and inference details of our method (Sec. 3.3).

3.1 Autoregressive Set Generation

Let $\mathcal{X} = \{\mathcal{X}_1, \ldots, \mathcal{X}_N\}$ denote a collection of scenes where each $\mathcal{X}_i = \left(\mathcal{O}_i, \mathbf{F}^i\right)$ comprises the unordered set of objects in the scene $\mathcal{O}_i = \{o_j^i\}_{j=1}^{M}$ and its floor layout $\mathbf{F}^i$. To compute the likelihood of generating $\mathcal{O}_i$ we need to accumulate the likelihood of generating $\{o_j^i\}_{j=1}^{M}$ autoregressively in *any*

order. This is formally written as

$$p_\theta(\mathcal{O}_i|\mathbf{F}^i) = \sum_{\hat{\mathcal{O}}\in\pi(\mathcal{O}_i)} \prod_{j\in\hat{\mathcal{O}}} p_\theta(o^i_j \mid o^i_{<j}, \mathbf{F}^i), \tag{1}$$

where $p_\theta(o^i_j \mid o^i_{<j}, \mathbf{F}^i)$ is the probability of the j-th object, conditioned on the previously generated objects and the floor layout, and $\pi(\cdot)$ is a permutation function that computes the set of permutations of all objects in the scene. As a result, the log-likelihood of the whole collection $\mathcal{X}$ is

$$\log p_\theta(\mathcal{X}) = \sum_{i=1}^{N} \log \left(\sum_{\hat{\mathcal{O}}\in\pi(\mathcal{O}_i)} \prod_{j\in\hat{\mathcal{O}}} p_\theta(o^i_j \mid o^i_{<j}, \mathbf{F}^i) \right). \tag{2}$$

However, training our generative model to maximize the log-likelihood of (2) poses two problems: (a) the summation over all permutations is intractable and (b) (2) does not ensure that all orderings will have high probability. The second problem is crucial because we want our generative model to be able to complete *any partial set* in a plausible way, namely we want any generation order to have high probability. To this end, instead of maximizing (2), we maximize the likelihood of generating a scene in all possible orderings, $\hat{p}_\theta(\cdot)$, which is defined as

$$\log \hat{p}_\theta(\mathcal{X}) = \sum_{i=1}^{N} \log \left(\prod_{\hat{\mathcal{O}}\in\pi(\mathcal{O}_i)} \prod_{j\in\hat{\mathcal{O}}} p_\theta(o^i_j \mid o^i_{<j}, \mathbf{F}^i) \right) = \sum_{i=1}^{N} \sum_{\hat{\mathcal{O}}\in\pi(\mathcal{O}_i)} \sum_{j\in\hat{\mathcal{O}}} \log p_\theta(o^i_j \mid o^i_{<j}, \mathbf{F}^i). \tag{3}$$

Note that training our generative model with (3) allows us to approximate the summation over all permutations using Monte Carlo sampling thus solving both problems of (2).

Modelling Object Attributes: We represent objects in a scene as labeled 3D bounding boxes and model them with four random variables that describe their category, size, orientation and location, $o_j = \{\mathbf{c}_j, \mathbf{s}_j, \mathbf{t}_j, \mathbf{r}_j\}$. The category $\mathbf{c}_j$ is modeled using a categorical variable over the total number of object categories C in the dataset. For the size $\mathbf{s}_j \in \mathbb{R}^3$, the location $\mathbf{t}_j \in \mathbb{R}^3$ and the orientation $\mathbf{r}_j \in \mathbb{R}^1$, we follow [62, 70] and model them with mixture of logistics distributions

$$\mathbf{s}_j \sim \sum_{k=1}^{K} \pi^s_k \text{logistic}(\mu^s_k, \sigma^s_k) \quad \mathbf{t}_j \sim \sum_{k=1}^{K} \pi^t_k \text{logistic}(\mu^t_k, \sigma^t_k) \quad \mathbf{r}_j \sim \sum_{k=1}^{K} \pi^r_k \text{logistic}(\mu^r_k, \sigma^r_k) \tag{4}$$

where π^s_k, μ^s_k and σ^s_k denote the weight, mean and variance of the k-th logistic distribution used for modeling the size. Similarly, π^t_k, μ^t_k and σ^t_k and π^r_k, μ^r_k ans σ^r_k refer to the weight, mean and variance of the k-th logistic of the location and orientation, respectively. In our setup, the orientation is the angle of rotation around the up vector and the location is the 3D centroid of the bounding box.

Similar to prior work [61, 74], we predict the object attributes in an autoregressive manner: object category first, followed by position, orientation and size as follows:

$$p_\theta(o_j \mid o_{<j}, \mathbf{F}) = p_\theta(\mathbf{c}_j|o_{<j}, \mathbf{F}) p_\theta(\mathbf{t}_j|\mathbf{c}_j, o_{<j}, \mathbf{F}) p_\theta(\mathbf{r}_j|\mathbf{c}_j, \mathbf{t}_j, o_{<j}, \mathbf{F}) p_\theta(\mathbf{s}_j|\mathbf{c}_j, \mathbf{t}_j, \mathbf{r}_j, o_{<j}, \mathbf{F}). \tag{5}$$

This is a natural choice, since we want our model to consider the object class before reasoning about the size and the position of an object. To avoid notation clutter, we omit the scene index i from (5).

3.2 Network Architecture

The input to our model is a collection of scenes in the form of 3D labeled bounding boxes with their corresponding room shape. Our network consists of four main components: (i) the *layout encoder* that maps the room shape to a global feature representation $\mathbf{F}$, (ii) the *structure encoder* h_θ that maps the M objects in the scene into per-object context embeddings $\mathbf{C} = \{\mathbf{C}_j\}_{j=1}^{M}$, (iii) the *transformer encoder* τ_θ that takes $\mathbf{F}$, $\mathbf{C}$ and a query embedding $\mathbf{q}$ and predicts the features $\hat{\mathbf{q}}$ for the next object to be generated and (iv) the *attribute extractor* that predicts the attributes of the next object. Our model is illustrated in Fig. 2. The layout encoder is simply a ResNet-18 [25] that extracts a feature representation $\mathbf{F} \in \mathbb{R}^{64}$ for the top-down orthographic projection of the floor.

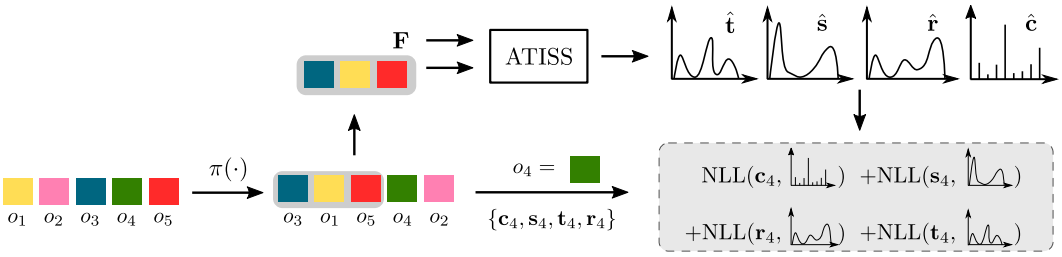


Figure 3: Training Overview: Given a scene with M objects (coloured squares), we first randomly permute them and then keep the first T objects (here $T = 3$). We task our network to predict the next object to be added in the scene given the subset of kept objects (highlighted with grey) and its floor layout feature $\mathbf{F}$. Our loss function is the negative log-likelihood (NLL) of the next object in the permuted sequence (green square).

Structure Encoder: The structure encoder h_θ maps the attributes of the j-th object into a per-object context embedding $\mathbf{C}_j$ as follows:

$$\begin{aligned} h_\theta : \mathbb{R}^C \times \mathbb{R}^3 \times \mathbb{R}^3 \times \mathbb{R}^1 &\rightarrow \mathbb{R}^{L_c} \times \mathbb{R}^{L_s} \times \mathbb{R}^{L_t} \times \mathbb{R}^{L_r} \\ (\mathbf{c}, \mathbf{s}, \mathbf{t}, \mathbf{r}) &\mapsto [\lambda(\mathbf{c}); \gamma(\mathbf{s}); \gamma(\mathbf{t}); \gamma(\mathbf{r})] \end{aligned} \tag{6}$$

where L_c, L_s, L_t, L_r are the output dimensionalities of the embeddings used to map the category, the size, the location and the orientation into a higher dimensional space respectively and $[\cdot\,;\cdot]$ denotes concatenation. For the object category $\mathbf{c}_j$ we use a learnable embedding $\lambda(\cdot)$, whereas for the size $\mathbf{s}_j$, the position $\mathbf{t}_j$ and the orientation $\mathbf{r}_j$, we use the positional encoding of [71] as follows

$$\gamma(p) = (\sin(2^0\pi p), \cos(2^0\pi p), \ldots, \sin(2^{L-1}\pi p), \cos(2^{L-1}\pi p)) \tag{7}$$

where p can be any of the size, position or orientation attributes and $\gamma(\cdot)$ is applied separately in each attribute's dimension. The set of per-object context vectors synthesizes the context embedding $\mathbf{C}$ that encapsulates information for the existing objects in the scene and is used to condition the next object to be generated. Before passing the output of (6) to the transformer encoder, we map each $\mathbf{C}_j$ to 64 dimensions using a linear projection.

Transformer Encoder: We follow [71, 12] and implement our encoder τ_θ as a multi-head attention transformer without any positional encoding. This allows us to learn a parametric function that computes features that are invariant to the order of $\mathbf{C}_j$ in $\mathbf{C}$. We use these features to predict the next object to be added in the scene, creating an autoregressive model. The input set of the transformer is $\mathbf{I} = \{\mathbf{F}\} \cup \{\mathbf{C}_j\}_{j=1}^{M} \cup \mathbf{q}$, with M the number of objects in the scene. $\mathbf{q} \in \mathbb{R}^{64}$ is a learnable object query vector that allows the transformer to predict output features $\hat{\mathbf{q}} \in \mathbb{R}^{64}$ used for generating the next object to be added in the scene. The use of a query token is akin to the use of a mask embedding in Masked Language Modelling [12] or the class embedding for the Vision Transformer [63].

Attribute Extractor: We autoregressively predict the attributes of the next object to be added in the scene using one MLP for each attribute. More formally, the attribute extractor is defined as follows:

$$c_\theta : \mathbb{R}^{64} \rightarrow \mathbb{R}^C \qquad \hat{\mathbf{q}} \mapsto \hat{\mathbf{c}} \tag{8}$$

$$t_\theta : \mathbb{R}^{64} \times \mathbb{R}^{L_c} \rightarrow \mathbb{R}^{3\times3\times K} \qquad (\hat{\mathbf{q}}, \lambda(\mathbf{c})) \mapsto \hat{\mathbf{t}} \tag{9}$$

$$r_\theta : \mathbb{R}^{64} \times \mathbb{R}^{L_c} \times \mathbb{R}^{L_t} \rightarrow \mathbb{R}^{1\times3\times K} \qquad (\hat{\mathbf{q}}, \lambda(\mathbf{c}), \gamma(\mathbf{t})) \mapsto \hat{\mathbf{r}} \tag{10}$$

$$s_\theta : \mathbb{R}^{64} \times \mathbb{R}^{L_c} \times \mathbb{R}^{L_t} \times \mathbb{R}^{L_r} \rightarrow \mathbb{R}^{3\times3\times K} \qquad (\hat{\mathbf{q}}, \lambda(\mathbf{c}), \gamma(\mathbf{t}), \gamma(\mathbf{r})) \mapsto \hat{\mathbf{s}} \tag{11}$$

where $\hat{\mathbf{c}}, \hat{\mathbf{s}}, \hat{\mathbf{t}}, \hat{\mathbf{r}}$ are the predicted attribute distributions and c_θ, t_θ, r_θ and s_θ are mappings between the latent space and the low-dimensional space of attributes. For the object category, c_θ predicts C class probabilities, whereas, t_θ, r_θ and s_θ predict the mean, variance and mixing coefficient for the K logistic distributions for each attribute. To predict the object properties in an autoregressive manner, we need to condition the prediction of a property on the previously predicted properties. Thus, instead of only passing $\hat{\mathbf{q}}$ to each MLP, we concatenate it with the previously predicted attributes, mapped in a higher-dimensional space using the embeddings $\lambda(\cdot)$ and $\gamma(\cdot)$ from (6).

3.3 Training and Inference

During training, we choose a scene from the dataset and apply a random permutation $\pi(\cdot)$ on its M objects. Then, we randomly select the first T objects to compute the context embedding $\mathbf{C}$.

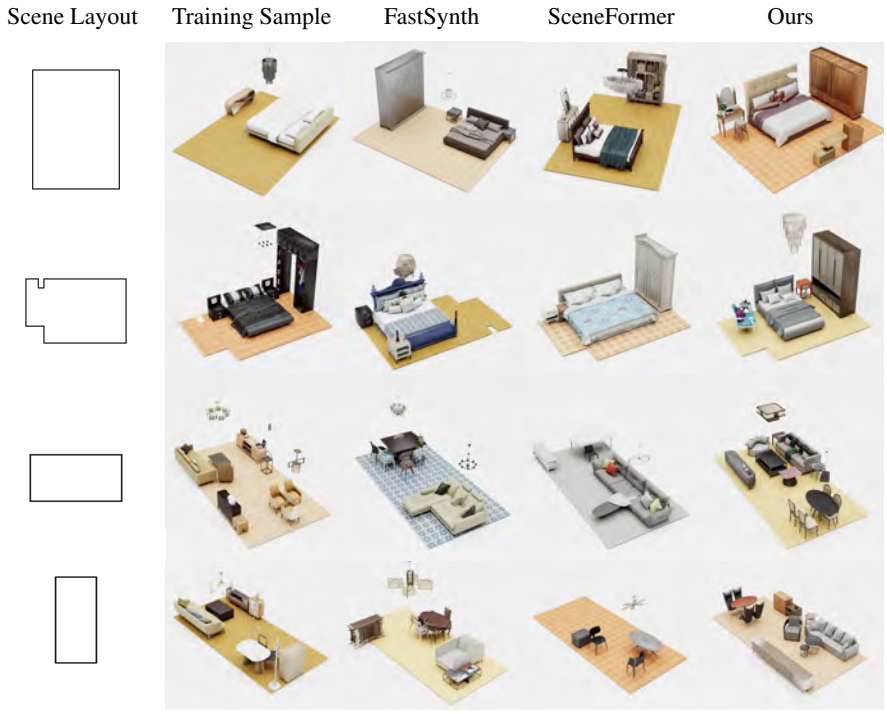


Figure 4: Qualitative Scene Synthesis Results. Synthesized scenes for three room types: bedrooms (1st+2nd row), living room (3rd row), dining room (4th row) using FastSynth, SceneFormer and our method. To showcase the generalization abilities of our model we also show the closest scene from the training set (2nd column).

Conditioned on $\mathbf{C}$ and $\mathbf{F}$, our network predicts the attribute distributions of the next object to be added in the scene and is trained to maximize the log-likelihood of the $T+1$ object from the permuted scene. A pictorial representation of the training process is provided in Fig. 3. To indicate the end of sequence, we augment the C object classes with an additional class, which we refer to as *end symbol*.

During inference, we start with an empty context embedding $\mathbf{C} = \emptyset$ and the floor representation $\mathbf{F}$ of the room to be populated and autoregressively sample attribute values from the predicted distributions of (8)-(11) for the next object. Once a new object is generated, it is appended to the context $\mathbf{C}$ to be used in the next step of the generation process until the *end symbol* is generated. A pictorial representation of the generation process can be found in Fig. 2. In order to transform the predicted labeled bounding boxes to 3D models we use object retrieval. In particular, we retrieve the closest object from the dataset in terms of the euclidean distance of the bounding box dimensions.

4 Experimental Evaluation

In this section, we provide an extensive evaluation of our method, comparing it to existing baselines. We further showcase several interactive use cases enabled by our method, not previously possible. Additional results as well as implementation details are provided in the supplementary.

Datasets: We train our model on the 3D-FRONT dataset [21] which contains a collection of $6,813$ houses with roughly $14,629$ rooms, populated with 3D furniture objects from the 3D-FUTURE dataset [22]. In our evaluation, we focus on four room types: (i) bedrooms, (ii) living rooms, (iii) dining rooms and (iv) libraries. After pre-processing to filter out uncommon object arrangements and rooms with unnatural sizes, we obtained 5996 bedrooms, 2962 living rooms, 2625 dining rooms and 622 libraries. We use 21 object categories for the bedrooms, 24 for the living and dining rooms and 25 for the libraries. The preprocessing steps are discussed in the supplementary.

Baselines: We compare our approach to FastSynth [61] and SceneFormer [74] using the authors' implementations. Note that both approaches were originally evaluated on the SUNCG dataset [65], which is now unavailable. Thus, we retrained both on 3D-FRONT. We also compare with a variant of our model that generates scenes as ordered sequences of objects (Ours+Order). To incorporate the order information to the input, we utilize a positional embedding [71] and a fixed ordering based on the object frequency as described in [74].

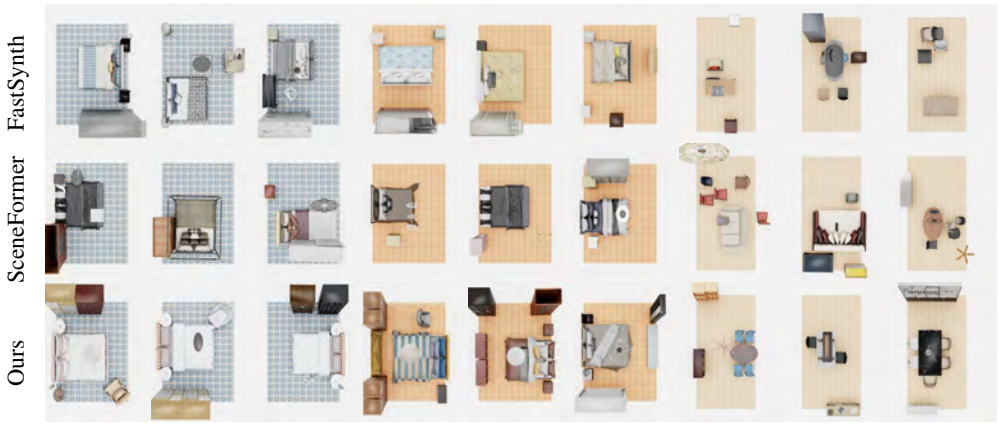


Figure 5: Scene Diversity. We show three generated scenes conditioned on three different floor plans for bedrooms and dining rooms. Every triplet of columns corresponds to a different floor plan.

	FID Score (↓)				Scene Classification Accuracy				Category KL Divergence (↓)			
	FastSynth	SceneFormer	Ours+Order	Ours	FastSynth	SceneFormer	Ours+Order	Ours	FastSynth	SceneFormer	Ours+Order	Ours
Bedrooms	40.89	43.17	38.67	**38.39**	0.883	0.945	0.760	**0.562**	0.0064	**0.0052**	0.0533	0.0085
Living	61.67	69.54	35.37	**33.14**	0.945	0.972	0.694	**0.516**	0.0176	0.0313	0.0372	**0.0034**
Dining	55.83	67.04	35.79	**29.23**	0.935	0.941	0.623	**0.477**	0.0518	0.0368	0.0278	**0.0061**
Library	37.72	55.34	35.60	**35.24**	0.815	0.880	0.572	**0.521**	0.0431	0.0232	0.0183	**0.0098**

Table 1: Quantitative Comparison. We report the FID score (↓) at 256^2 pixels, the KL divergence (↓) between the distribution of object categories of synthesized and real scenes and the real vs. synthetic classification accuracy for all methods. Classification accuracy closer to 0.5 is better.

Evaluation Metrics: To measure the realism of the generated scenes, we follow prior work [61] and report the KL divergence between the object category distributions of synthesized and real scenes from the test set and the classification accuracy of a classifier trained to discriminate real from synthetic scenes. We also report the FID [26] between top-down orthographic projections of synthesized and real scenes from the test set, which we compute using [51] on 256^2 images. We repeat the metric computation for FID and classification accuracy 10 times and report the average.

4.1 Scene Synthesis

We start by evaluating the performance of our model on generating plausible object configurations for various room types, conditioned on different floor plans. Fig. 4 provides a qualitative comparison of four scenes generated with our model and baselines. In some cases, both [61, 74] generate invalid room layouts with objects positioned outside room boundaries or overlapping. Instead, our model consistently synthesizes realistic object arrangements. We validate this quantitatively in Tab. 1, where we compare the generated scenes wrt. their similarity to the original data from 3D-FRONT. Synthesized scenes sampled from our model are almost indistinguishable from scenes from the test set, as indicated by the classification accuracy in Tab. 1, which is consistently around 50%. Our model also achieves lower FID scores for all room types and generates category distributions that are more faithful to the category distributions of the test set, expressed as lower KL divergence.

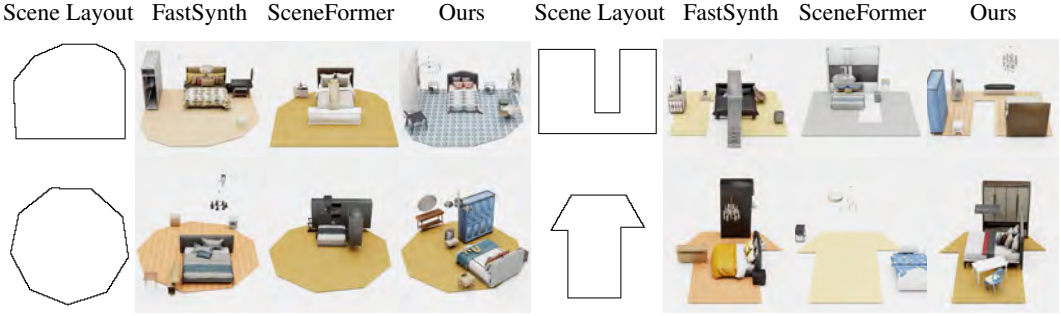


Figure 6: Generalization Beyond Training Data. We show four synthesized bedrooms conditioned on four room layouts that we manually designed.

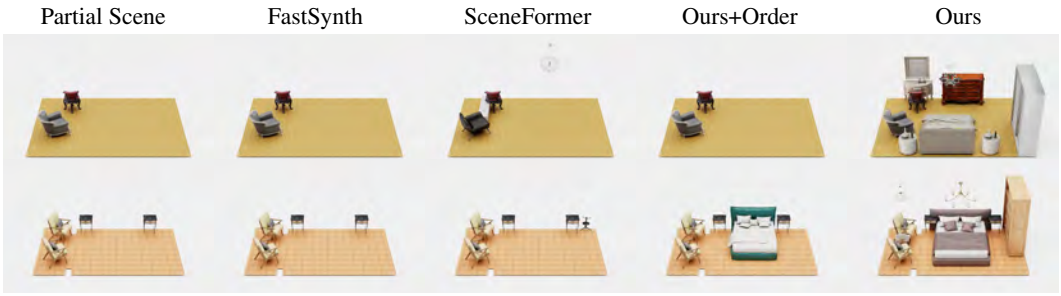


Figure 7: Scene Completion. Given a partial scene (left column), we visualize scene completions using our model and our baselines. Our model consistently generates plausible layouts.

	Bedroom	Living	Dining	Library
FastSynth [61]	13193.77	30578.54	26596.08	10813.87
SceneFormer [74]	849.37	731.84	901.17	369.74
Ours [61]	**102.38**	**201.59**	**201.84**	**88.24**

Table 2: Generation Time Comparison. We measure time (ms) to generate a scene, conditioned on a floor plan.

FastSynth [61]	SceneFormer [74]	Ours
38.180	129.298	36.053

Table 3: Network Parameters Comparison. We report the number of network parameters in millions.

To showcase that our model generates diverse object arrangements we visualize 3 generated scenes conditioned on the same floor plan for all methods (Fig. 5). We observe that our generated scenes are consistently valid and contain diverse object arrangements. In comparison [61, 74] struggle to generate plausible layouts particularly for the case of living rooms and dining rooms. We hypothesize that these rooms are more challenging than bedrooms, for the baselines, due to their significantly smaller volume of training data, and the large number of constituent objects per scene (20 on average, as opposed to 8). To investigate whether our model also generates plausible layouts conditioned on floor plans with uncommon shapes that are not in the training set, we manually design unconventional floor plans (Fig. 6) and generate bedroom layouts. While both [61, 74] fail to generate valid scenes, our model synthesizes diverse object layouts that are consistent with the floor plan. Finally, we compare the computational requirements of our architecture to [61, 74]. Our model is significantly faster (Tab. 2), while having fewer parameters (Tab. 3) than both [61, 74]. Note that [61] is orders of magnitude slower because it requires rendering every individual object added in the scene.

4.2 Applications

In this section, we present three applications that greatly benefit by our unordered set formulation and are crucial for creating an interactive scene synthesis tool.

Scene Completion: Starting from a partial scene, the task is to populate the empty space in a meaningful way. Since both [61, 74] are trained on sorted sequences of objects, they first generate frequent objects (e.g. beds, wardrobes) followed by less common objects. As a result, incomplete scenes that contain less common objects cannot be correctly populated. This is illustrated in Fig. 7, where [61, 74] either fail to add any objects in the scene or place furnitures in unatural positions, thus resulting in bedrooms without beds (see 1st row Fig. 7) and scenes with overlapping furniture (see 2nd row Fig. 7). In contrary, our model successfullly generates plausible completions with multiple objects such as lamps, wardrobes and dressing tables.

Figure 8: Failure Case Detection and Correction. We use a partial room with unnatural object arrangements. Our model identifies the problematic objects (first row, in green) and relocates them into meaningful positions.

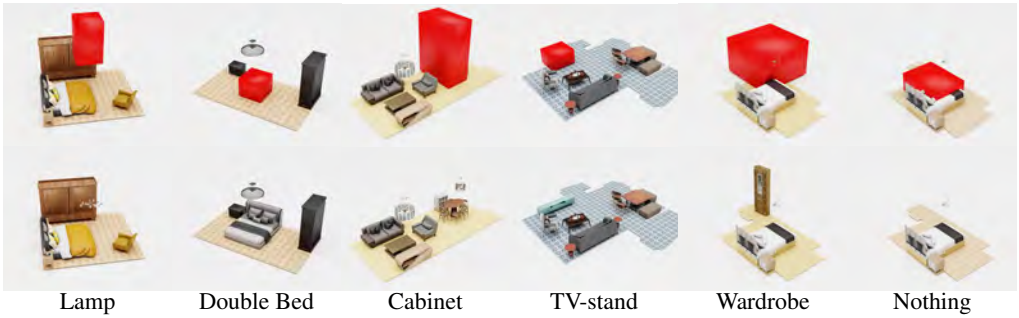


Figure 9: Object Suggestion. A user specifies a region of acceptable positions to place an object (marked as red boxes, 1st row) and our model suggests suitable objects (2nd row) to be placed in this location.

Failure Case Detection and Correction: We showcase that our model is able to identify and correct unnatural object arrangements. Given a scene, we compute the likelihood of each object, according to our model, conditioned on the other objects in the scene. We identify problematic objects as those with low likelihood and sample a new location from our generative model to rearrange it. We test our model in various scenarios such as overlapping objects, objects outside the room boundaries and objects in unnatural positions and show that it successfully identifies problematic objects (highlighted in green in Fig. 8) and rearranges them into a more plausible position. Note that this task cannot be performed by methods that consider ordering because they assign very low likelihood to common objects appearing after rare objects e.g. beds after cabinets.

Object Suggestion: We now test the ability of our model to provide object suggestions given a scene and user specified location constraints. To perform this task we sample objects from our generative model and accept the ones that fullfill the constraints provided by the user. Fig. 9 shows examples of location constraints (red box in top row) and the corresponding objects suggested (bottom row). Note that even when the user provided region is partially outside the room boundaries (4th, 5th column), suggested objects always reside in the room. Moreover, if the acceptable region overlaps with another object, our model suggests adding nothing (6th column). This task requires computing the likelihood of an object conditioned on an arbitrary scene, which [74, 61] cannot perform due to ordering.

4.3 Perceptual Study

We conducted two paired Amazon Mechanical Turk perceptual studies to evaluate the quality of our generated layouts against [61] and [74]. We sample 6 bedroom layouts for each method from the same 211 test set floor plans. Users saw 2 different rotating 3D scenes per method randomly selected from 6 pre-rendered layouts. Random layouts for each floor plan were assessed by 5 different workers to evaluate agreement and diversity across samples for a total of 1055 question sets per paired study. Generated scenes of [61] were judged to contain errors like interpenetrating furniture 41.4% of the time, nearly twice as frequently as our method, while [74] performs significantly worse (Tab. 4). Regarding realism, the scenes of [61] were more realistic than ours in only 26.9% of the cases. We conclude that our method outperforms the baselines in the key metric, generation of realistic indoor scenes, by a large margin. Additional details are provided in the supplementary.

Method	Condition	Mean Error Frequency ↓	More ↑ Realistic	Realism CI 99%
FastSynth [61]	vs. Ours	0.414	0.269	$[0.235, 0.306]$
SceneFormer [74]	vs. Ours	0.713	0.165	$[0.138, 0.196]$
Ours	vs. Both	**0.232**	**0.783**	$[0.759, 0.805]$

Table 4: Perceptual Study Results. Aggregated results for two A/B paired tests. Our method was judged more realistic with high confidence (binomial confidence interval with $\alpha = 0.01$ reported) and contained fewer errors.

5 Conclusion

We introduced ATISS, a novel autoregressive transformer architecture for synthesizing 3D rooms as unordered sets of objects. Our method generates realistic scenes that advance the state-of-the-art for scene synthesis. In addition, our novel formulation enables new interactive applications for semi-automated scene authoring, such as general scene completion, object suggestions, anomaly

detection and more. We believe that our model is an important step not only toward automating the generation of 3D environments, with impact on simulation and virtual testing, but also toward a new generation of tools for user-driven content generation. By accepting a wide range of user inputs, our model mitigates societal risks of task automation, and promises to usher in tools that enhance the workflow of skilled laborers, rather than replacing them. In future work, we plan to extend order invariance to object attributes to further expand interactive possibilities of this model, and to incorporate style information. As any machine learning model, our model can introduce learned biases for indoor scenes, and we plan to investigate learning from less structured and more widely available data sources to make this model applicable to a wider range of cultures and environments.

References

[1] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *Proc. of the International Conf. on Learning Representations (ICLR)*, 2019.

[2] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2020.

[4] Angel X. Chang, Manolis Savva, and Christopher D. Manning. Learning spatial knowledge for text to 3d scene generation. 2014.

[5] Siddhartha Chaudhuri, Evangelos Kalogerakis, Stephen Giguere, and Thomas A. Funkhouser. Attribit: content creation with semantic attributes. In *The 26th Annual ACM Symposium on User Interface Software and Technology, UIST'13, St. Andrews, United Kingdom, October 8-11, 2013*, 2013.

[6] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *Proc. of the International Conf. on Machine learning (ICML)*, 2020.

[7] Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers. *arXiv.org*, 2019.

[8] Yunjey Choi, Min-Je Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[9] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[10] Jia Deng, Wei Dong, Richard Socher, Li jia Li, Kai Li, and Li Fei-fei. Imagenet: A large-scale hierarchical image database. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[11] Jeevan Devaranjan, Amlan Kar, and Sanja Fidler. Meta-sim2: Unsupervised learning of scene structure for synthetic data generation. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2020.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019.

[13] Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. Jukebox: A generative model for music. *arXiv.org*, 2020.

[14] Xinhan Di, Pengqian Yu, Hong Zhu, Lei Cai, Qiuyan Sheng, and Changyu Sun. Structural plan of indoor scenes with personalized preferences. *arXiv.org*, 2020.

[15] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real NVP. In *Proc. of the International Conf. on Learning Representations (ICLR)*, 2017.

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proc. of the International Conf. on Learning Representations (ICLR)*, 2021.

[17] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[18] Matthew Fisher, Daniel Ritchie, Manolis Savva, Thomas A. Funkhouser, and Pat Hanrahan. Example-based synthesis of 3d object arrangements. 2012.

[19] Matthew Fisher, Manolis Savva, and Pat Hanrahan. Characterizing structural relationships in scenes using graph kernels. *ACM Trans. on Graphics*, 2011.

[20] Matthew Fisher, Manolis Savva, Yangyan Li, Pat Hanrahan, and Matthias Nießner. Activity-centric scene synthesis for functional 3d scene modeling. *ACM Trans. on Graphics*, 2015.

[21] Huan Fu, Bowen Cai, Lin Gao, Lingxiao Zhang, Cao Li, Zengqi Xun, Chengyue Sun, Yiyun Fei, Yu Zheng, Ying Li, Yi Liu, Peng Liu, Lin Ma, Le Weng, Xiaohang Hu, Xin Ma, Qian Qian, Rongfei Jia, Binqiang Zhao, and Hao Zhang. 3d-front: 3d furnished rooms with layouts and semantics. *arXiv.org*, abs/2011.09127, 2020.

[22] Huan Fu, Rongfei Jia, Lin Gao, Mingming Gong, Binqiang Zhao, Steve J. Maybank, and Dacheng Tao. 3d-future: 3d furniture shape with texture. *arXiv.org*, abs/2009.09633, 2020.

[23] 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 Trans. on Graphics*, 2017.

[24] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[26] 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. In *Advances in Neural Information Processing Systems (NIPS)*, 2017.

[27] Akash Abdu Jyothi, Thibaut Durand, Jiawei He, Leonid Sigal, and Greg Mori. Layoutvae: Stochastic scene layout generation from a label set. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2019.

[28] Amlan Kar, Aayush Prakash, Ming-Yu Liu, Eric Cameracci, Justin Yuan, Matt Rusiniak, David Acuna, Antonio Torralba, and Sanja Fidler. Meta-sim: Learning to generate synthetic datasets. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2019.

[29] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[30] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. 2020.

[31] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *Proc. of the International Conf. on Machine learning (ICML)*, 2020.

[32] Mohammad Keshavarzi, Aakash Parikh, Xiyu Zhai, Melody Mao, Luisa Caldas, and Allen Y. Yang. Scenegen: Generative contextual scene augmentation using scene graph priors. *arXiv.org*, 2020.

[33] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proc. of the International Conf. on Learning Representations (ICLR)*, 2015.

[34] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *Proc. of the International Conf. on Learning Representations (ICLR)*, 2014.

[35] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems (NIPS)*, 2018.

[36] Adam R. Kosiorek, Hyunjik Kim, and Danilo J. Rezende. Conditional set generation with transformers. *arXiv.org*, 2020.

[37] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2012.

[38] Matt J Kusner, Brooks Paige, and José Miguel Hernández-Lobato. Grammar variational autoencoder. In *Proc. of the International Conf. on Machine learning (ICML)*, pages 1945–1954. PMLR, 2017.

[39] Juho Lee, Yoonho Lee, Jungtaek Kim, Adam R. Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *Proc. of the International Conf. on Machine learning (ICML)*, 2019.

[40] Lingyun Luke Li, Bin Yang, Ming Liang, Wenyuan Zeng, Mengye Ren, Sean Segal, and Raquel Urtasun. End-to-end contextual perception and prediction with interaction transformer. In *Proc. IEEE International Conf. on Intelligent Robots and Systems (IROS)*, pages 5784–5791, 2020.

[41] Manyi Li, Akshay Gadi Patil, Kai Xu, Siddhartha Chaudhuri, Owais Khan, Ariel Shamir, Changhe Tu, Baoquan Chen, Daniel Cohen-Or, and Hao (Richard) Zhang. GRAINS: generative recursive autoencoders for indoor scenes. *ACM Trans. on Graphics*, 2019.

[42] Andrew Luo, Zhoutong Zhang, Jiajun Wu, and Joshua B. Tenenbaum. End-to-end optimization of scene layout. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[43] Rui Ma, Akshay Gadi Patil, Matthew Fisher, Manyi Li, Sören Pirk, Binh-Son Hua, Sai-Kit Yeung, Xin Tong, Leonidas J. Guibas, and Hao Zhang. Language-driven synthesis of 3d scenes from scene databases. *ACM Trans. on Graphics*, 2018.

[44] Paul Merrell, Eric Schkufza, Zeyang Li, Maneesh Agrawala, and Vladlen Koltun. Interactive furniture layout using interior design guidelines. *ACM Trans. on Graphics*, 2011.

[45] Kaichun Mo, Paul Guerrero, Li Yi, Hao Su, Peter Wonka, Niloy Mitra, and Leonidas Guibas. Structurenet: Hierarchical graph networks for 3d shape generation. In *ACM Trans. on Graphics*, 2019.

[46] Kaichun Mo, Leonidas J. Guibas, Mustafa Mukadam, Abhinav Gupta, and Shubham Tulsiani. Where2act: From pixels to actions for articulated 3d objects. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2021.

[47] Pascal Müller, Peter Wonka, Simon Haegler, Andreas Ulmer, and Luc Van Gool. Procedural modeling of buildings. *ACM Trans. on Graphics*, 2006.

[48] Charlie Nash, Yaroslav Ganin, S. M. Ali Eslami, and Peter W. Battaglia. Polygen: An autoregressive generative model of 3d meshes. In *Proc. of the International Conf. on Machine learning (ICML)*, 2020.

[49] Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. Scaling neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, 2018.

[50] Yoav I. H. Parish and Pascal Müller. Procedural modeling of cities. In *ACM Trans. on Graphics*, 2001.

[51] Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On buggy resizing libraries and surprising subtleties in FID calculation. *arXiv.org*, 2021.

[52] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In *Proc. of the International Conf. on Machine learning (ICML)*, 2018.

[53] Despoina Paschalidou, Angelos Katharopoulos, Andreas Geiger, and Sanja Fidler. Neural parts: Learning expressive 3d shape abstractions with invertible neural networks. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[54] Despoina Paschalidou, Ali Osman Ulusoy, and Andreas Geiger. Superquadrics revisited: Learning 3d shape parsing beyond cuboids. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[55] Despoina Paschalidou, Luc van Gool, and Andreas Geiger. Learning unsupervised hierarchical part decomposition of 3d objects from a single rgb image. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[56] Adam Paszke, Abhishek Chaurasia, Sangpil Kim, and Eugenio Culurciello. Enet: A deep neural network architecture for real-time semantic segmentation. *arXiv.org*, 1606.02147, 2016.

[57] Aayush Prakash, Shaad Boochoon, Mark Brophy, David Acuna, Eric Cameracci, Gavriel State, Omer Shapira, and Stan Birchfield. Structured domain randomization: Bridging the reality gap by context-aware synthetic data. In *Proc. IEEE International Conf. on Robotics and Automation (ICRA)*, 2019.

[58] Pulak Purkait, Christopher Zach, and Ian Reid. SG-VAE: scene grammar variational autoencoder to generate new indoor scenes. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2020.

[59] Siyuan Qi, Yixin Zhu, Siyuan Huang, Chenfanfu Jiang, and Song-Chun Zhu. Human-centric indoor scene synthesis using stochastic grammar. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[60] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *arXiv.org*, 2019.

[61] Daniel Ritchie, Kai Wang, and Yu-An Lin. Fast and flexible indoor scene synthesis via deep convolutional generative models. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[62] Tim Salimans, Andrej Karpathy, Xi Chen, and Diederik P. Kingma. Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications. In *Proc. of the International Conf. on Learning Representations (ICLR)*, 2017.

[63] Gilad Sharir, Asaf Noy, and Lihi Zelnik-Manor. An image is worth 16x16 words, what is a video worth? In *Proc. of the International Conf. on Learning Representations (ICLR)*, 2021.

[64] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, 2018.

[65] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[66] Jerry O. Talton, Yu Lou, Steve Lesser, Jared Duke, Radomír Mech, and Vladlen Koltun. Metropolis procedural modeling. *ACM Trans. on Graphics*, 2011.

[67] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. *arXiv.org*, 2020.

[68] Shubham Tulsiani and Abhinav Gupta. Pixeltransformer: Sample conditioned signal generation. In *Proc. of the International Conf. on Machine learning (ICML)*, 2021.

[69] Shubham Tulsiani, Hao Su, Leonidas J. Guibas, Alexei A. Efros, and Jitendra Malik. Learning shape abstractions by assembling volumetric primitives. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[70] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W. Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. In *The 9th ISCA Speech Synthesis Workshop, Sunnyvale, CA, USA, 13-15 September 2016*, 2016.

[71] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NIPS)*, pages 5998–6008, 2017.

[72] 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 Trans. on Graphics*, 2019.

[73] Kai Wang, Manolis Savva, Angel X. Chang, and Daniel Ritchie. Deep convolutional priors for indoor scene synthesis. *ACM Trans. on Graphics*, 37(4):70:1–70:14, 2018.

[74] Xinpeng Wang, Chandan Yeshwanth, and Matthias Nießner. Sceneformer: Indoor scene generation with transformers. *arXiv.org*, 2020.

[75] Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. Cross-modal self-attention network for referring image segmentation. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[76] 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 Trans. on Graphics*, 2011.

[77] Lap-Fai Yu, Sai Kit Yeung, and Demetri Terzopoulos. The clutterpalette: An interactive tool for detailing indoor scenes. *IEEE Trans. Vis. Comput. Graph.*, 2016.

[78] Song-Hai Zhang, Shaokui Zhang, Wei-Yu Xie, Cheng-Yang Luo, and Hong-Bo Fu. Fast 3d indoor scene synthesis with discrete and exact layout pattern extraction. *arXiv.org*, 2020.

[79] 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 Trans. on Graphics*, 2020.

[80] Yang Zhou, Zachary While, and Evangelos Kalogerakis. Scenegraphnet: Neural message passing for 3d indoor scene augmentation. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2019.

[81] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: deformable transformers for end-to-end object detection. *Proc. of the International Conf. on Learning Representations (ICLR)*, 2021.

Supplementary Material for ATISS: Autoregressive Transformers for Indoor Scene Synthesis

Abstract

In this **supplementary document**, we provide a detailed overview of our network architecture and the training procedure. Subsequently, we describe the preprocessing steps that we followed to filter out problematic rooms from the 3D-FRONT dataset [21]. Next, we provide ablations on how different components of our system impact the performance of our model on the scene synthesis task and we compare ATISS with various transformer models that consider ordering. Finally, we provide additional qualitative and quantitative results as well as additional details for our perceptual study presented in Sec 4.3 in our main submission.

A Implementation Details

In this section, we provide a detailed description of our network architecture. We then describe our training protocol and provide details on the metrics computation during training and testing. Finally, we also provide additional details regarding our baselines.

A.1 Network Architecture

Here we describe the architecture of each individual component of our model (from Fig. 2 in the main submission). Our architecture comprises four components: (i) the *layout encoder* that maps the room shape to a global feature representation $\mathbf{F}$, (ii) the *structure encoder* that maps the M objects in a scene into per-object context embeddings $\mathbf{C} = \{\mathbf{C}_j\}_{j=1}^{M}$, (iii) the *transformer encoder* that takes $\mathbf{F}$, $\mathbf{C}$ and a query embedding $\mathbf{q}$ and predicts the features $\hat{\mathbf{q}}$ for the next object to be generated and (iv) the *attribute extractor* that autoregressively predicts the attributes of the next object.

Layout Encoder: The first part of our architecture is the *layout encoder* that is used to map the room's floor into a global feature representation $\mathbf{F}$. We follow [73] and we model the floor plan with its top-down orthographic projection. This projection maps the floor plan into an image, where pixel values of 1 indicate regions inside the room and pixel values of 0 otherwise. The layout encoder is implemented with a ResNet-18 architecture [25] that is not pre-trained on ImageNet [10]. We empirically observed that using a pre-trained ResNet resulted in worse performance. From the original architecture, we remove the final fully connected layer and replace it with a linear projection to 64 dimensions, after average pooling.

Structure Encoder: The structure encoder maps the attributes of each object into a per-object context embedding $\mathbf{C}_j$. For the object category $\mathbf{c}_j$, we use a learnable embedding, which is simply a matrix of size $C \times 64$, that stores a per-object category vector, for all C object categories in the dataset. For the size $\mathbf{s}_j$, the position $\mathbf{t}_j$ and the orientation $\mathbf{r}_j$, we use the positional encoding of [71] as follows

$$\gamma(p) = (\sin(2^0\pi p), \cos(2^0\pi p), \ldots, \sin(2^{L-1}\pi p), \cos(2^{L-1}\pi p)) \tag{12}$$

where p can be any of the size, position or orientation attributes and $\gamma(\cdot)$ is applied separately in each attribute's dimension. In our experiments, L is set to 32. The output of each embedding layer, used to map the category, size, location and orientation in a higher dimensional space, are concatenated into an 512-dimensional feature vector, which is then mapped to the per-object context embedding. A pictorial representation of the structure encoder is provided in Fig. 10.

Transformer Encoder: We follow [71, 12] and implement our transformer encoder as a multi-head attention transformer without any positional encoding. Our transformer consists of 4 layers

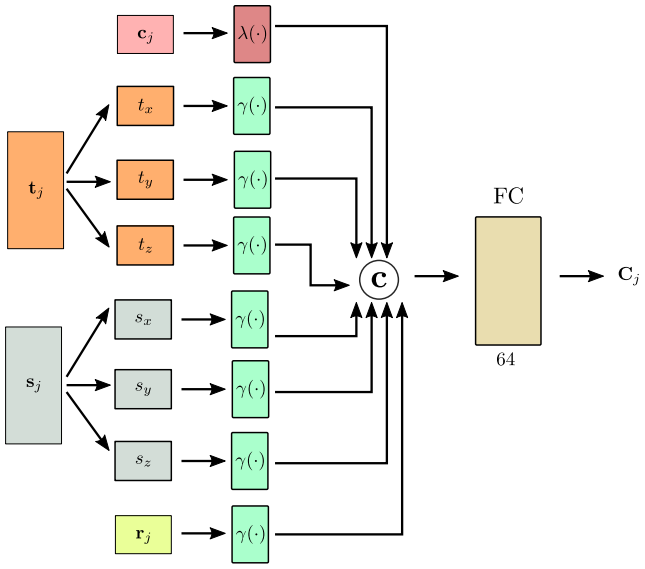
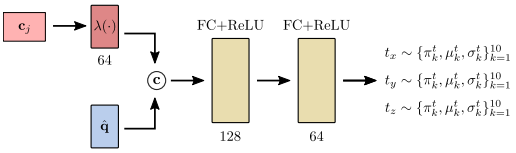
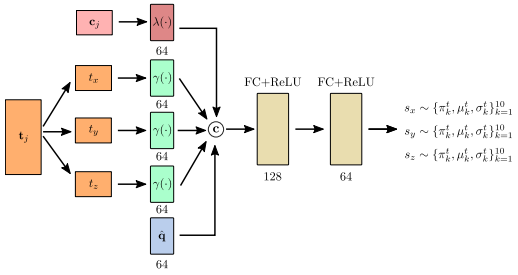


Figure 10: Structure Encoder. The structure encoder predicts the per-object context embeddings $\mathbf{C}_j$ conditioned on the object attributes. For the object category $\mathbf{c}_j$, we use a learnable embedding $\lambda(\cdot)$, whereas for the location $\mathbf{t}_j$, the size $\mathbf{s}_j$ and orientation $\mathbf{r}_j$ we employ the positional encoding from (12). Note that the positional encoding $\gamma(\cdot)$ is applied separately in each dimension of $\mathbf{t}_j$ and $\mathbf{s}_j$.

with 8 attention heads. The queries, keys and values have 64 dimensions and the intermediate representations for the MLPs have 1024 dimensions. To implement the transformer architecture we use the transformer library provided by Katharopoulos et al. [31][1]. The input set of the transformer is $\mathbf{I} = \{\mathbf{F}\} \cup \{\mathbf{C}_j\}_{j=1}^{M} \cup \mathbf{q}$, where M denotes the number of objects in the scene and $\mathbf{q} \in \mathbb{R}^{64}$ is a learnable object query vector that allows the transformer to predict output features $\hat{\mathbf{q}} \in \mathbb{R}^{64}$ used for generating the next object to be added in the scene.



(a) $t_\theta(\cdot)$ predicts the parameters of the mixture of logistics distribution for the location $\mathbf{t}$.



(b) $s_\theta(\cdot)$ predicts the parameters of the mixture of logistics distribution for the size $\mathbf{s}$.

Figure 11: Attribute Extractor. The attribute extractor consists of four MLPs that autoregressively predict the object attributes. Here we visualize the MLP $t_\theta(\cdot)$ for the location attribute (left side) and the MLP $s_\theta(\cdot)$ for the size attribute (right side).

Attribute Extractor: The attribute extractor autoregressively predicts the attributes of the next object to be added in the scene. The MLP for the object category is a linear layer with 64 hidden dimensions that predicts C class probabilities per object. The MLPs for the location, orientation and size predict the mean, variance and mixing coefficient for the K logistic distributions for each attribute. In our experiments we set $K = 10$. The size, location and orientation attributes are predicted using a 2-layer MLP with RELU non-linearities with hidden size 128 and output size 64. A pictorial representation for the MLPs $t_\theta(\cdot)$ and $\sigma_\theta(\cdot)$ used to predict the parameters of the mixture of logistics distribution for the location and the size is provided in Fig. 11. Note that r_θ is defined in a similar manner.

[1] https://github.com/idiap/fast-transformers

A.2 Object Retrieval

During inference, we select 3D models from the 3D-FUTURE dataset [22] to be placed in the scene based on the predicted category, location, orientation and size. In particular, we perform nearest neighbor search through the 3D-FUTURE dataset[22] to find the closest model in terms of object dimensions. While prior work [61, 74] explored more complex object retrieval schemes based on object dimensions and object cooccurrences (i.e. favor 3D model of objects that frequently co-occur in the dataset), we note that our simple object retrieval strategy consistently resulted in visually plausible rooms. We leave more advanced object retrieval schemes for future research.

A.3 Training Protocol

In all our experiments, we use the Adam optimizer [33] with learning rate $\eta = 10^{-4}$ and no weight decay. For the other hyperparameters of Adam we use the PyTorch defaults: $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-8}$. We train all models with a batch size of 128 for 100k iterations. During training, we perform rotation augmentation with random rotations between $[0, 360]$ degrees. To determine when to stop training, we follow common practice and evaluate the validation metric every 1000 iterations and use the model that performed best as our final model.

A.4 Metrics Computation

As mentioned in our main submission, we evaluate our model and our baselines using the KL divergence between the object category distributions of synthesized and real scenes and the classification accuracy of a classifier trained to discriminate real from synthetic scenes as well as the FID [26] score between 256^2 top-down orthographic projections of synthesized and real scenes using the code provided by Parmar et al. [51][2]. For the metrics computation, we generate the same amount of scenes as in the test set and we compute each metric using real scenes from the test set. In particular, for the KL divergence, we measure the frequency of object category occurrences in the generated scenes and compare it with the frequency of object occurrences in real scenes. Regarding the scene classification accuracy, we train a classifier to distinguish real from generated scenes. Our classifier is an Alexnet [37] pre-trained on ImageNet, that takes as input a 256^2 top-down image-based representation of a room and predicts whether this scene is real or synthetic. Both for the FID and the classification accuracy, we repeat the metric computation 10 times and report the average.

A.5 Baselines

In this section, we provide additional details regarding our baselines. We compare our model with FastSynth [61] and SceneFormer [74]. Both methods were originally evaluated on the SUNCG dataset [65], which is currently unavailable, thus we retrained both on 3D-FRONT using the augmentation techniques described in the original papers. To ensure fair comparison, we use the same object retrieval for all methods and no rule-based post-processing on the generated layouts.

FastSynth: In FastSynth [61], the authors employ a series of image-based CNNs to sequentially predict the attributes of the next object to be added in the scene. In addition to 2D labeled bounding boxes they have auxiliary supervision in the form of object segmentation masks, depth maps, wall masks etc. For more details, we refer the reader to [73]. During training, they assume that there exists an ordering of objects in each scene, based on the average size of each category multiplied by its frequency of occurrences in the dataset. Each CNN module is trained separately and the object properties are predicted in an autoregressive manner: object category first, followed by location, orientation and size. We train [61][3] using the provided PyTorch [56] implementation with the default parameters until convergence.

SceneFormer: In SceneFormer [74], the authors utilize a series of transformers to autoregressively add objects in the scene, similar to [61]. In particular, they train a separate transformer for each attribute and they predict the object properties in an autoregressive manner: object category first, followed by orientation, location and size. Similar to [61], they also treat scenes as ordered sequences

[2] https://github.com/GaParmar/clean-fid
[3] https://github.com/brownvc/fast-synth

of objects ordered by the frequency of their categories. We train [74][4] using the provided PyTorch [56] implementation with the default parameters until convergence.

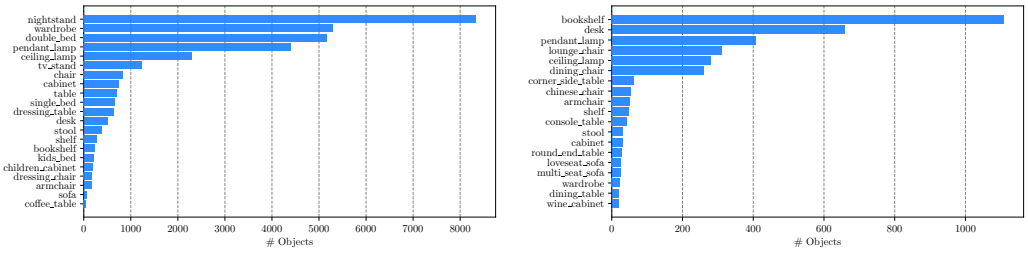
B 3D-FRONT Dataset Filtering

We evaluate our model on the 3D-FRONT dataset [21], which is one of the few available datasets that contain indoor environments. 3D-FRONT contains a collection of 6813 houses with roughly 14629 designed rooms, populated with 3D furniture objects from the 3D-FUTURE dataset [22]. In our experiments, we focused on four room types: (i) bedrooms, (ii) living rooms, (iii) dining rooms and (iv) libraries. Unfortunately, 3D-FRONT contains multiple problematic rooms with unnatural sizes, misclassified objects as well as objects in unnatural positions e.g. outside the room boundaries, lamps on the floor, overlapping objects etc. Therefore, in order to be able to use it, we had to perform thorough filtering to remove problematic scenes. In this section, we present in detail the pre-processing steps for each room type. We plan to release the names/ids of the filtered rooms, when the paper is published.

The 3D-FRONT dataset provides scenes for the following room types: *bedroom*, *diningroom*, *elderly-room*, *kidsroom*, *library*, *livingdiningroom*, *livingroom*, *masterbedroom*, *nannyroom*, *secondbedroom* that contain 2287, 3233, 233, 951, 967, 2672, 1095, 3313, 16 and 2534 rooms respectively. Since some room types have very few rooms we do not consider them in our evaluation.

Bedroom: To create training and test data for bedroom scenes, we consider rooms of type *bedroom, secondbedroom* and *masterbedroom*, which amounts to 8134 rooms in total. We start by removing rooms of unnatural sizes, namely rooms that are larger than 6m $\times$ 6m in floor size and taller than 4m. Next, we remove infrequent objects that appear in less than 15 rooms, such as chaise lounge sofa, l-shaped sofa, barstool, wine cabinet etc. Subsequently, we filter out rooms that contain fewer than 3 and more than 13 objects, since they amount to a small portion of the dataset. Since the original dataset contained various rooms with problematic object arrangements such overlapping objects, we also remove rooms that have objects that are overlapping as well as misclassified objects e.g. beds being classified as wardrobes. This results in 5996 bedrooms with 21 object categories in total. Fig. 12a illustrates the number of appearances of each object category in the 5996 bedroom scenes and we remark that the most common category is the nightstand with 8337 occurrences and the least common is the coffee table with 45.

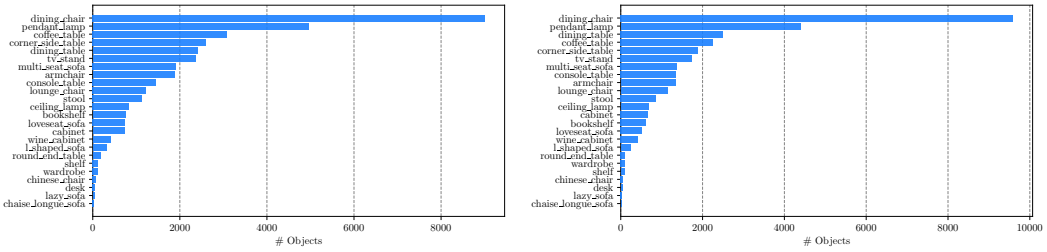
Library: We consider rooms of type *library* that amounts to 967 scenes in total. For the case of libraries, we start by filtering out rooms with unnatural sizes that are larger than 6m $\times$ 6m in floor size and taller than 4m. Again we remove rooms that contain overlapping objects, objects positioned outside the room boundaries as well as rooms with unnatural layouts e.g. single chair positioned in the center of the room. We also filter out rooms that contain less than 3 objects and more than 12 objects since they appear less frequently. Our pre-processing resulted in 622 rooms with 19 object categories in total. Fig. 12b shows the number of appearances of each object category in the 622 libraries. The most common category is the bookshelf with 1109 occurrences and the least common is the wine cabinet with 19.



(a) Bedrooms **(b) Libraries**

Figure 12: Number of object occurrences in Bedrooms and Libraries.

[4] https://github.com/cy94/sceneformer



(a) Living Rooms **(b) Dining Rooms**

Figure 13: Number of object occurrences in Living Rooms and Dining Rooms.

Living Room: For the living rooms, we consider rooms of type *livingroom* and *livingdiningroom*, which amounts to 3767 rooms. We follow a similar process as before and we start by filtering out rooms with unnatural sizes. In particular, we discard rooms that are larger than 12m $\times$ 12m in floor size and taller than 4m. We also remove uncommon objects that appear in less than 15 rooms such as bed and bed frame. Next, we filter out rooms that contain less than 3 objects and more than 13 objects, since they are significantly less frequent. For the case of living rooms, we observed that some of the original scenes contained multiple lamps without having any other furniture. Since this is unnatural, we also removed these scenes together with some rooms that had either overlapping objects or objects positioned outside the room boundaries. Finally, we also remove any scenes that contain any kind of bed e.g. double bed, single bed, kid bed etc. After our pre-processing, we ended up with 2962 living rooms with 24 object categories in total. Fig. 13a visualizes the number of occurrences of each object category in the living rooms. We observe that the most common category is the dining chair with 9009 occurrences and the least common is the chaise lounge sofa with 30.

Dining Room: For the dining rooms, we consider rooms of type *diningroom* and *livingdiningroom*, since the *diningroom* scenes amount to only 233 scenes. This results in 3233 rooms in total. For the dining rooms, we follow the same filtering process as for the living rooms and we keep 2625 rooms with 24 objects in total. Fig. 13b shows the number of occurrences of each object category in the dining rooms. The most common category is the dining chair with 9589 occurrences and the least common is the chaise lounge sofa with 19.

To generate the train, test and validation splits, we split the preprocessed rooms such that 70% is used for training, 20% for testing and 10% for validation. Note that the 3D-FRONT dataset comprises multiple houses that may contain the same room, e.g the exact same object arrangement might appear in multiple houses. Thus splitting train and test scenes solely based on whether they belong to different houses could result in the same room appearing both in train and test scenes. Therefore, instea of randomly selecting rooms from houses but we select from the set of rooms with distinct object arrangements.

C Ablation Study

In this section, we investigate how various components of our model affect its performance on the scene synthesis task. In Sec. C.1, we investigate the impact of the number of logistic distributions in the performance of our model. Next, in Sec. C.2, we examine the impact of the architecture of the layout encoder. In Sec. C.3, we compare ATISS with two variants of our model that consider ordered sets of objects. Unless stated otherwise, all ablations are conducted on the bedroom scenes of the 3D-FRONT [21] dataset.

C.1 Mixture of Logistic distributions

We represent objects in a scene as labeled 3D bounding boxes and model them with four random variables that describe their category, size, orientation and location, $o_j = \{\mathbf{c}_j, \mathbf{s}_j, \mathbf{t}_j, \mathbf{r}_j\}$. The category $\mathbf{c}_j$ is modeled using a categorical variable over the total number of object categories C in the dataset. For the size $\mathbf{s}_j \in \mathbb{R}^3$, the location $\mathbf{t}_j \in \mathbb{R}^3$ and the orientation $\mathbf{r}_j \in \mathbb{R}^1$, we follow [62, 70]

and model them with a mixture of logistic distributions

$$\mathbf{s}_j \sim \sum_{k=1}^{K} \pi_k^s \text{logistic}(\mu_k^s, \sigma_k^s) \quad \mathbf{t}_j \sim \sum_{k=1}^{K} \pi_k^t \text{logistic}(\mu_k^t, \sigma_k^t) \quad \mathbf{r}_j \sim \sum_{k=1}^{K} \pi_k^r \text{logistic}(\mu_k^r, \sigma_k^r) \quad (13)$$

where π_k^s, μ_k^s and σ_k^s denote the weight, mean and variance of the k-th logistic distribution used for modeling the size. Similarly, π_k^t, μ_k^t and σ_k^t and π_k^r, μ_k^r ans σ_k^r refer to the weight, mean and variance of the k-th logistic of the location and orientation, respectively.

In this experiment, we test our model with different numbers for logistic distributions for modelling the object attributes. Results are summarized in Tab. 5.

	FID (↓)	Classification Accuracy (↓)	Category Distribution (↓)
$K = 1$	41.71 ± 0.4008	0.7826 ± 0.0080	0.0491
$K = 5$	40.41 ± 0.2491	0.5667 ± 0.0405	0.0105
$K = 10$	**38.39** ± 0.3392	**0.5620** ± 0.0228	0.0085
$K = 15$	40.41 ± 0.4504	0.5980 ± 0.0074	0.0095
$K = 20$	40.39 ± 0.3964	0.6680 ± 0.0035	**0.0076**

Table 5: Ablation Study on the Number of Logistic Distributions. This table shows a quantitative comparison of our approach with different numbers of K logistic distributions for modelling the size, the location and the orientation of each object.

As it is expected, using a single logistic distribution (first row in Tab. 5) results in worse performance, since it does not have enough representation capacity for modelling the object attributes. We also note that increasing the number of logistic distributions beyond 10 hurts performance wrt. FID and classification accuracy. We hypothesize that this is due to overfitting. In our experiments we set $K = 10$.

C.2 Layout Encoder

We further examine the impact of the layout encoder on the performance of our model. To this end, we replace the ResNet-18 architecture [25], with an AlexNet [37]. From the original architecture, we remove the final classifier layers and keep only the feature vector of length 9216 after max pooling. We project this feature vector to 64 dimensions with a linear projection layer. Similar to our vanilla model, we do not use an AlexNet pre-trained on ImageNet because we empirically observed that it resulted in worse performance.

	FID (↓)	Classification Accuracy (↓)	Category Distribution (↓)
AlexNet	40.40 ± 0.2637	0.6083 ± 0.0034	**0.0064**
ResNet-18	**38.39** ± 0.3392	**0.5620** ± 0.0228	0.0085

Table 6: Ablation Study on the Layout Encoder Architecture. This table shows a quantitative comparison of ATISS with two different layout encoders.

Tab. 6 compares the two variants of our model wrt. to the FID score, the classification accuracy and the KL-divergence. We remark that our method is not particularly sensitive to the choice of the layout encoder. However, using an AlexNet results in slightly worse performance, hence we utilize a ResNet-18 in all our experiments.

C.3 Transformers with Ordering

In this section, we analyse the benefits of synthesizing rooms as unordered sets of objects in contrast to ordered sequences. To this end, we train two variants of our model that utilize a positional embedding [71] to incorporate order information to the input. The first variant is trained with random permutations of the input (Ours+Perm+Order), similar to our model, whereas the second with a fixed ordering based on the object frequency (Ours+Order) as described in [61, 74]. We compare these variants to our model on the scene synthesis task and observe that the variant with the fixed ordering (second row Tab. 7) performs significantly worse as the classifier can identify synthesized scenes with 76% accuracy. Moreover, we remark that besides enabling all the applications presented in our main submission, training with random permutations also improves the quality of the synthesized

	FID (↓)	Classification Accuracy (↓)	Category Distribution (↓)
Ours+Perm+Order	40.18 ± 0.2831	0.6019 ± 0.0060	0.0089
Ours+Order	38.67 ± 0.5552	0.7603 ± 0.0010	0.0533
Ours	**38.39** ± 0.3392	**0.5620** ± 0.0228	0.0085

Table 7: Ablation Study on Ordering. This table shows a quantitative comparison of our approach wrt. two variants of our model that represent rooms as ordered sequence of objects.

Figure 14: Failure Case Detection and Correction. Starting from a room with an unnatural object arrangement, our model identifies the problematic objects (first row and third row, in green) and relocates them into meaningful positions (second and fourth row).

scenes (first row Tab. 7). However, our model that is permutation invariant, namely the prediction is the same regardless of the order of the partial scene, performs even better (third row Tab. 7). We conjecture that the invariance of our model will be more even more crucial for training with either larger datasets or larger scenes i.e. scenes with more objects, because observing a single order allows reasoning about all permutations of the partial scene.

D Applications

In this section, we provide additional qualitative results for various interactive applications that benefit greatly by our unordered set formulation.

D.1 Failure Case Detection And Correction

In this experiment, we investigate whether our model is able to identify unnatural furniture layouts and reposition the problematic objects such that they preserve their functional properties. As we described in our main submission, we identify problematic objects as those with low likelihood and as soon as a problematic object is identified, we sample a new location from our generative model to reposition it. Fig. 14 shows additional qualitative results on this task. The first and third row show examples of unnatural object arrangements, together with the problematic object, highlighted in green, for each scenario. We note that our model successfully identifies objects in unnatural positions e.g. flying bed (first row, first column Fig. 14), light inside the bed (first row, third column Fig. 14) or table outside the room boundaries (third row, fourth column Fig. 14) as well as problematic objects that do not necessarily look unnatural, such as a cabinet blocking the corridor (first row, sixth column Fig. 14), a chair facing the wall (third row, first column Fig. 14) or a lamp being too close to the table (third row, third column Fig. 14). After having identified the problematic object, our model consistently repositions it at plausible position.

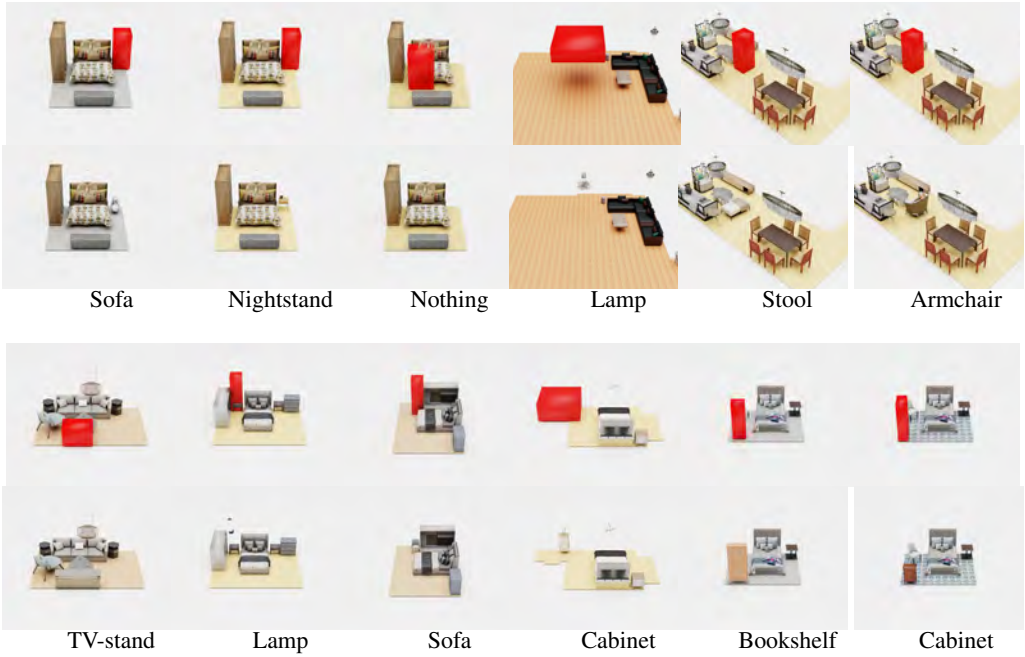


Figure 15: Object Suggestion. A user specifies a region of acceptable positions to place an object (marked as red boxes, first and third row) and our model suggests suitable objects (second and fourth row) to be placed in this location.

D.2 Object Suggestion

For this task, we examine the ability of our model to provide object suggestions given a scene and user specified location constraints. For this experiment, the user only provides location constraints, namely valid positions for the centroid of the object to be generated. Fig. 15 shows examples of the location constraints, marked with red boxes, (first and third row) and the corresponding objects suggested by our model (second and fourth row). We observe that our model consistently makes plausible suggestions, and for the cases that a user specifies a region that overlaps with other objects in the scene, our model suggests adding nothing (first row, third column Fig. 15). In Fig. 15, we also provide two examples, where our model makes different suggestions based on the same location constraints, such as sofa and nightstand for the scenario illustrated in the first and second column and stool and armchair for the scenario illustrated in the fifth and sixth column in the first row.

D.3 Scene Completion

Starting from a partial scene, we want to evaluate the ability of our model to generate plausible object arrangements. To generate the partial scenes, we randomly sample scenes from the test set and remove the majority of the objects in them. Fig. 16 shows examples for various partial rooms (first row Fig. 16), as well as two alternative scene completions using our model (second and third row Fig. 16). We observe that our model generates diverse arrangements of objects that are consistently meaningful. For example, for the case where the partial scene consists of a chair and a bed (last column Fig. 16), our model generates completions that have nightstands surrounding the bed as well as a desk in front of the chair.

D.4 Object Placement

Finally, we showcase the ability of our model to add a specific object in a scene on demand. Fig. 17 illustrates the original scene (first row) and the complete scene (second row) using the user specified object (third row). To perform this task, we condition on the given scene and instead of sampling from the predicted object category distribution, we use the user provided object category and sample the rest of the object attributes i.e. translation, size and orientation. Also in this task, we note that the generated objects are realistic and match the room layout.

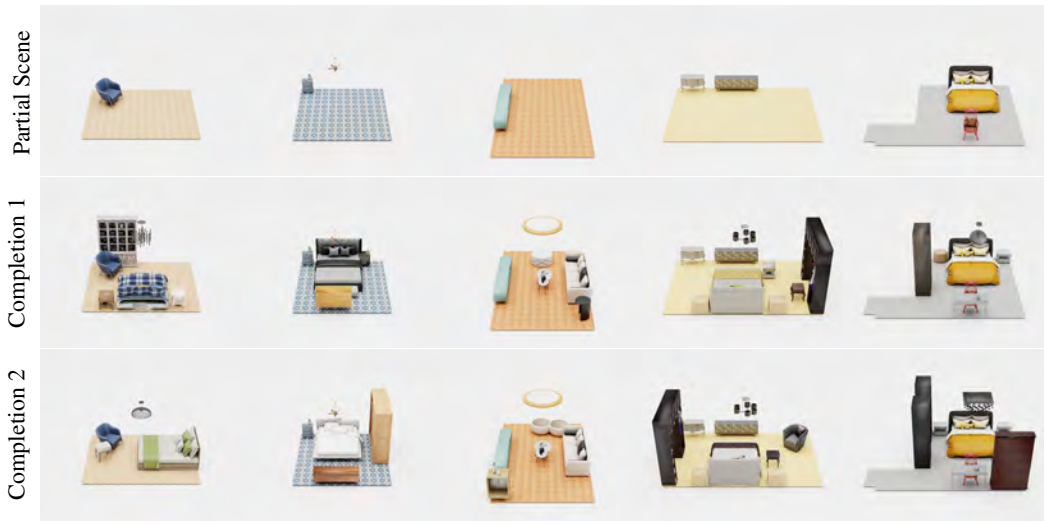


Figure 16: Scene Completion. Starting from a partially complete scene (first row), we visualize two examples of scene completions using our model (second and third row).

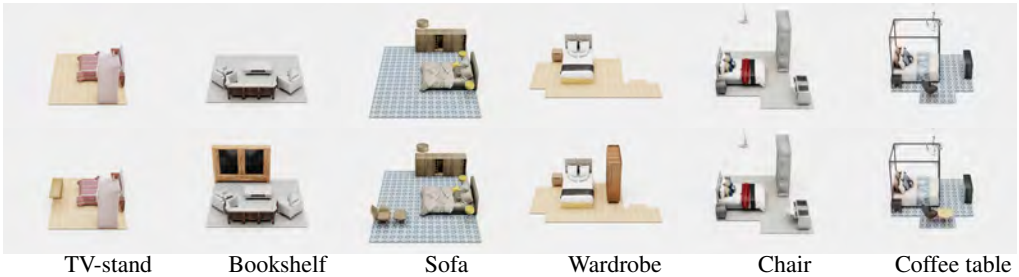


Figure 17: Object Placement. Starting from a partially complete scene, the user specifies an object to be added in the scene and our model places it at a reasonable position. The first rows illustrates the starting scene and the second row the generated scened using the user specified object (third row).

E Scene Synthesis

In this section, we provide additional qualitative results for our scene synthesis experiment on the four 3D-FRONT rooms. Moreover, since, we repeat the FID score and classification accuracy computation 10 times, in Tab. 8, we also report the standard deviation for completeness.

	FID Score (↓)			Scene Classification Accuracy			Category KL Divergence (↓)		
	FastSynth	SceneFormer	Ours	FastSynth	SceneFormer	Ours	FastSynth	SceneFormer	Ours
Bedrooms	40.89 ± 0.5098	43.17 ± 0.6921	**38.39** ± 0.3392	0.883 ± 0.0010	0.945 ± 0.0009	**0.562** ± 0.0228	0.0064	**0.0052**	0.0085
Living	61.67 ± 1.2136	69.54 ± 0.9542	**33.14** ± 0.4204	0.945 ± 0.0010	0.972 ± 0.0010	**0.516** ± 0.0075	0.0176	0.0313	**0.0034**
Dining	55.83 ± 1.0078	67.04 ± 1.3043	**29.23** ± 0.3533	0.935 ± 0.0019	0.941 ± 0.0008	**0.477** ± 0.0027	0.0518	0.0368	**0.0061**
Library	37.72 ± 0.4501	55.34 ±0.1056	**35.24** ± 0.2683	0.815 ± 0.0032	0.880 ± 0.0009	**0.521** ± 0.0048	0.0431	0.0232	**0.0098**

Table 8: Quantitative Comparison. We report the FID score (↓) at 256^2 pixels, the KL divergence (↓) between the distribution of object categories of synthesized and real scenes and the real vs. synthetic classification accuracy for all methods. Classification accuracy closer to 0.5 is better.

Conditioned on a floor plan, we evaluate the performance of our model on generating plausible furniture arrangements and compare with FastSynth [61] and SceneFormer [74]. Fig. 29 provides a qualitative comparison of generated bedroom scenes conditioned on the same floor layout using our model and our baselines. We observe that in contrast to [61, 74], our model consistently generates layouts with more diverse objects. In particular, [74] typically generates bedrooms that consist only of a bed, a wardrobe and less frequently also a nightstand, whereas both our model and FastSynth synthesize rooms with more diverse objects. Similarly generated scenes for living rooms and dining rooms are provided in Fig. 30 and Fig. 31 respectively. We observe that for the case of living rooms and dining rooms both baselines struggle to generate plausible object arrangements, namely generated

objects are positioned outside the room boundaries, have unnatural sizes or populate a small part of the scene. We hypothesize that this might be related to the significantly smaller amount of training data compared to bedrooms. Instead our model, generates realistic living rooms and dining rooms. For the case of libraries (see Fig. 32), again both [61, 74] struggle to generate functional rooms.

E.1 Object Co-occurrence

To further validate the ability of our model to reproduce the probabilities of object co-occurrence in the real scenes, we compare the probabilities of object co-occurrence of synthesized scenes using our model, FastSynth [61] and SceneFormer [74] for all room types. In particular, in this experiment, we generate 5000 scenes using each method and report the difference between the probabilities of object co-occurrences between real and synthesized scenes. Fig. 18 summarizes the absolute differences for the bedroom scenes. We observe that our model better captures the object co-occurrence than baselines since the absolute differences for most object pairs are consistently smaller.

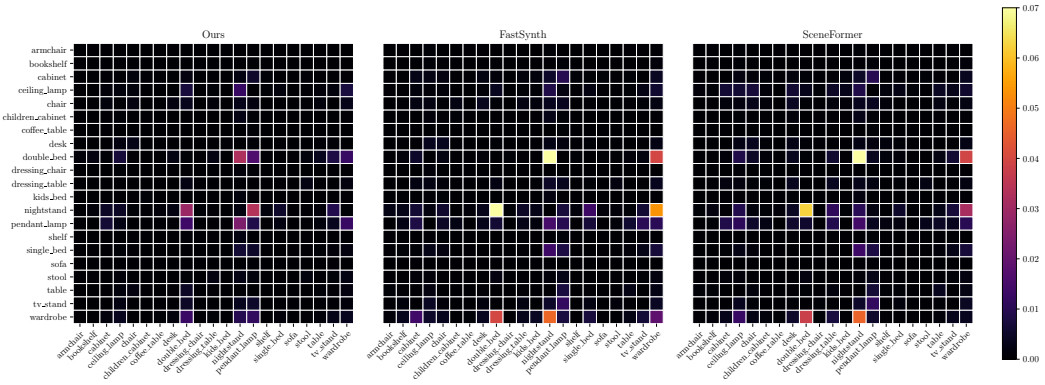


Figure 18: Absolute Difference between Object Co-occurrence in Bedrooms. We visualize the absolute difference of the probabilities of object co-occurrence computed between real and synthesized scenes using ATISS (left), FastSynth (middle) and SceneFormer (right). Larger differences correspond to warmer colors and are worse.

This is also validated for the case of living rooms (Fig. 19), dining rooms (Fig. 20) and libraries (Fig. 21), where our model better captures the object co-occurrences than both FastSynth [61] and SceneFormer [74]. Note that from our analysis it becomes evident that while our method better reproduces the probabilities of object co-occurrence from the real scenes, all methods are able to

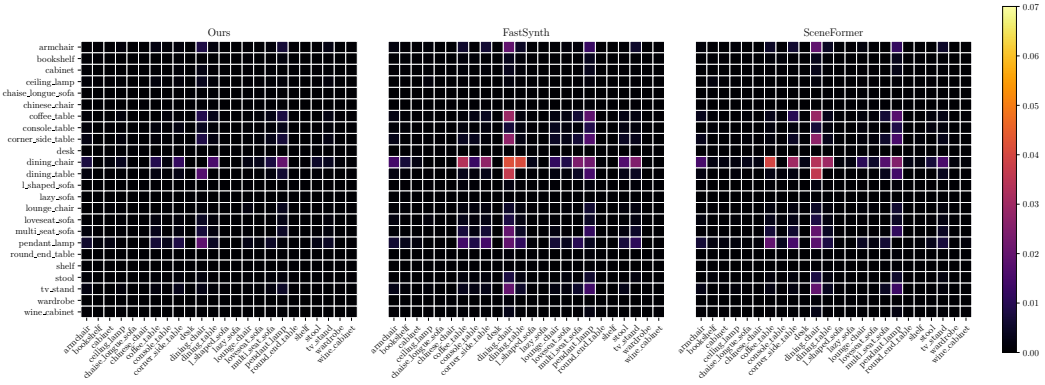


Figure 19: Absolute Difference between Object Co-occurrence in Living Rooms. We visualize the absolute difference of the probabilities of object co-occurrence computed between real and synthesized scenes using ATISS (left-most column), FastSynth (middle column), SceneFormer (right-most column). Lower is better.

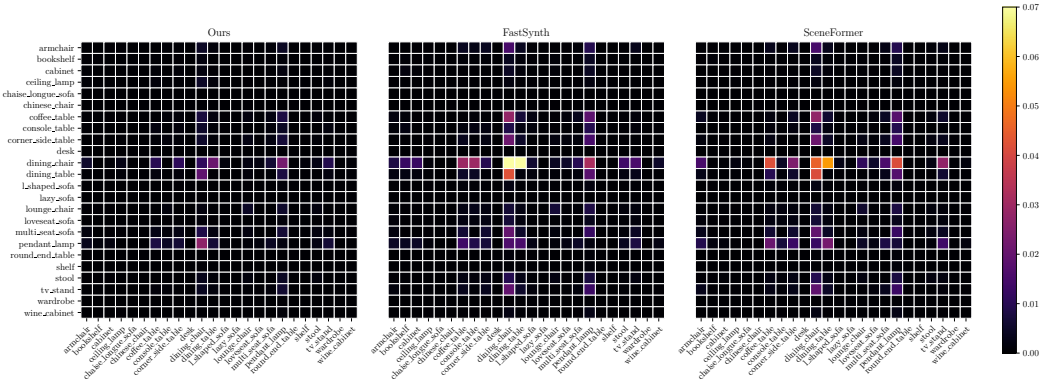


Figure 20: Absolute Difference between Object Co-occurrence in Dining Rooms. We visualize the absolute difference of the probabilities of object co-occurrence computed between real and synthesized scenes using ATISS (left-most column), FastSynth (middle column), SceneFormer (right-most column). Lower is better.

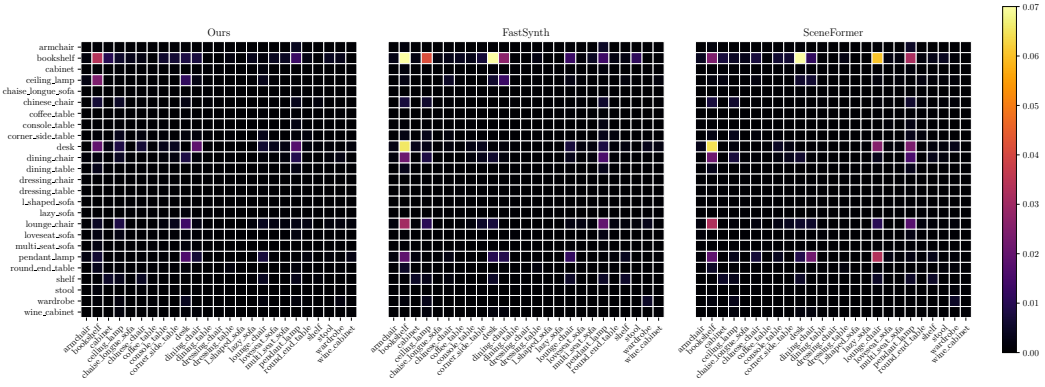


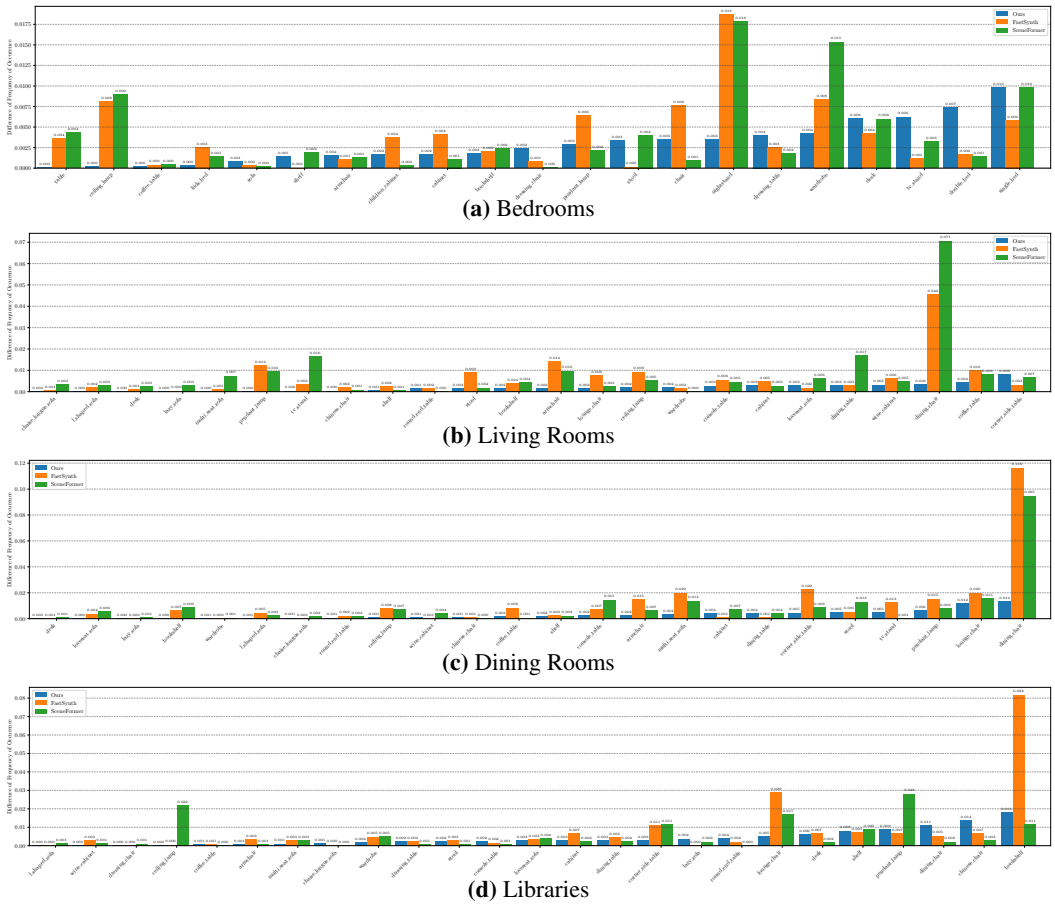
Figure 21: Absolute Difference between Object Co-occurrence in Libraries. We visualize the absolute difference of the probabilities of object co-occurrence computed between real and synthesized scenes using ATISS (left-most column), FastSynth (middle column), SceneFormer (right-most column). Lower is better.

generate scenes with plausible object co-occurrences. This is expected, since learning the categories of objects to be added in a scene is a significantly easier task in comparison to learning their sizes and positions in 3D space.

Finally, in Fig. 22, we visualize the per-object difference in frequency of occurrence between synthesized and real scenes from the test set for all room types. We observe that our model generates object arrangements with comparable per-object frequencies to real rooms. In particular, for the case of living rooms (22b), dining rooms (22c) and libraries (22d) that are more challenging rooms types due to their smaller size, our model has an even smaller discrepancy wrt. the per-object frequencies.

E.2 Visualizations of Predicted Distributions

In this section, we provide examples of the predicted location distributions for different input scenes. In particular, we randomly select 6 bedroom floor plans and conditioned on them we generate 5000 scenes conditioned on each floor plan. Based on the locations of the generated objects, we create scatter plots for the locations of various object categories i.e. chair (Fig. 23), desk (Fig. 24), nightstand (Fig. 25), wardrobe (Fig. 26). We observe that for all object categories the location distributions of the generated objects are consistently meaningful.



(a) Bedrooms

(b) Living Rooms

(c) Dining Rooms

(d) Libraries

Figure 22: Difference of Per-Object Frequencies. We visualize the absolute difference between the per-object frequency of generated and real scenes using our method, FastSynth [61] and SceneFormer [74] for all room types. Lower is better.

E.3 Computational Requirements

In this section, we provide additional details regarding the computational requirements of our method, presented in Table 2 and 3 in our main submission. We observe that ATISS requires significantly less time to generate a scene compared to [74, 61]. Note that the computational cost varies depending on the room type, due to the different average number of objects for each room type. Living rooms and dining rooms are typically larger in size, thus more objects need to be generated to cover the empty space. All reported timings are measured on a machine with an NVIDIA GeForce GTX 1080 Ti GPU.

Even though the implementations are not directly comparable, since we cannot guarantee that all have been equally optimized, our findings meet our expectations. Namely, FastSynth [61] requires rendering the scene each time a new object is added, thus it is expected to be significantly slower than both SceneFormer and our model. On the other hand, SceneFormer [74] utilizes four different transformer models for generating the attributes of each object, hence it is expected to be at least four times slower than our model, when generating the same number of objects.

F Perceptual Study

We conducted two paired Amazon Mechanical Turk perceptual studies to evaluate the quality of our generated layouts against FastSynth [61] and SceneFormer [74]. To this end, we first sampled 211 floor plans from the test set and generated 6 scenes per floor plan for each method; no filtering or post-processing was used, and samples were randomly and independently drawn for all methods.

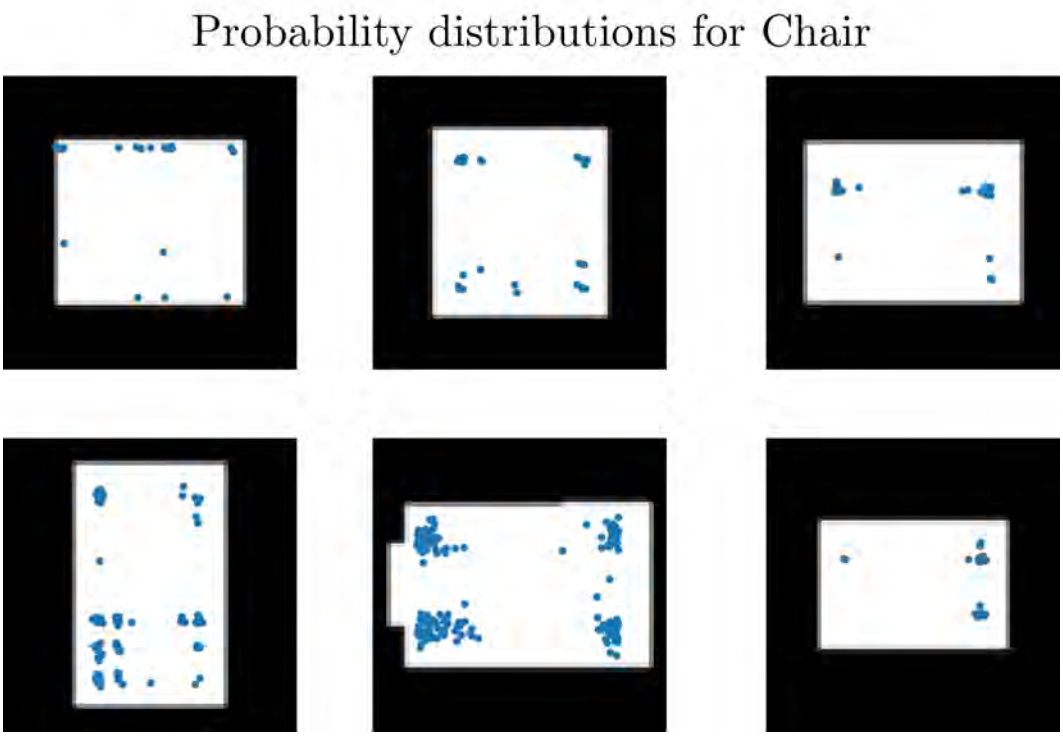


Figure 23: Location Distributions for Chair.

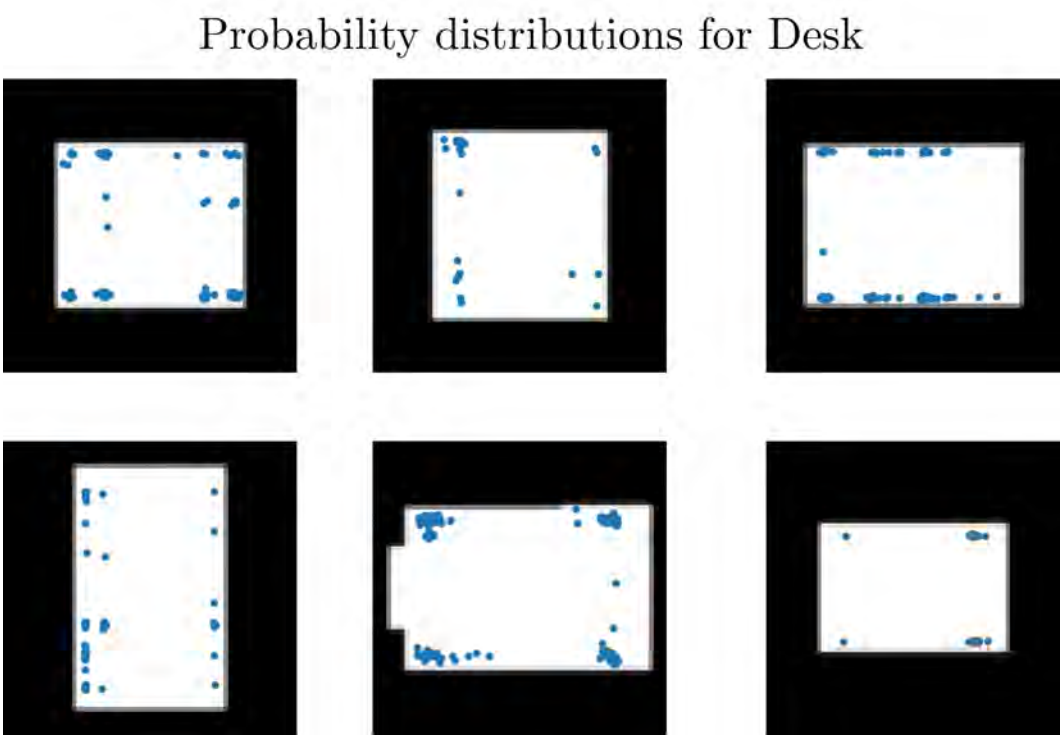


Figure 24: Location Distributions for Desk.

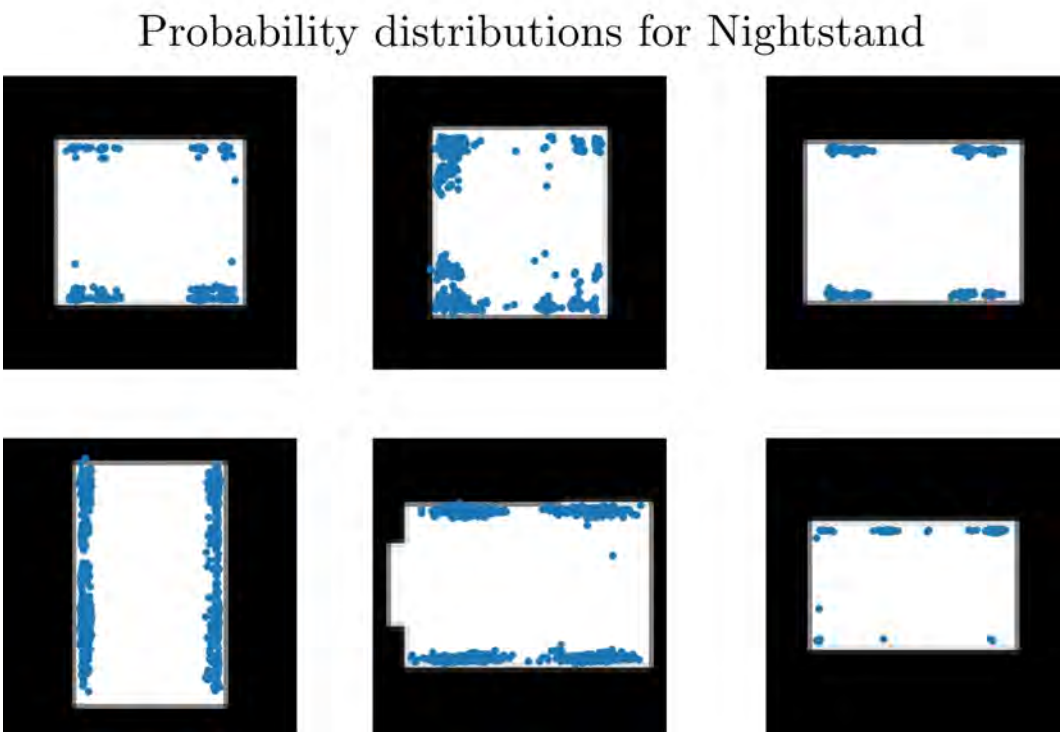


Figure 25: Location Distributions for Nightstand.

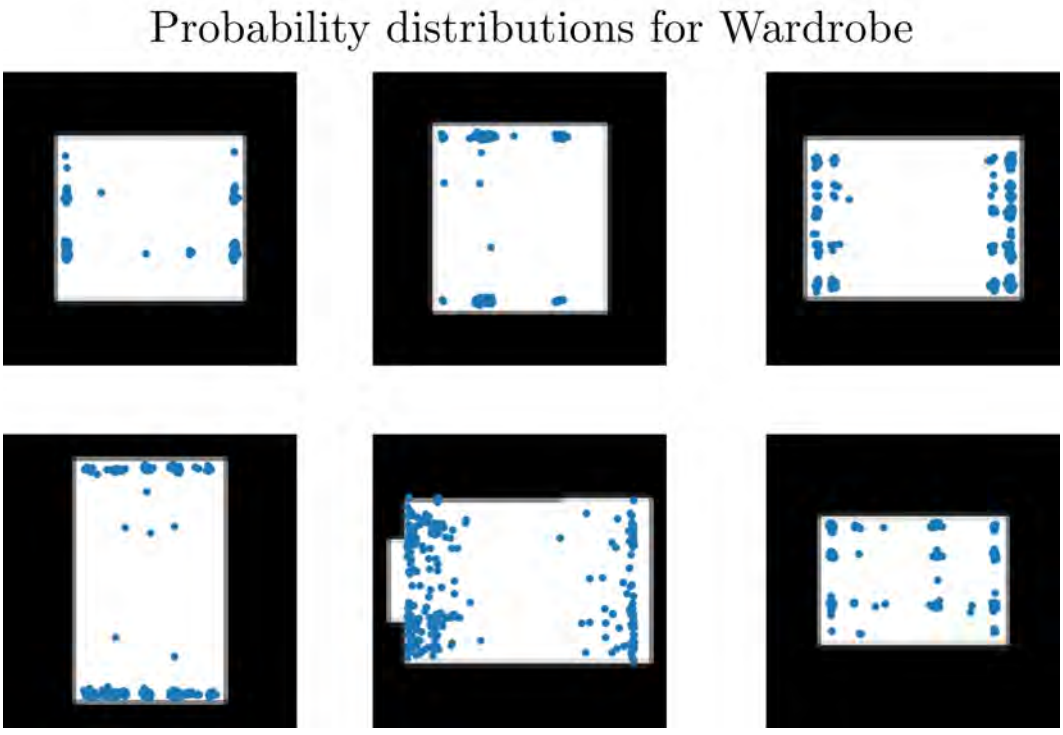


Figure 26: Location Distributions for Wardrobe.

Originally, we considered rendering the rooms with the same furniture objects for each floor plan to allow participants to only focus on the layout itself, which is the main focus of our work. However, since the object retrieval is done based on the object dimensions, rescaling the same furniture piece to fit all predicted dimensions would result in unrealistically deformed pieces that could skew perceptual judgements even more heavily. To avoid having participants focusing on the individual furniture pieces, we added prominent instructions to focus on the layout and **not** the properties of selected objects (see Fig. 27). Each 3D room was rendered as an animated gif using the same camera rotating around the room.

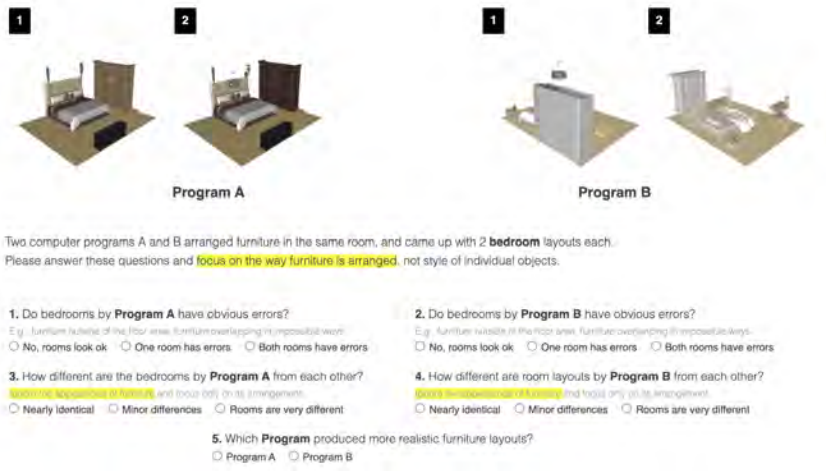


Figure 27: Perceptual Study UI. A/B paired questions with rotating 3D scenes (zoom in).

In each user study, users were shown paired question sets: two rooms generated using our method and two generated with the baseline conditioned on the same floor plan. We randomly selected two out of the 6 pre-rendered scenes for the given floor plan, and 5 different workers answered the question set about every floor plan. Namely, the majority of the 6 layouts were shown more than once on average. A / B order was randomized to avoid bias. The question sets posed the same two questions about scenes generated with program A and B, in order to let users focus on the details of the results and to assess errors of the generated layouts. The last question forced participants to choose between A or B, based on which scene looks more realistic.

Specifically, users were instructed to pay attention to errors like interpenetrating furniture and furniture outside of the floor area and answer if none, one or both layouts for each method had errors. We aggregated these statistics to obtain average error rate per layout, with our method performing nearly twice better than the best baseline [61]. The results on realism in Table 4. in our main submission (first and second row) specify the fraction of the times users chose the baseline over ours. For example, [61] was judged more realistic than ours only 26.9% of the time. Because there was no intermediate option, this means that 73.1% of the time our method was preferred. The last line in Table 4, in our main submission, aggregated preference for our method across both studies.

Workers were compensated $0.05 per question set for a total of USD $106. The participation risks involved only the regular risks associated with the use of a computer.

G Additional Related Work on Indoor Synthesis

In this section, we discuss alternative lines of work on indoor scene synthesis. Fisher et al. [19] propose to represent scenes using relationship graphs that encode spatial and semantic relationships between objects in a scene as well as the identity and semantic classification of each object. Then, they introduce a graph kernel-based scene comparison operator that allows for retrieving similar scenes, performing context-based model search etc. Such representations have been subsequently adopted in models that generate scenes conditioned on user provided constraints and interior design guidelines [44] or rely on a set of example images for generating plausible room layouts [18]. Another line of research [20, 23] leverage activity-associated object relation graphs for generating semantically meaningful object arrangements. Finally, another line of research [4, 43] parses text descriptions into a scene relationship graph that is subsequently used for arranging objects in a 3D scene.

H Discussion and Limitations

Figure 28: Failure Cases. We visualize various failure cases of our model for different toom types.

Lastly, we discuss the limitations of our model and show some examples of failure cases in Fig. 28. One type of failure case that is illustrated in Fig. 28 is overlapping objects, in particular chairs for the case of living rooms and dining rooms (see second and third column in Fig. 28). As we already discussed in Sec. B, to be able to use the 3D-FRONT dataset, we performed intense filtering to remove objects that intersect with each other. However, we found out that not all problematic arrangements were removed from the dataset, which we hypothesize is the reason for such failure cases. Another type of failure case that we observed, which is also related to the existence of problematic rooms in our training data, is the unnatural orientation of objects (e.g. chair facing the bookshelf in first column of Fig. 28 or chair facing opposite of the table in last column of Fig. 28.) Note that these failure cases are quite rare, as also indicated by our quantitative analysis in Sec. 4.1 in the main submission as well as the perceptual study in Sec. 4.3, but our method does not guarantee error-free layouts and there is room for improvement.

Our approach is currently limited to generating object properties using a specific ordering (category first, followed by location, then orientation and lastly size). To further expand the interactive possibilities of our model, we believe that also the object attributes should be generated in an order invariant fashion, similar to the objects in the scene. Furthermore, in our current formulation, the object retrieval is disconnected from the attribute generation. As a result we cannot guarantee that the retrieved objects would match with existing objects in the scene. To address this, in the future, we plan to also incorporate style as an additional object attribute to allow for improved object retrieval. Incorporating style information, would also allows us to generate rooms conditioned on a specific style. Another exciting research direction that we would like to explore is combining ATISS with existing compositional representations of objects [69, 54, 45, 55, 53, 46]. This will allow us to generate 3D scenes with control over the object arrangement, object parts and part relationships. Due to the unique characteristics of compositional representations representations, our generated scenes will be fully controllable i.e. it will be possible to manipulate objects and object parts, edit specific parts of the scene etc.

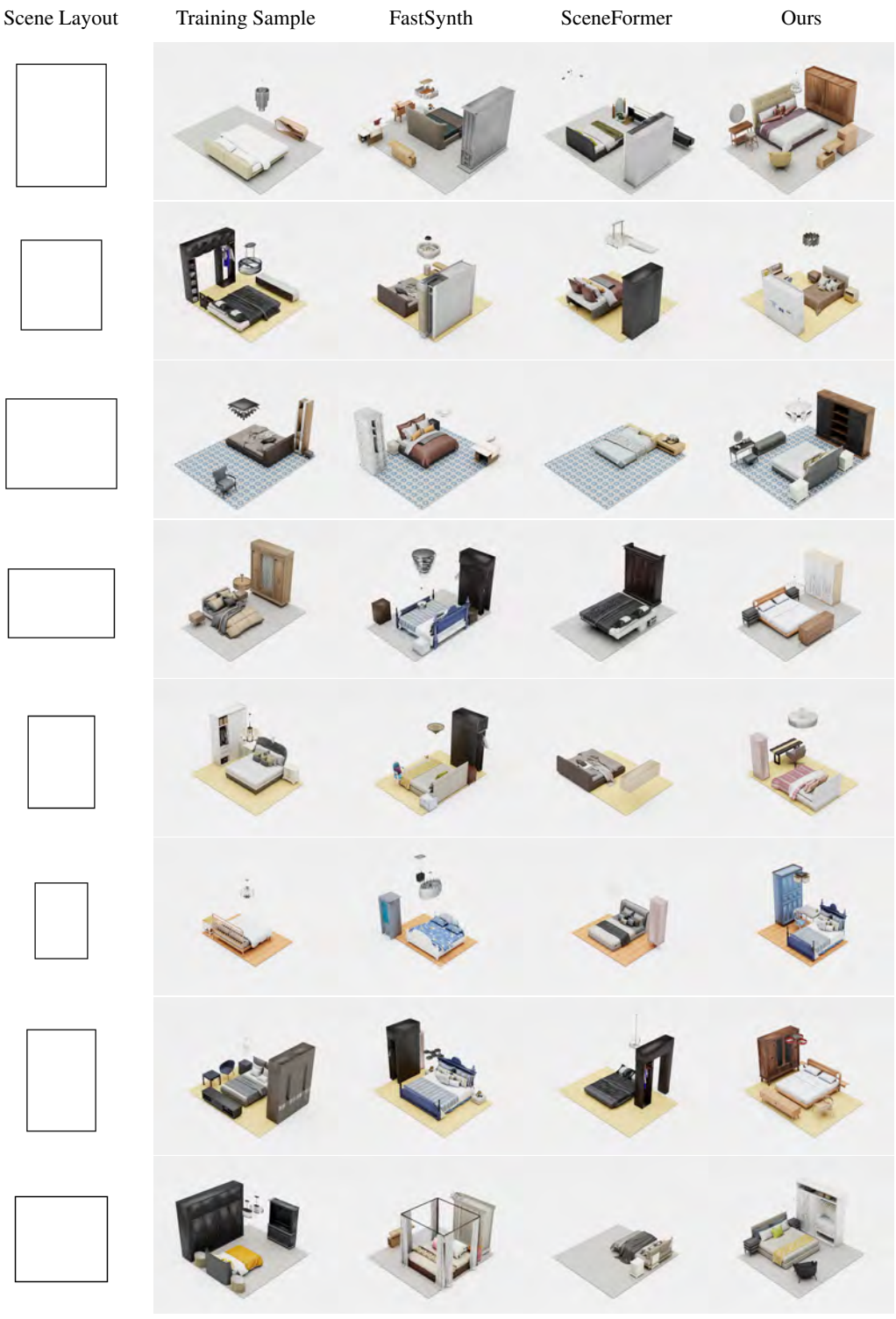


Figure 29: Qualitative Scene Synthesis Results on Bedrooms. Generated scenes for bedrooms using FastSynth, SceneFormer and our method. To showcase the generalization abilities of our model we also show the closest scene from the training set (2nd column).

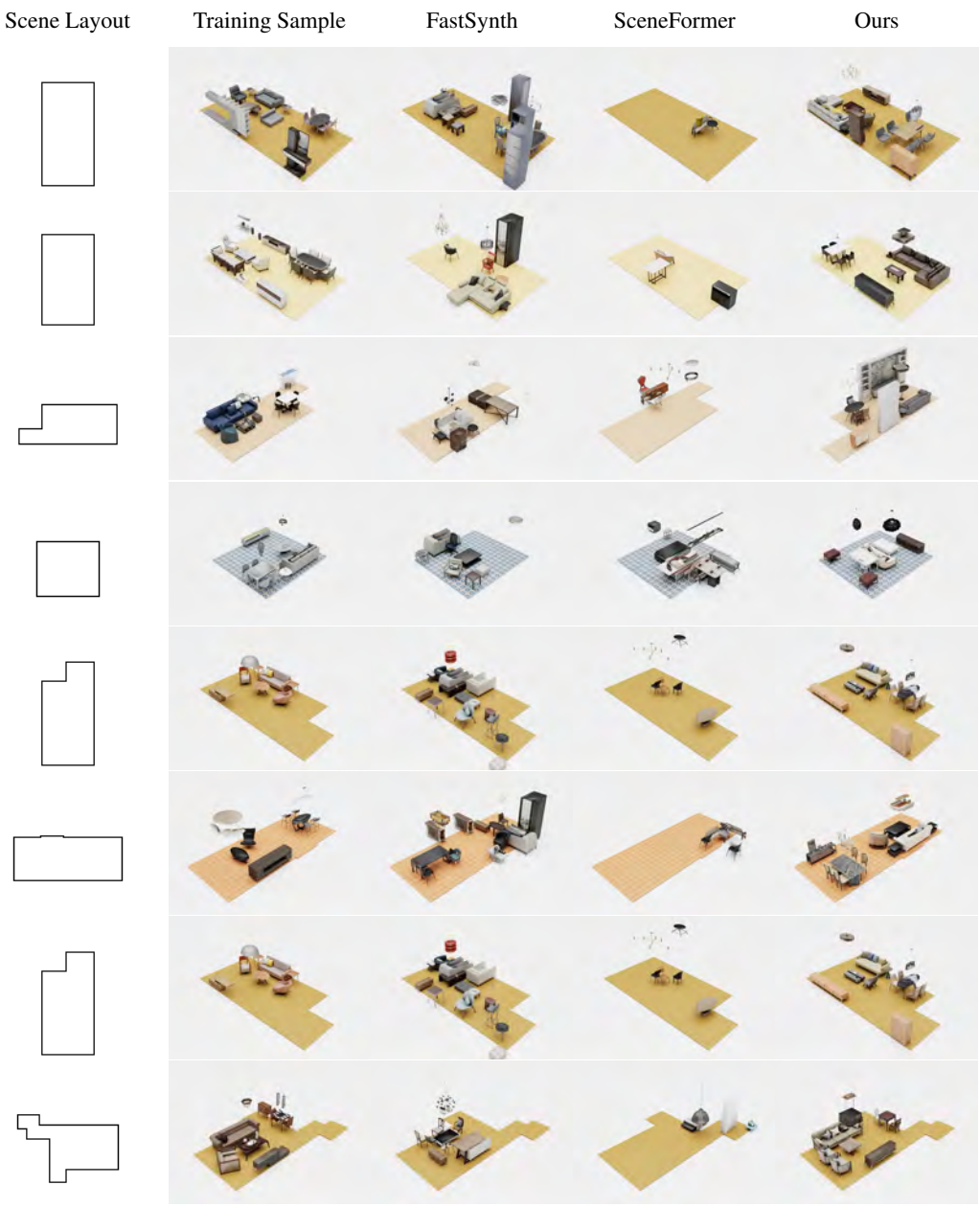


Figure 30: Qualitative Scene Synthesis Results on Living Rooms. Generated scenes for living rooms using FastSynth, SceneFormer and our method. To showcase the generalization abilities of our model we also show the closest scene from the training set (2nd column).

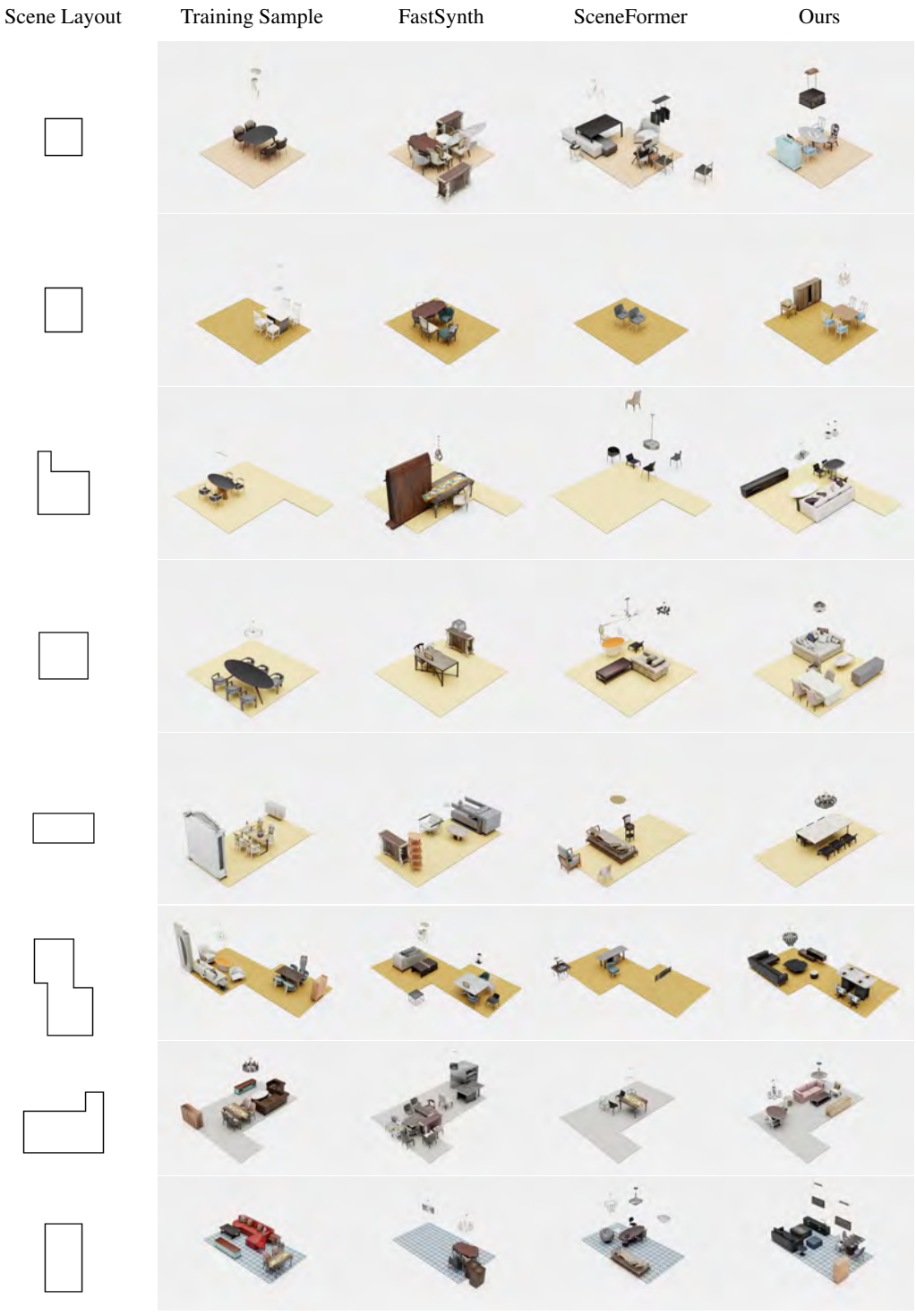


Figure 31: Qualitative Scene Synthesis Results on Dining Rooms. Generated scenes for dining rooms using FastSynth, SceneFormer and our method. To showcase the generalization abilities of our model we also show the closest scene from the training set (2nd column).

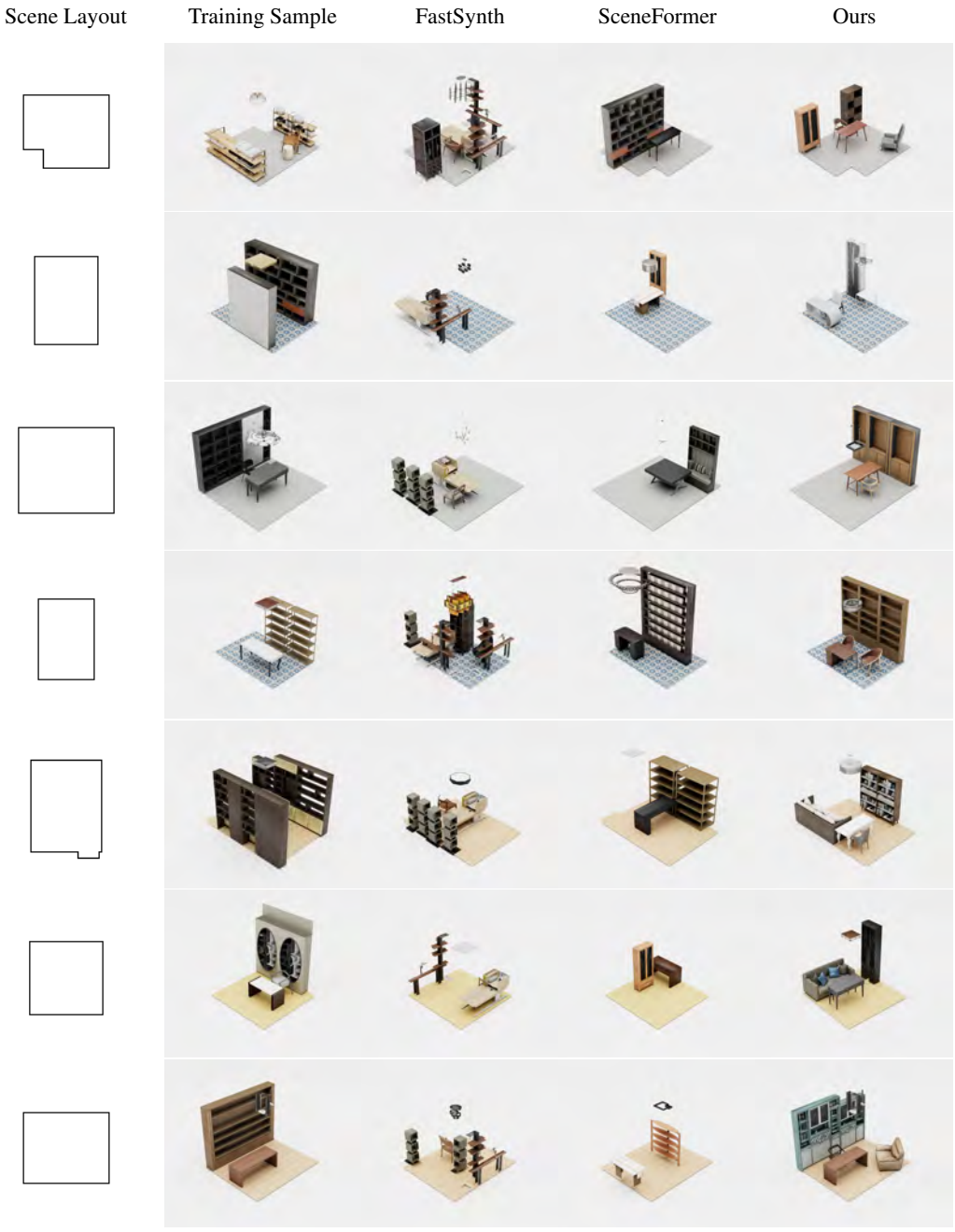


Figure 32: Qualitative Scene Synthesis Results on Libraries. Generated scenes for libraries using FastSynth, SceneFormer and our method. To showcase the generalization abilities of our model we also show the closest scene from the training set (2nd column).